Data Reduction and Feature Selection for Chemometric Analysis

by

Lawrence Asamoah Adutwum

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry
University of Alberta

Abstract

Advancements in data acquisition technologies and the desire for rich data has led to an increase in the size of data collected from modern analytical instruments. With the aid of chemometric techniques, researchers are still able to glean more useful information from these kinds of data than they can with conventional interpretation tools. These chemometric models also benefit immensely from methods that eliminate redundant information.

To make these feature selection methods efficient, strategies to reduce the size of the data prior to their implementation are also desirable. However, in attempting to reduce the data volume, there is an associated risk of information loss or distortion. In chromatography, where multivariate detectors such as mass spectrometers are used, data reduction methods currently available generally resort to elimination of some dimension of the data.

This dissertation presents new approaches to data size reduction for chromatographic data where multivariate detectors are used. The Unique Ion Filter (UIF) was developed as a data reduction strategy for reducing data size without altering the multivariate nature as well as the chemical information in the data. Two types of UIF were developed namely, UIF1D and UIF2D

for one-dimensional and comprehensive chromatography where multivariate detectors are employed. UIF1D and UIF2D were successfully applied to complex data and were found to be very useful. Segmented total ion spectrum (STIS) was also developed to achieve data reduction with partial preservation of retention information for gas chromatography data. STIS is presented as an alignment-free data reduction method which allows inter-laboratory comparison of chromatograms so long as the same anchor compounds are used.

Cluster resolution feature selection (CR-FS) was developed as an objective feature selection algorithm. Hitherto, there existed no guidance to the determination of the two main parameters needed for full automation of CR-FS. This has prevented true automation of the implementation of this algorithm. The development of an empirical approach to guide the selection of these two critical parameters is also accomplished in this dissertation.

Applications of feature selection tools beyond the realm of chromatography are also explored. It is the desire of X- ray crystallographers to be able to predict the crystal structure of crystalline compounds from their elemental compositions. A machine learning approach to this problem was also explored using CR-FS to determine elemental properties that can guide such predictions.

Rapid identification of micro-organism is highly desirable. This task increases in difficulty as one moves down the taxonomic rank. Feature selection with CR-FS in combination with matrix assisted laser desorption ionization mass spectroscopy (MALDI-TOFMS) data presents an opportunity for high throughput and automated method for bacterial identification. The potential of this approach is also explored.

Preface

A version of Chapter Two has been published as *Unique Ion filter; A strategy for GC-MS data processing prior to Chemometric Analysis,* Adutwum L. A., Harynuk J. J., Anal. Chem. 2014, 86(15), 7726 -7733. I was responsible for the UIF software development, implementation, evaluation and data analysis. The method development and analysis of green tea samples by comprehensive gas chromatography-mass spectrometry (GC×GC-MS) were also done by me. In addition to manuscript preparation and editing. Harynuk J. J. was involved in concept formation, guidance, manuscript preparation and editing.

A version of Chapter Three is under review for publication in Journal of Forensic Science as *Comparison of Total Ion Spectra and Segmented Total Ion Spectra as Preprocessing Tools for Gas Chromatography-Mass Spectrometry Data for the Chemometric Analysis of Casework Fire Debris Samples,* Adutwum L. A., Abel J. Robin and Harynuk J. J. (reference number: JOFS-17-303). I was responsible for the segmented total ion spectra (STIS) software development and implementation. In addition to the chemometric analysis of total ion spectra (TIS) and STIS data. In addition to manuscript preparation and editing. Abel J. Robin was involved in forensic analysis of misclassified samples as well as manuscript editing. Harynuk J. J. was involved in concept formation, guidance, manuscript preparation and editing.

A version of Chapter Four is under review for publication in Analytical and Bioanalytical Chemistry as *Estimation of Start and Stop Numbers for Cluster Resolution Feature Selection Algorithm; An Empirical Approach using Null Distribution Analysis of Fisher Ratios,* Adutwum L. A., de La Mata A. P., Bean H. D., Hill, E. J. and Harynuk J. J. (reference number:ABC-01259-2017). I was responsible for software development, implementation, evaluation and data analysis. In addition to manuscript preparation and edits, de La Mata A. P. was responsible for the GC×GC-TOFMS analysis of fabric volatiles compounds from fabric. Bean H. D. and Hill E. J. provided the bacterial data in this study. Harynuk J. J. was involved in concept formation, guidance, manuscript preparation and editing.

Portions of Chapter Five have been published as *Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis,* Oliynyk A. O., Adutwum L. A., Harynuk J. J., Mar A., Chem. Mater., 2016, 28(18), 6672-6681. I was responsible for software development for data organization, importation and data analysis as well as manuscript preparation and edits. Oliynyk A. O. was involved in data extraction, synthesis and characterization as well as manuscript preparation. Harynuk J. J. and Mar A. were the supervisory authors and were involved in concept formation, guidance, manuscript preparation and editing.

Portions of Chapter Five has been submitted for publication as *How to Confuse a Machine: Structure Prediction and Polymorphism through Machine Learning,* Oliynyk A. O., Adutwum L. A., Rudyk B. W., Pisavadia H., Tehrani M. A., Hukhyy V., Harynuk J. J., Mar A., Brgoch J. (reference number: ja-2017-08460p). I was responsible for software development for data organization, importation and data analysis as well as manuscript edits. Oliynyk A. O.'s contribution to this study includes data extraction and processing, synthesis and characterization and preparing the initial draft of the manuscript. Rudyk B. W. was involved with synthesis and characterization of the samples using the X-ray photoelectron spectroscopy (XPS) technique. Pisavadia H. contributed by extracting the crystal structure data from the various data bases used in this study. Tehrani M. A. contributed by performing density functional theory (DFT) calculations of total energy. Hukhyy V. was involved with magnetic property measurements and discussion. Harynuk J. J., Mar A. and Brgoch J. performed the supervisory role on the study.

Dedication

to the memory of my dear mother Ama Nyarko

Acknowledgments

The completion of the work in this dissertation could not have been accomplished without the support of many individuals and funding agencies. To them I owe a lot of gratitude.

I would like to express my heartfelt gratitude to my research supervisor, Dr. James J. Harynuk for his leadership, direction and guidance throughout the period of my PhD studies. I never believed I could accomplish this when I started but you believed in me, for that I am very grateful. My sincere thanks also goes to members of my advisory and examining committees, Dr. Charles Lucy, Dr. John Veinot, Dr. Gabriel Hanna, Dr. Chris Le, Dr. Vincent Bouchard and Dr. Sarah Rutan, your comments, corrections and inputs are appreciated. Much thanks also goes to my collaborators who provided data.

To current and past members of the Harynuk Research Group, I couldn't have done this without your encouragements during various stages of my studies. Much thanks to the Department of Chemistry's Mass Spectrometry Laboratory and Technical Shops and Services for their assistance. Many thanks to my friends in Canada and Ghana for their encouragement at various states of my academic pursuit.

I would like to acknowledge the financial support from the University of Alberta, the Natural Sciences and Engineering Research Council of Canada (NSERC), Chromaleont and Alberta Innovates Technology Futures for funding at various stages of my studies.

My final gratitude is reserved for my siblings Gertrude, Salomey, Hustinx and Joyce, I owe you this. To my dear wife Linda, and my kids Akosua and Papa Kwadwo, you all have been a blessing to me. To my late mum Ama Nyarko, thank you for believing in me, I dedicate this work to your memory.

Contents

Figures

Tables

Abbreviations

^{1}D	First Dimension
^{2}D	Second Dimension
acc	Accuracy of Classification Model
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
COW	Correlation Optimized Warping
CR	Cluster Resolution
CR-FS	Cluster Resolution Feature Selection
$cr_{(max)}$	Cluster Resolution of an External Validation set
EDA	Exploratory Data Analysis
EP	Earthly Paradise
f	Fisher Ratio
f_{NULL}	Null F-Ratio
f_{TRUE}	True F-Ratio
GA	Genetic Algorithm
GB	Gigabyte
GC	Gas Chromatography
GC-MS	Gas Chromatography - Mass Spectrometry

GC×GC	Comprehensive Gas Chromatography
HS-MS	Head Space - Mass Spectrometry
IL	Ignitable Liquid
JD	Jasmine Dragon Tears
KNN	*k* Nearest Neighbour
LC	Liquid Chromatography
LC-MS	Liquid Chromatography Mass Spectrometry
LDA	Linear Discriminant Analysis
LIS	Local Ion Signature
LV	Latent Value
MALDI	Matrix Assisted Laser Desorption Ionization
MS	Mass Spectrometry
NIR	Near Infrared
NMR	Nuclear Magnetic Resonance
OM	Organic Makaibari
OVL	Overlapping Coefficient
PC	Principal Component
PCA	Principal Component Analysis
PDF	Probability Density Function
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
QDA	Quadratic Discriminant Analysis

qMS	Quadrupole Mass Spectrometer
RAFFT	Recursive Alignment by Fast Fourier Transform
SBE	Sequential Backward Elimination
sens	Sensitivity of Classification Model
SFS	Sequential Forward Selection
spec	Specificity of Classification Model
SPME	Solid Phase Microextraction
SR	Selectivity Ratio
SS	Spring Sencha
STIS	Segmented Total Ion Spectrum
SVM	Support Vector Machines
TM	Tamaryoku Cha
TIC	Total Ion Current/Total Ion Chromatogram
TIS	Total Ion Spectrum
TOF	Time of Flight
UIF	Unique Ion Filter
UIF$_1$D	Unique Ion Filter for GC-MS
UIF$_2$D	Unique Ion Filter for GC$\times$GC-MS
UV	Ultraviolet
VIP	Variable Importance to Projection

Symbols

${}^2t_r$	Second dimension retention time
f'	Smoothing window for Savitsky Golay filter
$f_{\mathbf{N}}$	Optimum density function for null F-ratios
$f_{\mathbf{T}}$	Optimum density function for true F-ratios
m	Number of modulations
$\mathbf{M}$	Mask for a raw GC-MS/GC $\times$ GC-MS data
m/z	Mass to charge ratio
$n_{\mathbf{ST}}$	Start number for backward elimination
$n_{\mathbf{SP}}$	Stop number for forward elimination
p	Number of unique ions
um/z	Unique mass-to-charge ratio
$\mathbf{U}$	Null matrix same size as that of GC-MS/GC$\times$GC-MS chromatogram
$\mathbf{v}$	Relative abundance vector
w	Number of scans
$\bar{x}$	Sample mean
$\mathbf{X}$	Matrix containing X dataset
$\mathbf{Y}_{\mathbf{MS}}$	Matrix of Mass spectra for identified peaks
σ	Population Standard deviation
σ^2	Variance

"The significant problems we have cannot be solved at the same level of thinking with which we created them"

Albert Einstein

1

General Introduction

1.1 Motivation

Experiments relying on modern instrumental analyses across many fields generate huge amounts of data because of the advances made in data generation and high-throughput acquisition technologies.[1,2] These kinds of data are rich with information that can enable scientists to ask increasingly difficult questions or probe complex systems. Application of chemometric techniques to glean useful information from data is becoming increasingly important.[3–14] The huge amount of data contains the underlying chemical information in addition to irrelevant variables or noise. More often than not, the number of irrelevant and redundant variables is commensurate with the amount of data.[15,16] Hence, the data can easily overwhelm the underlying information. Prior to chemometric analysis, it is important to eliminate redundant and/or irrelevant variables/noise from the data.[15,16]

This is because their inclusion could be detrimental to the discriminating power of the chemometric models.[15] The ability to obtain useful information out of any kind of data relies heavily on one's ability to sift through and eliminate the irrelevant subset.[17] This presents the proverbial 'needle in a haystack problem'. Even for a small number of variables (i.e., a few hundred), identifying the relevant subset is not a simple problem, and it becomes much more challenging as the number of variables increases (e.g., 1 million).

Cluster resolution (CR) is a model quality parameter developed to estimate the quality of principal component analysis (PCA) models.[18] When CR is combined with a variable space search algorithm such as sequential backward elimination (SBE) and sequential forward selection (SFS), a feature selection algorithm is realized. This algorithm is termed cluster resolution-feature selection (CR-FS).[18] CR-FS has been used as a feature selection method for the classification of gasoline and fire debris.[18–21] CR-FS requires two main inputs which are referred to as the start and stop number for the SBE and SFS stages, respectively. Hitherto, no guidance as to the setting of these parameters existed other than trial and error. This increases model optimization time, introduces subjectivity in its implementation and prevents the true automation of the feature selection process. As the start and stop number varies with the data, it is important to estimate these numbers using some statistical parameters from the data.

The motivation of this dissertation is to contribute to the field of data reduction and feature selection. On the data reduction front, the development of a reduction tool that retains the advantages of currently available methods but eliminates some, if not all, of their drawbacks was explored. On the feature selection front, a complete automation of CR-FS is sought by finding ways to estimate the start and stop number from the dataset. It is hoped that this will make the use of raw data more attractive to people that shy away from high complexity.

1.2 Chemometric Analysis

In 1971, Wold coined the term *'Chemometrics'*, describing it as 'the art of extracting chemically relevant information from data produced in a chemical experiment'.[22] Chemometrics has since been defined as, *the chemical discipline that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments, and to provide maximum chemical information by analyzing chemical data.*[3,4,23–25] To test a hypothesis, experiments are designed to generate some measurements (i.e., data) which are then analyzed to convert the data into information. In that context, chemometrics guides the number and types of experiments, as well as the extraction of useful information from the generated data towards new discoveries or to gain better understanding.[26]

Over the years, several reviews have been published highlighting the development and application of chemometric techniques in several areas.[3–14] The field of metabolomics has evolved to become an exploration of data-rich analytical chemistry measurements with chemometrics.[12,27,28] Recently, chemometric techniques have been very instrumental for biomarker discovery.[28–34]

1.2.1 Pattern Recognition in Chemical Data

Searching for patterns in data is a fundamental endeavor. Pattern recognition problems are present in diverse forms in the world and are extremely important in most decision making processes. Commonly known techniques such as fingerprint and handwriting analysis are essentially pattern recognition problems. As the name implies, this is the application of techniques to identify inherent patterns (information) in data. These patterns are usually not as apparent in multivariate data as they are with fingerprints, for example. In chemometrics, the goal of pattern recognition techniques is ultimately classification of an object/sample.[6,11] Classification involves finding a mathematical model with the capability to recognize the membership of an object (i.e., sample) and assign it to the proper class/group.[35,36] Some of these approaches have been used for source identification of jet fuels,[37] chemical fingerprinting of gasoline,[38–41] tracking and weathering of oil spills,[42,43] classification of casework arson samples,[20]

classification of vinegars and wines,[44–46] biomarker identification,[47,48] drug discovery and verification of herbal medicines,[49,50] amongst others. In metabolomics, pattern recognition techniques have aided breath analysis towards biomarker discovery.[28–34] Pattern recognitions are categorized into two main groups namely unsupervised and supervised.[51]

1.2.2 Unsupervised Pattern Recognition

Unsupervised pattern recognition, also referred to as explorative data analysis, encompasses all approaches employed with several goals including maximizing insights into data, uncovering underlying structures/patterns and identification of outliers.[36,52,53] These techniques are designated unsupervised in that they do not take into account the assigned class or labels of samples. These techniques aid in exploring the otherwise buried patterns in the data with the primary goals of visualization, clustering and projection.[36] Visualization provides a graphical presentation to give a qualitative understanding of information content.[54] Clustering involves organizing the unlabeled data into groups/clusters. A cluster is a group of samples/objects that are more similar to each other than they are to objects in other clusters.[36] Hence cluster analysis is a general term for several algorithms that are used to achieve this goal. The *k*-means algorithm is arguably the most popular clustering technique. Like most clustering algorithms, a class centroid is located amongst the data with samples grouped according to

their distance from the centroid.[36] Distance measures such as Euclidean, Mahalanobis and Manhattan distances are used to guide the grouping of samples in cluster analysis.[26] Hierarchical clustering involves finding clusters in clusters in order to find their relationships.[55–57] Clusters are ranked in a hierarchy to identify those that are subsets of larger clusters. Projection refers to the decomposition of the original data into a lower-dimensional space without altering the structure of the data.[52,58,59] Principal component analysis (PCA) is probably the oldest projection technique; first reported in 1901.[26,59,60] PCA offers an approximation of a data matrix $\mathbf{X}$, into two smaller matrices termed scores ($\mathbf{T}$) and loadings ($\mathbf{P}'$). Visualization by a plot of columns of $\mathbf{T}$ (score plot) and rows of $\mathbf{P}'$ (loading plot) provides the main patterns in objects/samples and variables, respectively.[61] Unsupervised techniques are very useful for probing data to gain new understanding.

1.2.3 Supervised Pattern Recognition

The problem of assigning objects/samples in a data set to groups/classes also known as classification, is a very interesting area of research. Recently, several advances in metabolomics have focused on the use of some mathematical model to classify subjects into groups such as healthy or sick.[28,62] Supervised pattern recognition techniques require the use of data of known class/group assignment to train classifiers (usually some discriminant function) with the aim of assigning new samples to the correct classes.[23]

Unlike unsupervised techniques, supervised pattern recognition methods use the class assignment or labels of samples to obtain the optimum model. A number of discriminant functions have been developed to solve classification problems.[63] The majority of discriminant functions are boundary based methods. Boundary based methods classify samples by finding the optimal boundary that separates groups of samples.[63] Linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), partial least squares–discriminant analysis (PLS-DA), and support vector machines (SVM) are among the most widely used methods for classification.[28,64,65] The selection of an optimal discriminant method for a given data set is guided by the distribution of the samples.[63] Goodness of classification methods is estimated using prediction sensitivity, specificity and accuracy.[66]

1.3 Data Preprocessing and Pretreatment

Data collected from analytical instruments combines signal generated by the chemical information in the sample, as well as artefacts and noise resulting from stochastic variations in experimental conditions and instruments itself. This makes preprocessing a prerequisite for chemometric analysis. Preprocessing involves a series of mathematical transformations performed on the data with each one aimed at mitigating the impact of a particular artefact to improve the quality of the data.[67] In other cases, preprocessing

is meant to eliminate some offset in the data.[68] Preprocessing can have a significant effect on the outcome of the analysis, so it is important to know the sources of artefacts in a particular data set to select the appropriate processing steps.[26,69] Thus, the chemical analytical technique used as well as the research question being probed guides the selection of preprocessing methods. There are various preprocessing techniques available for various kinds of analytical data.[70] These include: baseline drift correction, smoothing, data reduction (binning/bucketing), scaling and centering amongst others.[68–73]

1.3.1 Noise Filtering and Baseline Drift Correction

Noise filtering is aimed at eliminating the background signal from matrix, instrumental interference, measurement noise or baseline distortions from the data.[74] The nature of the measurement signal and noise to be mitigated, the signal-to-noise ratio, as well as the computational resources available influence the choice of noise reduction method.[75] Several signal smoothing algorithms are available. Arguably, the two most popular are moving averages and Savitzky-Golay filters. The Savitzky-Golay filter is generally preferable for chromatographic data since it preserves the shapes of the peaks.[76,77] With chromatographic data, generation of a peak table is one of the simplest approaches to eliminating noisy regions of the data. Peak tables are generated with the aid of a peak finding algorithm which usually incorporates signal smoothing. Baseline drift (i.e., baseline distortion) is a common occurrence

in both chromatographic and spectroscopic data. This is when the ideally flat baseline has a non-zero slope. The presence of drifts in a baseline can influence peak detection and thus adversely affect the accurate quantification of compounds (i.e., signal-to-noise ratio).[78] In gas chromatography (GC), drift may be caused by column bleed because of stationary phase degradation. It may also be caused by contamination of the stationary phase. In liquid chromatography (LC), the primary cause of baseline drift, especially in gradient elution, is the change in the mobile phase composition caused by the changing ratio of solvents. Elimination of baseline drift is an important preprocessing step. There are several methods for correction of baseline drifts.[79–88] In general, baseline drift correction algorithms fit the baseline signal as a function of the x-variable (e.g., time or wavelength) and subtract this function from the original signal. The simplest and most common forms of baseline correction employ polynomial fitting to the baseline.[80,83,84] Wavelet transforms have also been employed for baseline correction,[85–87] as well other novel algorithms relying on partial least squares.[79,88] Recently, a probabilistic peak detection algorithm has been used to estimate the baseline for complex samples.[89]

1.3.2 Data Alignment

Multivariate analyses are performed on a data matrices with samples in rows and variables in columns. It is important to ensure that the

compounds/peaks for all samples in a dataset are aligned. This implies that the signal for a given compound is registered in the same column of the data matrix in all samples. For peak tables this is generally not difficult to achieve; however for raw signals it is somewhat more difficult. Spurious interpretation of results will occur if the data is not properly aligned. Subtle shifts in peak positions do occur in chromatographic and spectroscopic experiments. In chromatography, shifts in retention times are caused by temperature variations, column deterioration/contamination, and variations in mobile phase composition.[90] Stochastic variations in experimental conditions may also contribute to peak shifts. In spectroscopy, temperature variations and inhomogeneity in applied magnetic fields in nuclear magnetic resonance (NMR) coupled with variations in the matrix concentration in samples can cause these variations.[91–93] There are several strategies to correct shifts in peak positions in both spectroscopic and chromatographic data.[94] Alignment algorithms can be classified as peak-based or raw-data based.[95] Peak-based algorithms align peak-tables generated from raw data. They depend on prior peak detection and in some cases deconvolution for true peak area estimations. If the raw data is multivariate, such as GC-MS, similarity scores based on mass spectral comparison are usually used to guide the correct identification and matching of peaks.[96,97] Raw data-based methods use the complete data for alignment. Raw data here may be a set of vectors such as spectra or univariate chromatograms such as GC-FID/NPD data, or it may be

a set of two-dimensional matrices such as with GC-MS data. Algorithms such as correlation optimized warping (COW),[98–100] Interval correlated shifting (Icoshift),[101] and recursive alignment by fast fourier transform (RAFFT)[102] warp the data with the aim of optimizing the correlation between a reference chromatogram (the target) and the sample data to be aligned. Alignment algorithms focused on optimizing correlation between a reference and sample chromatogram fail when there is lack of correlation between chromatograms. For example, GC-MS data from different fire scenes will often lack correlation since the sample matrix varies. For such samples, anchor based methods are generally used.[73,103,104] Anchor based methods use the location of known compounds as anchors to find the shifts between the target and the reference chromatograms. Where anchors may not be consistently located due to high sample-to-sample variability, compounds such as perdeuterated alkanes can be introduced to aid correct alignment.[20] For chromatographic data with MS detection, methods that employ pairwise similarity functions between the total ion chromatogram/total ion current (TIC) and/or mass spectra to find optimal similarity scores have also been reported.[105,106] Genetic algorithm (GA)-based peak alignment methods have been successfully used for aligning raw data.[92,93] Where shifts are minimal, techniques that segment the data and align sections are useful; however, when peak shifts are severe, global alignment algorithms are preferred. Peak-based alignment algorithms are generally faster than raw data based algorithms. Some warping algorithms

allow peak shifts in only one direction, so knowledge of expected shift character may be required to aid the selection of peak finding algorithm.[107]

1.3.3 Data Size Reduction

Reducing the size of the data simplifies multivariate analysis, especially where there are several thousands or millions of variables. For chromatographic data, generation of peak tables where areas estimated from identified peaks are very common.[37,108,109] The accuracy of peak area estimation relies on a robust peak-finding algorithm. Different peak finding algorithms depend on different parameters to identify true peaks, and these demonstrate varied false discovery rates.[110] The use of specific parameters may result in missing real peaks or the identification of false peaks. Non-parametric, probabilistic peak finding and multiple testing local maxima algorithms have also been developed to overcome some of these problems.[111–114] When a multivariate detector (such as MS, UV-diode array, etc.) is used, the resulting data have two dimensions (retention time and detector signal ordinate (e.g., m/z or λ); elimination of one dimension achieves tremendous data reduction. A GC-MS analysis with a run time of 60 min monitoring 100 m/z at a moderate 10 Hz yields 3.6 million variables. Eliminating either retention or m/z results in 100 or 36,000 data points, respectively. However, it would be advantageous if the data reduction was more selective, retaining some of each of the dimensions of the signal.

Development of a strategy to achieve this style of data reduction for GC-MS and comprehensive gas chromatography (GC×GC-MS) data is the subject of Chapter Two.

1.3.4 Centering and Scaling

The appropriate use of centering and scaling contributes to good models and model interpretation. In general, datasets for chemometrics are matrices with samples/objects in rows and variables in columns. Centering is performed on columns while scaling is performed on rows.[68,115] Metabolomic data may have fluctuations even under the same experimental conditions for the same subjects. Different subjects may have different magnitudes among measured analytes.[115] Centering involves the removal of offsets in the columns of the data. Mean centering is the most common centering approach in multivariate analysis. Mean centering implies the mean of each column is subtracted from each element in the column. This converts all responses to variations around zero instead of the mean. This adjusts for the differences in offset between high and low response values (i.e., abundances or concentrations).[115] Centering must be avoided when analyzing response data (i.e., calibration curves).[68,70,115] Multivariate techniques are by design aimed at evaluating the variance/covariance within the data to measure the (dis)similarity in the data. This makes it imperative to know the variance in the data before multivariate analysis is commenced.[70]

Scaling methods multiply/divide individual variables by a factor, thereby adjusting the contribution of different variables to the model. In metabolomics, where concentrations differ by orders of magnitudes, the choice of a good scaling technique improves the data interpretation. Projection methods (e.g., PCA, partial least squares (PLS), etc.), rely heavily on the choice of appropriate scaling of the data.[94,116] Several scaling methods are available in the literature with each aimed at a specific goal.[68,94,115] Auto-scaling, where each column is mean-centered and then divided by the standard deviation of the column is the most commonly used scaling method. This scaling approach tends to make all variables equally important and so each has an equal chance of influencing the model. Pareto scaling differs from autoscaling in that it uses division of each column by the square root of its standard deviation. This reduces and increases the influence of very intense and weaker signals, respectively.[94] Pareto scaling decreases the influence of noise and artefacts which leads to an improvement in the predictive power of chemometric model.[117] Other data transformation techniques, such as range, level and vast scaling as well as log and power transformations, are designed to counter other variations in the data. A comprehensive assessment of these techniques comparing their advantages and disadvantages has been made.[115]

1.4 Feature Selection

Feature selection, also referred to as variable/variable subset selection, involves the choosing of a subset of features to enable the construction of robust models while eliminating as much irrelevant and redundant features from the data as possible.[15,16,118–120] Selection of the relevant feature subset contributes positively to the accuracy and efficiency of chemometric models.[121] Advantages of feature selection include improving performance of machine learning algorithms, aiding visualization of data, reducing the data size to limit storage requirements and associated costs, as well as simplified models.[16,121] 'Relevance' has several meanings within chemometric and machine learning, the simplest of which describes a feature as relevant with respect to a target function, task or research problem. If altering that feature in a sample will alter the output of the target function, the feature is relevant to the function.[119]

Other concepts such as 'strongly relevant', 'weakly relevant' and 'incrementally useful' have also been described.[119,122,123] Feature selection algorithms fall into three main categories namely, filter, wrapper and embedded methods.[15,17,119,121,124] Filter methods employ some characteristics of the training set data to select and exclude some features. Filter algorithms assess the merits of features from the data without the inputs of any induction/learning algorithm.[15,118,119,121] Thus, learning is performed

independent of a classifier. Filter methods are computationally fast, provide a general variable selection independent of any learning algorithm and avoid overfitting.[15] The lack of interaction with classifiers implies that feature dependencies are ignored. Some of the popular filter methods employed are the RELIEF algorithm,[121,125] the FOCUS algorithm,[126] and correlation-based filters.[127–129] Filter feature selection algorithms are merely variable ranking techniques in that these algorithms compute a ranking metric which connotes potential importance of the variables. Hence metrics such as t-test, Mann-Whitney test, mutual information, Pearson's correlation, as well as others can be used.[70]

Wrappers derive their name from the notion that the feature subset selection occurs as a wrapper around the induction/learning algorithm.[15,118,123,124] Thus, the induction algorithm is used as part of the feature selection process as an evaluation function to help estimate the worth of each feature. The feature subspace can be searched via sequential forward selection (SFS) or sequential backward elimination (SBE).[15,124] Randomized subspaces can be employed as well, as is the case with genetic algorithms (e.g., GA-SVM, GA-KNN).[130–133] Wrapper methods generally perform better than filter methods since there is an actual evaluation of the features with a learning algorithm. However, there is a significant risk of overfitting as the feature selection is tuned by the induction algorithm to the specific dataset.[121] Wrappers are also computationally more expensive.

In embedded methods, the feature selection is built into the induction/learning algorithm.[15,134,135] Thus as part of the learning process, weights may be applied to features to achieve optimum performance. Typical examples are the weighting vector of SVM, variable importance to projection (VIP) and, selectivity ratio (SR) in PLS-DA.[136–138] Thus the features selected are only relevant to that specific algorithm. Vieira *et al.* described a fourth group which were referred to as hybrid methods which combine filter and wrapper methods.[135] Here a population of variables of potential importance is first obtained from a filter method and then this population is used by the wrapper method to optimize the feature selection.[18,139]

1.4.1 Cluster Resolution Feature Selection (CR-FS) Algorithm

CR is a model quality metric developed for estimating the separation between two or more clusters of samples in PCA space.[18,21] To determine the CR, confidence ellipses/ellipsoids are generated around the points representing samples of each class in PCA score space. For a pair of classes, the size of the confidence ellipses (both calculated at the same confidence interval (0 to 1)) is adjusted until the ellipses are just touching, without intersecting. For an *n*-class model, pairwise CR values are calculated for each pairing of classes, and the product of all pairwise CR values yields a single metric bounded by 0 to 1 which estimates the overall model quality. CR can be used as an objective function to evaluate the impact of variables on PCA

models. In combination with a SBE and SFS, partial or the entire variable space can be searched for relevant variables. This produces a hybrid (filter and wrapper) feature selection algorithm termed CR-feature selection (CR-FS). The use of CR-FS requires an initial variable subset for the SBE step. This is obtained by first ranking the variables in order of probable relevance by a ranking algorithm. Ranking algorithms such as Fisher ratio (F-ratio) from analysis of variance (ANOVA) and SR for discriminant variable analysis are usually used.[137,138] During the SBE stage, an initial subset of variables which is defined by the start number (i.e., the number of variables used for the SBE) is used to construct a PCA model and the CR is evaluated. The effect of the elimination of the lowest-ranked feature on a new model is evaluated iteratively through all initial features from the lowest-ranked to the highest-ranked. Features whose elimination does not negatively affect the model are eliminated. During the SFS, features that were not considered in the backward elimination are sequentially added, starting with the highest-ranked variable not included in the initial set of variables. After each addition, the effect of adding a variable is evaluated based on CR. If a variable improves CR, it is retained. CR-FS can be performed in both 2D (PC1 vs. PC2) or 3D (PC1 vs. PC2 vs. PC3) score space.[19] CR-FS has been applied to various kinds of data yielding very good results.[18–21,140,141]

1.5 Scope of Dissertation

1.5.1 Unique Ion Filter: A Data Reduction Tool for GC-MS Data Preprocessing for Chemometric Analysis

A routine to eliminate noise and reduce redundancy in data prior to the application of feature selection leads to faster and improved feature selection and better chemometric models. In Chapter Two, a technique termed the Unique Ion Filter (UIF), a data reduction strategy for raw GC-MS data, is reported. UIF achieves data reduction without compromising the chemical and multivariate information in the data. UIF was applied to a gasoline data where the samples were classified according to their octane rating. UIF was subsequently made amenable for the analysis of raw GC×GC-MS data (i.e., UIF 2D). UIF 2D was tested with green tea data to yield classification according to their country of origin.

1.5.2 Comparison of Total Ion Spectra and Segmented Total Ion Spectra as Preprocessing Tools for GC-MS Data for the Chemometric Analysis of Casework Fire Debris Samples

Total Ion Spectrum (TIS) was developed by Sigman *et al.* as an alignment-free GC-MS data processing technique which achieves significant data reduction, preserving the *m/z* information but eliminating the retention information.[142] In Chapter Three, a new implementation of TIS that

offers partial preservation of retention information in a technique termed Segmented Total Ion Spectrum (STIS) is discussed. A comparison of TIS and STIS was made by analysis of casework fire debris samples to be classified based on the presence or absence of gasoline.

1.5.3 Estimating CR-FS Start and Stop Number via Probability Density Function Analysis of True and Null Fisher Ratios

The CR-FS algorithm requires two main inputs: start and stop numbers for the SBE and SFS, respectively. Hitherto, no guidance as to the setting of these parameters exists other than trial and error. This increased analysis time and reduced the consistency of the results obtained, as subjectivity was introduced. The lack of an approach to estimate the start and stop numbers has also prevented the true automation of the feature selection process. Chapter Four discusses an approach to estimate the start and stop number for CR-FS to enable the full automation of the feature selection process is made.

1.5.4 Exploring the Application of CR-FS

The importance of feature selection as a pre-requisite for chemometric modelling cannot be overstated. The use of the right feature selection method can improve model prediction accuracies as well as provide new insights. One example is the prediction of the phase type of binary and ternary equiatomic compounds from elemental properties for inorganic chemists.

The ability to predict the strain of bacteria for the microbiologist from simple analytical experiments is another example. In Chapter Five, CR-FS in combination with SVM/PLS-DA are employed to predict phases of AB and AIB compounds from elemental properties. Strain-level prediction of *Lactobacillus reuteri* using matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) data is also explored.

There are sadistic scientists who hurry to hunt down errors instead of establishing the truth.

Marie Curie

2

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

2.1 Introduction

Gas chromatography (GC) is a versatile tool that has been applied in various fields of chemical analysis including environmental, pharmaceutical, petrochemical, forensics, amongst others. Comprehensive two-dimensional gas chromatography (GC×GC) is a much more powerful separation technique relying on two columns coupled serially by a modulator. Compounds separated in the first column/dimension (^{1}D) are subjected

Adutwum, L. A. and Harynuk J. J., *Anal. Chem.* 86.15 (2014): 7726-7733.

to further separation in the second column/dimension (^{2}D) allowing for a greatly improved peak capacity.[143–146] When these techniques (i.e., GC and GC×GC) are coupled with a mass spectrometer (MS), their separation potentials are further enhanced by the rich data generated from the detector. This makes GC-MS and GC×GC-MS arguably the two most powerful tools in the arsenals of analytical chemists for very complex samples. GC-MS for a long time has been the go-to analytical tool for the analysis of volatile and semi-volatile organic compounds. GC-MS is the standard analytical tool for the analysis petroleum products.[18,19,147,148] GC×GC-MS has been applied in various areas of analysis such as petroleum,[149–154] environmental,[155–159] metabolomics,[33,48,160,161] etc. Recent reviews have highlighted several areas where GC×GC-MS has been applied.[162–166] Mass spectrometers such as time-of-flight MS (TOFMS) 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 raw data are incredibly rich but the sheer data volume, coupled with irrelevant regions of the chromatogram often bury useful information.

Chemometric techniques involve the use of statistical and computational methods to extract useful information from complex chemical data and have become very useful.[22,25] 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 and GC×GC-MS data. Chemometric techniques have been used in the identification of jet fuels,[167] classification and chemical fingerprinting of gasoline,[18,19,38,39,41] tracking and weathering of oil spills,[42,43] classification of casework arson samples,[20] classification of vinegars and wines,[44–46] biomarker identification,[47,48] drug discovery and verification of herbal medicines,[49,50] compound identification[117,168] as well as metabolomics and breath analysis.[31,109,169–171]

Raw GC-MS data presents as a two-dimensional matrix with rows representing mass-to-charge ratio (m/z) and columns representing time (scan#). GC×GC-MS data on the other hand is presented as a three-dimensional with ^{1}D time vs. ^{2}D time vs. m/z. High data rate mass analyzers are 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.[172] In GC×GC separations, much higher data rates (50-500 spectra/s) are required due to the narrow (< 100 ms half-height width) peaks in ^{2}D. In both GC-MS and GC×GC-MS, a 10 min separation monitored at 100 Hz over a window of 50-350 m/z results in 1.8×10^6 variables acquired per sample. If the entire raw data is used for analysis, the number of variables will exceed the number of samples by several orders of magnitude. Furthermore, most of the variables are irrelevant to the intended chemometric

model (empty mass channels or regions of empty chromatographic space, or regions containing signals for unimportant compounds).

The data matrix is highly overdetermined and sparsely populated with mostly irrelevant data. Thus it is difficult (if not impossible) to construct a meaningful model without choosing a subset of the variables to consider as candidates for inclusion. The huge volume of data increases the computational demands, requiring computers with large memories and sometimes parallel computing techniques to enable data analysis.[173] 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.[68,174]

Data reduction is of particular importance for GC-MS and GC×GC-MS data due to the sheer number of variables. Common approaches to data reduction include the use of integrated peak areas based on total ion currents (TICs) or mass spectrally deconvoluted data.[34,37,38,43,108,109] 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.[48,175] 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). Additionally, regions of empty baseline in the chosen ion chromatograms are not excluded with this approach. Recently, a tile-based Fisher ratio analysis approach and local ion signature (LIS) have been demonstrated to achieve data reduction in the analysis of raw GC×GC-MS data.[176–178] These approaches have the added advantage of reducing the need for a strict chromatographic alignment. In the tile-based method, careful selection of the tile size is critical as it must be wide enough to capture the entire peak as much as possible as well as retention time variation in both first and second dimensions. Tiles that are too large could increase the impact of noise in the data and risk the inclusion of multiple peaks in a single tile. Tiles that are too small may not be able to accommodate slight variations in peak locations resulting in peaks being registered in different tiles in different chromatograms due to minor shifts in retention time. For LIS, significant shifts in chromatograms will lead to spurious interpretation of results.

The advantage of using the entire GC-MS chromatogram has been demonstrated and applied to very complex samples.[18–21,167] 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 raw GC×GC-MS data for chemometric analysis is not very popular due to the high data volume. Using 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.[15] To overcome this challenge, relevant variables are obtained using feature ranking and feature selection protocols.[135,138,167,179,180] Selection of the relevant feature subset contributes positively to the accuracy and efficiency of chemometric models.[121] 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.[167] The use of selectivity ratio as a feature ranking technique has also be reported.[19,138] 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.

Sinkov *et al.* developed a hybrid sequential backward-elimination/sequential forward-selection (SBE/SFS) algorithm relying on cluster resolution as the model quality metric for objective selection

of a subset of variables to include a chemometric model.[18,19,21] Briefly, the algorithm creates an initial model using a fraction of top-ranked variables (e.g., by F-ratio or SR). The quality of the model is evaluated using CR.[18] During the SBE 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-lower-ranked variable is tested. In the SFS 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 and GC×GC-MS data where high data rate detectors are used, as the vast majority of the variables are uninformative. Results from earlier studies showed that several hundreds or even thousands of variables were selected for a single chromatographic peak.[18,21] 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 would add stability to a model as the variables would 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 overfitting the training set data and/or confusion of the learning algorithm, in this case the feature selection process.[124,134,135,181] 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 research, I developed a preprocessing technique termed unique ion filter (UIF) for automated data reduction prior to chemometric analysis. UIFs developed are termed UIF1D and UIF2D for raw GC-MS and GC×GC-MS data respectively. In UIF1D for GC-MS, 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. In GC×GC-MS the presence of the modulator leads to the splitting of 1D peaks in smaller peaks termed sub-peaks. Hence UIF2D automated GC×GC-MS data reduction retains for each identified peak, a specified number of sub-peaks in ^{1}D, and for all retained sub-peaks a specified number of spectra in ^{2}D and few of the most abundant, unique ions *um/z* in the *m/z* dimension. Essentially, the UIF1D and UIF2D objectively filters each raw GC-MS/GC×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 artefacts, resulting in more parsimonious models.

2.2 Experimental Data

2.2.1 Data UIF1D for GC-MS

A dataset used for a previously published work[18] was used in the proof-of-principle work for UIF1D for GC-MS. 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# $\times m/z$) matrix. A detailed sample extraction and analysis for this data can be found in Appendix A.

2.2.2 Data for UIF2D for GC×GC-MS

The data used to test UIF2D consisted of five green tea samples from three different countries, namely Organic Makaibari (OM) - India; Spring Sencha (SS) - Japan; Tamaryoku Cha (TM) - Japan; Earthly Paradise Jasmine (EP) - China; and Jasmine Dragon Tears (JD) - China were obtained from specialty tea suppliers in Edmonton. The tea volatiles were extracted from the ground, dried tea leaves using headspace solid phase microextraction (SPME) and analyzed using a Pegasus®GC×GC-TOF MS system (Leco, St Joseph, MI, USA). A total of 12 chromatograms were collected for each sample except OM, where 16 chromatograms were collected. The individual raw chromatographic data from each sample was exported from ChromaTOF (version 4.50.8.0, Leco, St Joseph, MI, USA) in .csv format and subsequently imported into Matlab®. Each imported data consisted of 50.054 $\times 10^6$ variables, i.e., ^{1}D = 380 (number of modulations), ^{2}D = 500 (number of spectra) and m/z = 266. A data matrix obtained from entire dataset consisted of 64 samples× 50.54 $\times 10^6$. The samples were to be classified according to the country of origin, i.e., China (EP and JD), India (OM), and Japan (SS and TM). A detailed sample extraction, analysis and data alignment can be found in Appendix A.

2.3 Theory

UIF1D and UIF2D are additional preprocessing steps applied to each individual GC-MS or GC×GC-MS sample chromatogram prior to chemometric analysis. This step was applied before feature selection. Figure 2.3.1 shows the data analysis workflow with or without UIF. The fundamental principle behind unique ion determination is the same for UIF1D and UIF2D.

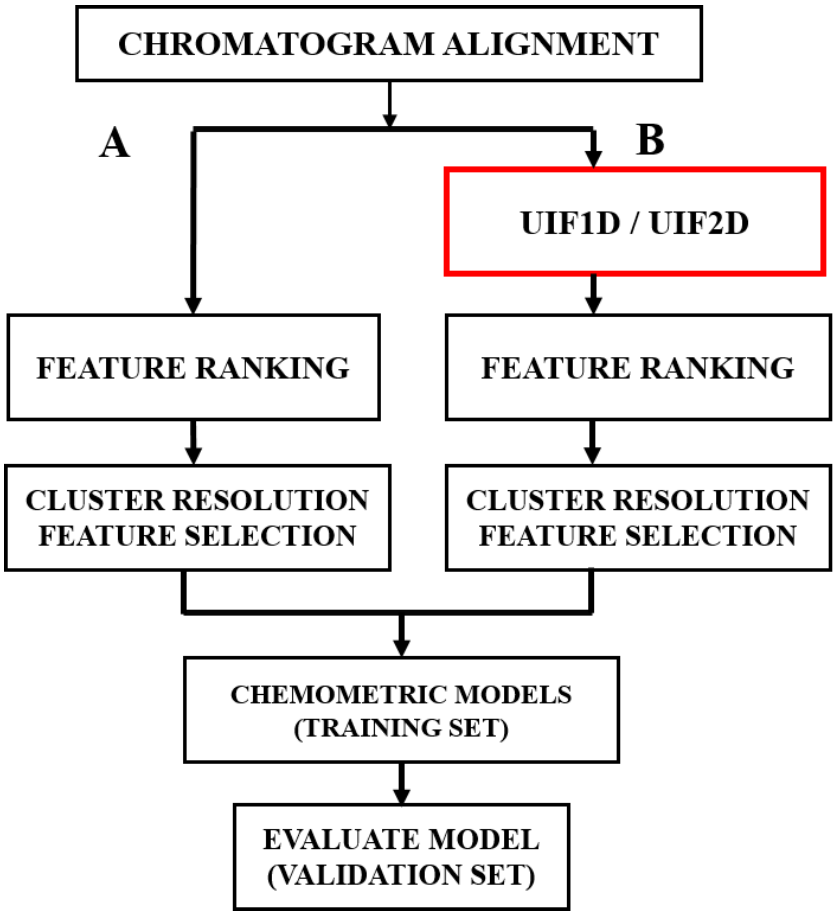


Fig. 2.3.1: Data analysis workflow without (A) and with (B) UIF.

2.3.1 Algorithm for UIF1D and UIF2D

There are two main inputs for UIF1D, 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. UIF2D, on the other hand requires three main inputs, which are the maximum number of sub-peaks, the maximum number of unique ions *um/z* to be retained for each sub-peak and the number of spectra surrounding the apex of each retained sub-peak to be included. Accurate peak detection is necessary for effective application of the UIF algorithm. This is because peak apex locations, peak widths, and for UIF2D, the number of sub-peaks are needed for the implementation of UIF. In principle, any peak detection algorithm that is capable of detecting peak apexes, starts, and stops can be used. In further discussion, the notation of $\mathbf{UIF1D}_{(p,\,w)}$ and $\mathbf{UIF2D}_{(m,\,p,\,w)}$ are used where m is the number of modulations in the case of GC$\times$GC-MS data, p is the number of unique ions to retain for each peak/sub-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 spectra (the peak apex plus two spectra to either side of the apex) would be retained.

2.3.2 Determination of peak parameters and peak groups for UIF1D

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 *m/z* dimension (Equation 2.1), where **X** is the raw chromatogram, **z** is the TIC vector, *i* is the spectra number, *j* is the *m/z* and *J* is the total number of ions.

$$\mathbf{z}_i = \sum_{j=1}^{J} \mathbf{X}_{(i,\,j)} \tag{2.1}$$

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 Equation 2.2, where **sdz** is the second derivative vector, **s** is the second-derivative Savitsky-Golay smoothing vector, **z** is the TIC, f' is the smoothing window and n is the length of **z**.

$$\mathbf{sdz}_i = \mathbf{s}^T \times \mathbf{z}_{\left(i-\frac{f'-1}{2} \,:\, i+\frac{f'-1}{2}\right)} \tag{2.2}$$

$$\frac{f'-1}{2} \leq i \leq \frac{f'+1}{2}$$

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 locations 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.

Three different types of peak groups can be identified from peak start and peak stop locations as shown in Fig. 2.3.2. Group A are resolved peaks, where peak start and peak stop locations do not overlap with any adjacent peaks. Group 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 spectra around peak apexes to be used are then passed to the UIF1D algorithm.

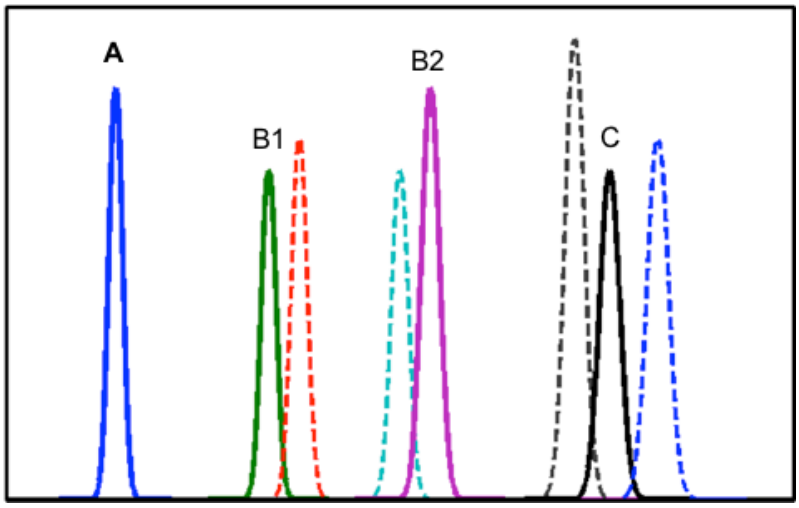


Fig. 2.3.2: Peak groups for UIF for GC-MS data. Peaks are grouped according to their resolution. A - resolved peaks, B1, peaks on the left of two co-eluting peaks, B2 - peaks on the right of two co-eluting peaks and C - peaks sandwiched between two peaks. The group determines how the *um/z* are identified.

2.3.3 Identification of Unique Ions for GC-MS chromatogram

The signals at all peak apexes for a chromatogram are extracted into a matrix ($\mathbf{Y}$) with dimensions of number peaks (N)$\times$ m/z. The extracted signals in $\mathbf{Y}$ are converted into a mass spectrum matrix, $\mathbf{Y_{MS}}$, according to Equation 2.3 where $\mathbf{Y_{MS}}$ are the mass spectra at the apexes, n is the peak number, and j is m/z.

$$\mathbf{Y}_{\mathbf{MS}(n,\,j)} = \frac{\mathbf{Y}_{(n,\,j)}}{\sum_{j=1}^{J} \mathbf{Y}_{(n,\,j)}} \tag{2.3}$$

The group (A, B, C) 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 $\mathbf{Y_{MS}}$). 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 $\mathbf{Y_{MS}}$ (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 Equations 2.4 or 2.5, respectively, where j is the m/z.

$$\mathbf{v} = \frac{\mathbf{Y}_{\mathbf{MS}(n,\,j)}}{\mathbf{Y}_{\mathbf{MS}(n-1,\,j)}} \tag{2.4}$$

$$\mathbf{v} = \frac{\mathbf{Y}_{\mathbf{MS}(n,\,j)}}{\mathbf{Y}_{\mathbf{MS}(n+1,\,j)}} \tag{2.5}$$

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 (Equation 2.4) or $(n+1)$ in (Equation 2.5). Truly unique ions in $\mathbf{v}$ will have a value of ∞, while pseudo-unique 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 co-elutant on both sides) two abundance vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ are calculated using equations 2.4 and 2.5, 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 comprised of zeros, with ones located at positions indicating ions that are unique (or pseudo-unique) to peak *n* in the cluster of three peaks. The coordinates of these *um/z* are set to a value of 1 in $\mathbf{U}$. The resulting matrix $\mathbf{U}$ is a sparse matrix of zeros and ones with the ones indicating the positions of *um/z* for each peak. A Hadamard product of $\mathbf{U}$ and $\mathbf{Y_{MS}}$ yields $\mathbf{V}$ ($\mathbf{V} = \mathbf{U} \circ \mathbf{Y_{MS}}$), a matrix of the raw abundance of each *um/z*. Based on 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.

2.3.4 Generation of new UIF1D filtered chromatogram

In the final step of the UIF, a mask of zeros, $\mathbf{M}$, 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 in the spectra direction, centered on the peak apex. A Hadamard product of $\mathbf{M}$ and the original data matrix $\mathbf{X}$ results in the unique ion filtered data, $\mathrm{UIF}_{(p,\,w)} = \mathbf{M} \circ \mathbf{X}$.

2.3.5 Determination of peak parameters and peak groups for UIF2D

A similar peak detection approach is used for the identification of peaks in GC×GC-MS data. This is because the raw data obtained from GC×GC-MS is actually a 2D data represented as m/z×spectra/time. Thus once the data is summed in the m/z dimension, the TIC obtained is analogous to that of 1D GC-MS. Thus a similar routine for peak detection applied to 1D GC-MS data earlier can be used (Equations 2.1 and 2.2). Fig. 2.3.3 shows a typical TIC obtained from the 2D data (m/z×scan) of a GC×GC-MS separation. However, since the peaks in the ^{1}D are split into subpeaks, it is important to identify the subpeaks which belong to the same compound. Thus after peaks and peak parameters have been determined, the chromatogram is folded into a 3D matrix ($\mathbf{G_A}$) with the aid of the modulation period and the data rate of the MS detector. For a peak modulated three times, three apexes will be identified in three successive modulations. Sub-peaks from sequential modulations belonging to the same compounds are identified based on second dimension retention time (^2t_r) and peak width comparison described by of Peters *et al.*[182] in addition to mass spectra matching using the weighted cosine correlation

score by Kim *et al.*.[183] Compared to a GC-MS, GC×GC separation provides a two-dimensional separation space in which a peak can theoretically be surrounded by more than two neighboring peaks. Thus for a GC×GC, the possible peak groups identified are shown in Fig. 2.3.4.

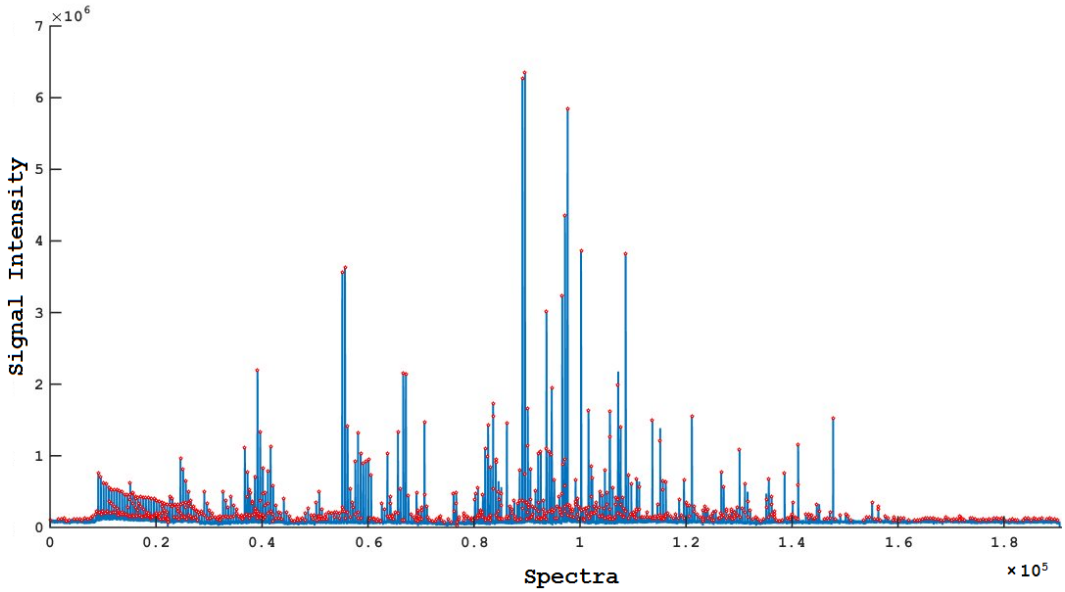


Fig. 2.3.3: A 1D version of a TIC generated from GC×GC-MS separation. The peak apexes are marked with red dots.

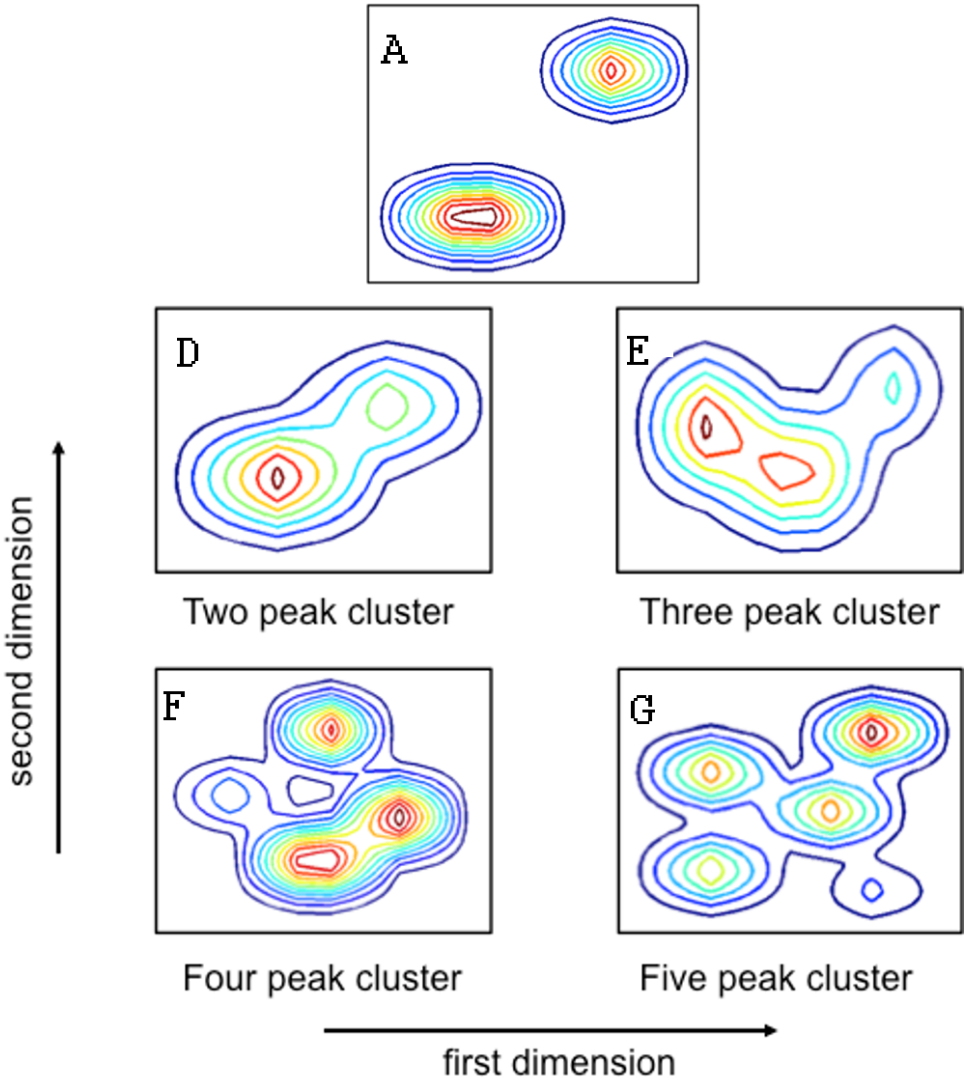


Fig. 2.3.4: Peak groups for UIF2D for GC×GC-MS identification.

2.3.6 Identification of Unique or Pseudo-unique Ions for UIF2D

Theoretically, in GC×GC separation an unresolved peak can be surrounded by several peaks, unlike GC-MS. The identification of *um/z* for D and E in Fig. 2.3.4 are analogous to that of B and C in Fig. 2.3.2. However for peak falling into groups F and G, pairwise comparisons of all mass spectra from the apex of all co-eluting peaks are used to determine the most unique or pseudo-unique ion.

2.3.7 Generation of UIF2D filtered chromatogram

To generate a new UIF2D filtered GC×GC chromatogram, a null matrix, **M**, the same size as the folded GC×GC-MS data is generated. **M** is modified such that ones are placed at the coordinates of the *m* sub-peaks centered on the base peak in ^{1}D, *p* most-abundant *um/z* in each of the *m* sub-peaks within a *w* scans/spectra in ^{2}D centered on the peak apexes of the sub-peaks being retained. **M** is thus a mask for the original data. A Hadamard product of **M** and the aligned original GC×GC chromatogram results in the unique ion filtered data.

2.3.8 Chemometric analysis

The GC-MS data consisted of gasoline samples to be classified according to their octane ratings. Each sample chromatogram was imported as a data matrix from .csv files and aligned using an algorithm written in-house

which is based on a piecewise alignment algorithm.[184] 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.

For the GC×GC-MS data, individual chromatograms were imported into Matlab® as the unfolded GC-MS chromatogram, a matrix of 190,000 rows ×266 columns (i.e., number of spectra × m/z). Each chromatogram was folded into a 3D cube using the modulation period (5 s) and the data rate (100 Hz) and aligned. A data set matrix consisting of 64 rows × 50.054 $\times 10^6$ columns (samples × variables) resulted. Columns representing m/z where no signals were detected above a set threshold (200 for this data set) for all samples were removed from further computation. As indicated earlier, the tea data was to be classified according to the country of origin.

For both UIF1D and UIF2D, two experimental pathways were explored to determine the effect of UIF on the data. Fig. 2.3.1. shows the two experimental routes explored. In pathway A (benchmark pathway), UIF was not implemented. In pathway B, UIF1D and UIF2D were applied to GC-MS and GC×GC-MS, respectively, before feature selection. Each dataset was split into two-thirds for training and one-third for validation sets. The training set data was used for variable ranking and optimization. Feature ranking was performed with the training set data using an ANOVA-based ranking technique reported earlier.[19,109]

Specificity, sensitivity, and accuracy of each optimized model were calculated based on validation data and used as an objective parameter in comparing model quality for both routes.[66] 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 model's ability to correctly classify or predict negative results/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.

2.4 Results and Discussion

The UIF1D offers a convenient approach for automated, objective reduction 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, 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 GC-MS 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, I benchmarked this work without the UIF 1D 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. I chose this approach because it was readily available and has demonstrated success in handling entire raw GC-MS chromatograms.[18–21] 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 Fig. 2.4.1, an 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. Increasing the maximum number of features tested also increases the computation time for the feature selection process.

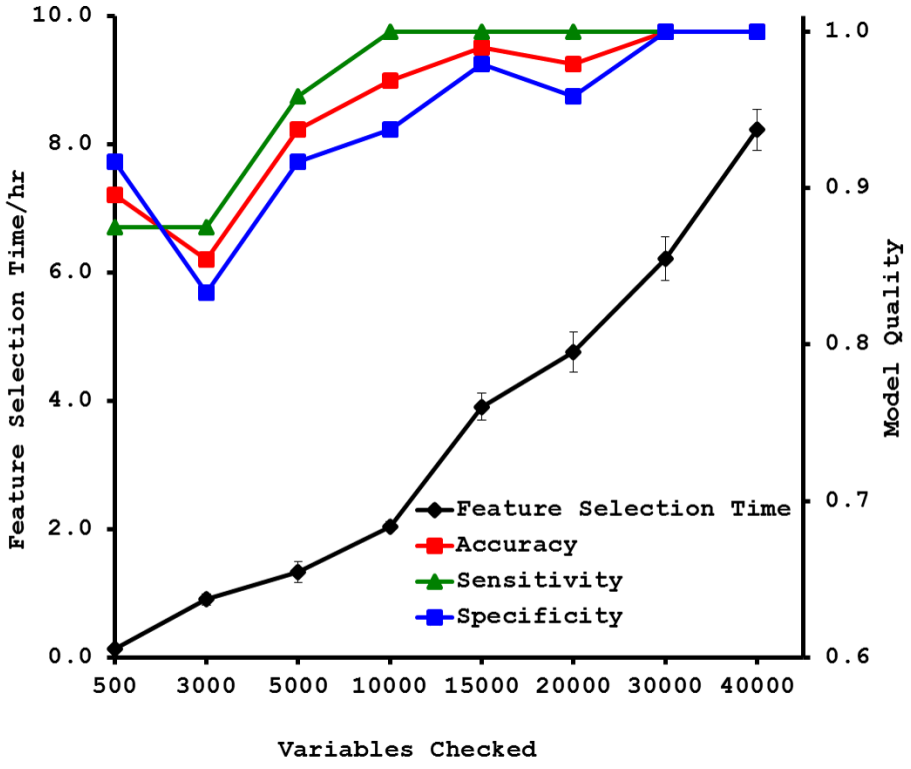


Fig. 2.4.1: Feature Selection time and model quality plot for benchmark pathway GC - MS. The number of variables evaluated during the application of CR-FS was increased until the PLS-DA prediction accuracy on the external validation set was 100%. The total number of variables evaluated until this was achieved was 30000. The error bars indicate the standard deviation for n = 5.

To compare the effect of UIF1D 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. p ranging from 1 to 10 and w 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 3,717 for $\mathbf{UIF1D}_{(1,\,1)}$ to 107,982 for $\mathbf{UIF1D}_{(10,\,17)}$. 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 $w = 1$ (i.e., only ions at the peak apex are retained), an increase in p considered does not improve the model (Fig. 2.4.2). However, 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 (2) *um/z* and three (3) spectra is necessary to achieve 100% model prediction sensitivity, specificity and accuracy.

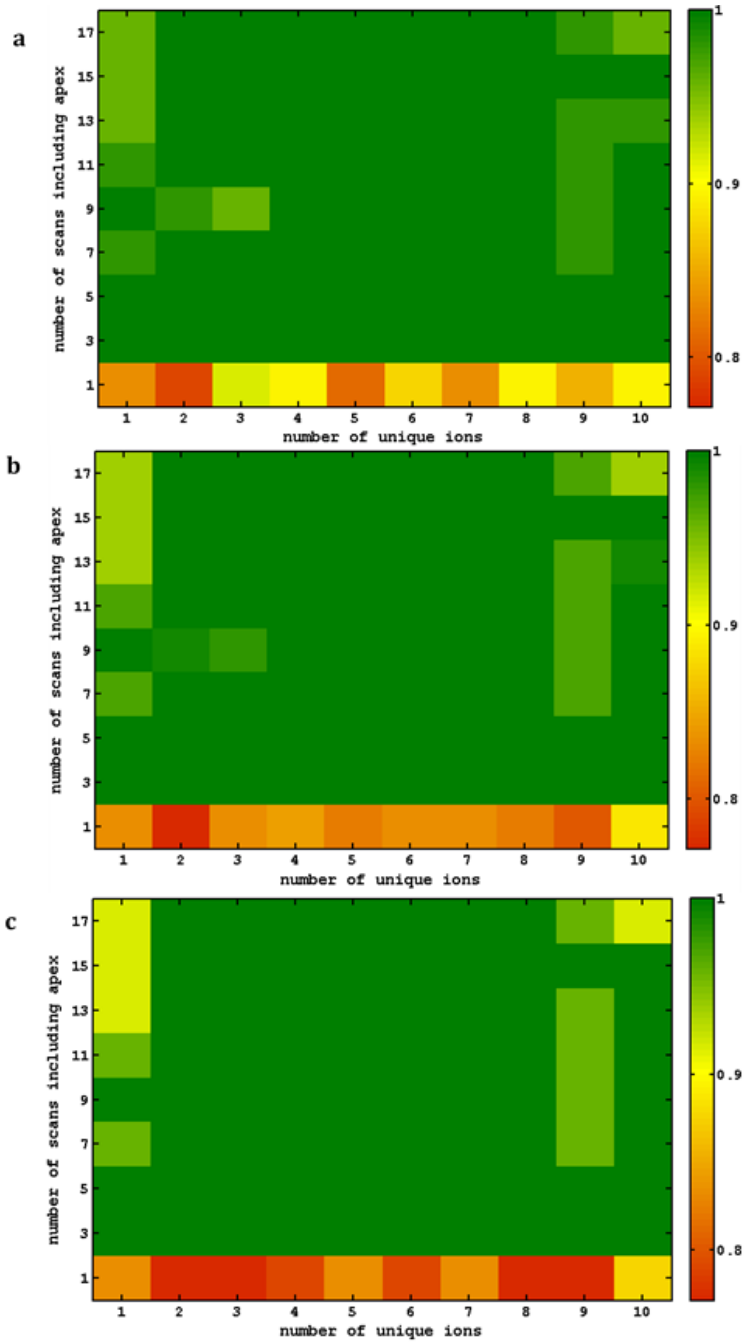


Fig. 2.4.2: Sensitivity (a), Specificity (b), and Accuracy (c) of UIF1D experiments. *m* and *w* were varied from 1 to 10 and 1 to 17, respectively. CR-FS was performed on the variables obtained after UIF application. The variables obtained after CR-FS were used to construct PLS-DA models. Sensitivity, specificity and accuracy were obtained from the PLS-DA model prediction of external validation set.

UIF1D$_{(2,\,5)}$ was chosen as the optimum to be compared to the benchmark pathway. The PLS-DA y-predicted plots for the three classes of samples when **UIF1D**$_{(2,\,5)}$ and the benchmark (no UIF1D) are shown in Fig. 2.4.3 a and b. These two results are comparable since they all demonstrate a model prediction sensitivity, specificity and accuracy of a 100%. However, the model presented in Fig. 2.4.3 b-1 to b-3 is likely a more robust model since the validation data for 87, 89 and 91 project further away from the class discrimination barrier (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. The number of variables retained for the the benchmark and **UIF1D**$_{(2,\,5)}$ were 3001 and 53, respectively.

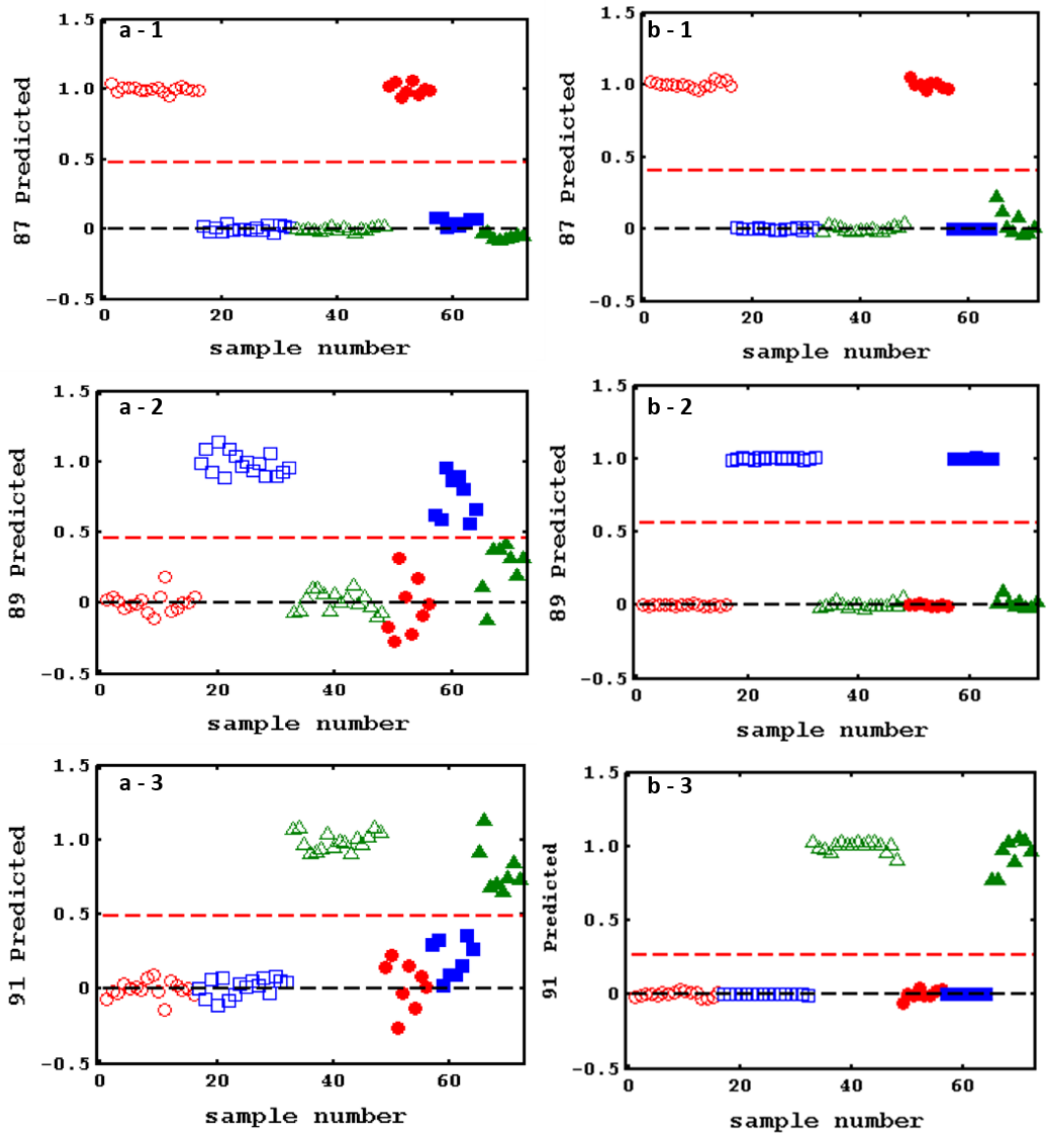


Fig. 2.4.3: PLS-DA y-predicted plot for benchmark pathway (a-1 to a-3) and optimum for UIF1D (**UIF1D**$_{(2,\,5)}$) (b-1 to b-3). Models were generated using features obtained after feature selection. The number of variables retained for the the benchmark and **UIF1D**$_{(2,\,5)}$ were 3001 and 53, respectively. 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.

The overall effect of applying $\mathbf{UIF}_{(2,5)}$ to a sample region of a chromatogram is shown in Fig. 2.4.4. The overall reduction in the number of candidate variables is obvious. The *m/z* dimension in Fig. 2.4.4b and 2.4.4c 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 Fig. 2.4.4, the other *um/z* is at a *m/z* value > 140.

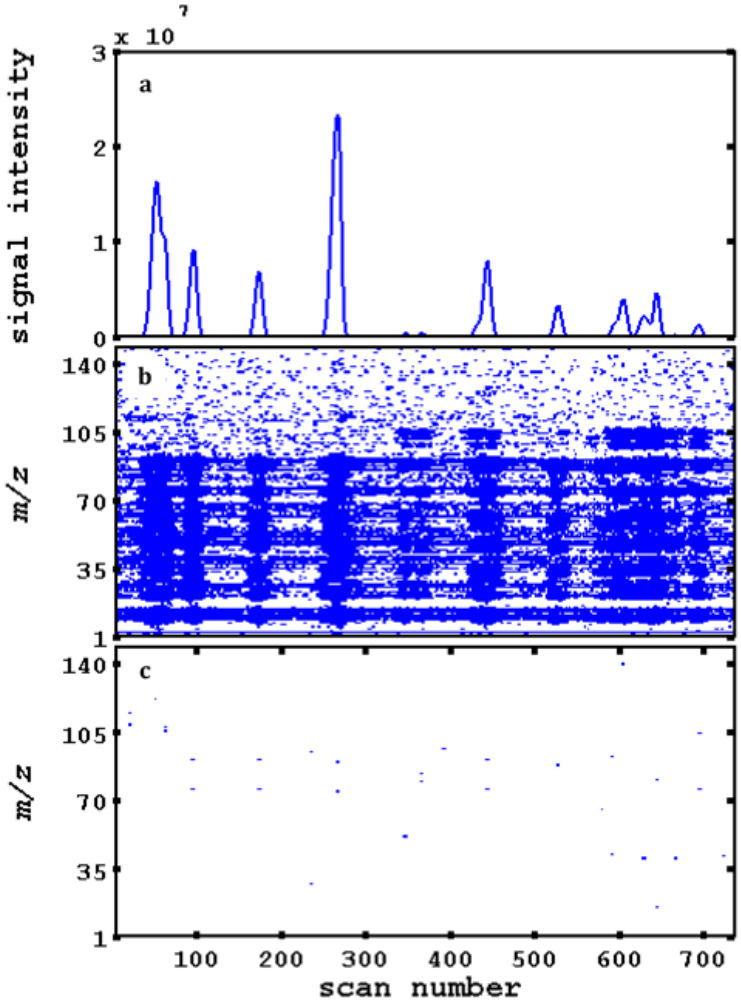


Fig. 2.4.4: Effect of $\mathbf{UIF1D}_{(2,5)}$ on an example segment of a chromatogram. (a) TIC trace, (b) unfiltered GC-MS data matrix, (c) data matrix after being filtered by $\mathbf{UIF1D}_{(2,5)}$.

Comparing the features selected with and without the application of the UIF1D, the features correspond largely to the same compounds (Fig 2.4.5). 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.

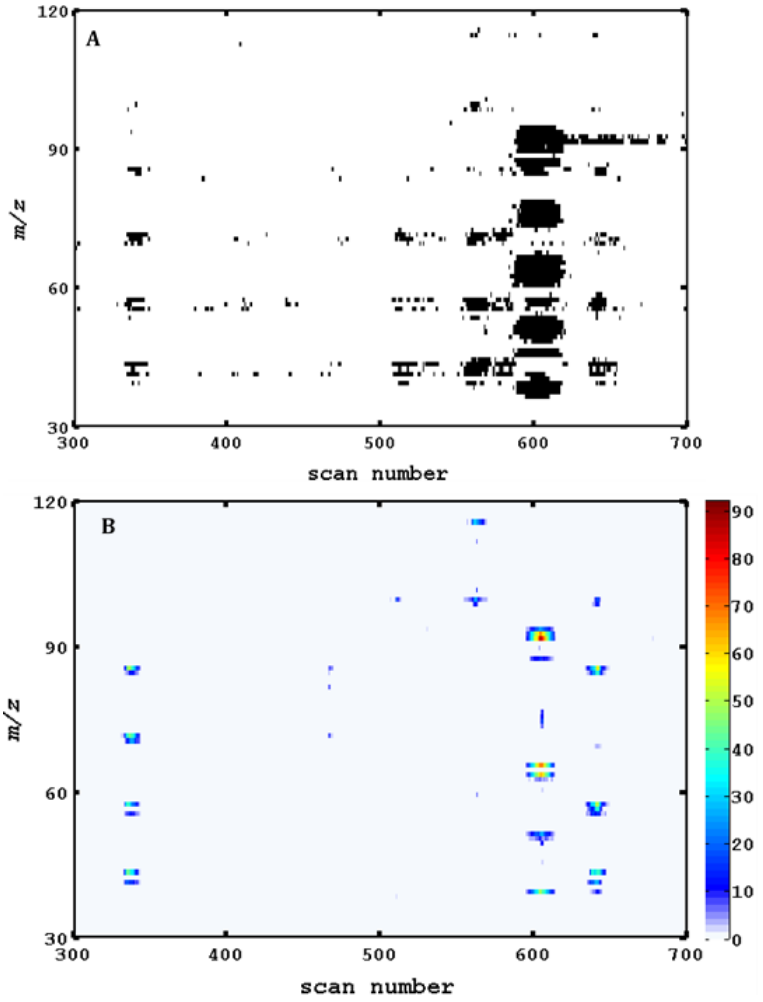


Fig. 2.4.5: Features selected by feature selection without(A) and with UIF1D(B).

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. UIF1D was also tested on a more challenging data set. The optimum UIF setting for this work (i.e., $\text{UIF1D}_{(2,\,5)}$), was applied to a data set comprising GC-MS chromatograms of casework fire debris samples from a previous study.[20] 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 Fig. 2.4.6.

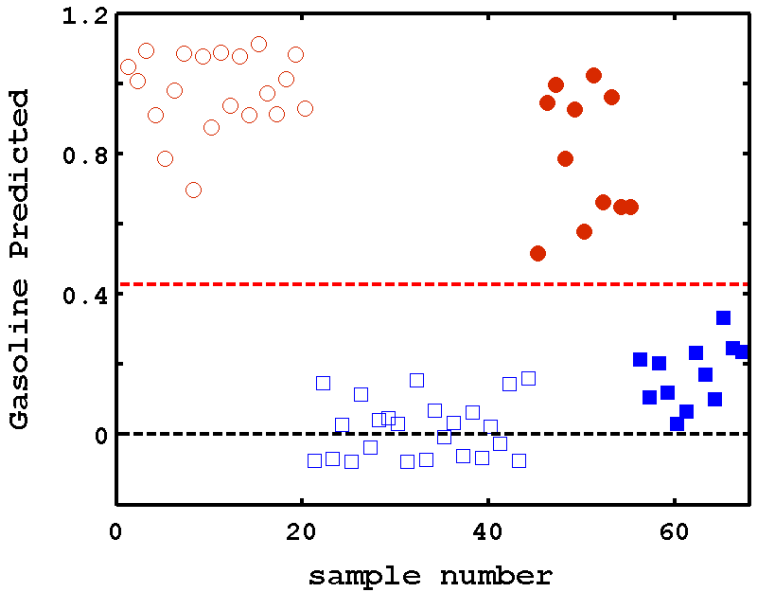


Fig. 2.4.6: PLS-DA y-predicted plots for predicting the presence or absence of gasoline in casework fire debris sample from a previous study.[20] Red circles indicate samples containing gasoline and blue squares indicate samples that do not contain gasoline, Hollow and filled markers are for training and validation sets, respectively.

Table 2.4.1 presents a comparison of the optimum benchmark and UIF1D conditions. Even though excellent model quality was achieved without the UIF1D, 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 UIF1D is applied is slightly longer than for the benchmark algorithm due to the additional computations applied by the UIF1D. However, the total number of candidate variables was reduced by two 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 UIF1D, 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 UIF1D. Results in Table 2.4.1 also show that without the use of the UIF, testing only the 500 top-ranked variables led to poorer overall model quality.

Table 2.4.1: Results of feature selection and model quality for optimum benchmark and UIF1D

Data unfolding		Feature Selection				Model Quality
Condition	**Time/sample(s)**	**Total**	**Checked**	**Passed**	**Time/min**	**Accuracy**
UIF1D$_{(2,\,5)}$	0.56 (0.02)*	0.01×10^{6}	500	53	9 (1)*	1.00
NO UIF	0.027 (0.002)*	1.67×10^{6}	500	116	8.4 (0.9)*	0.83
NO UIF	0.029 (0.004)*	1.67×10^{6}	30000	3001	370 (20)*	1.00

*mean and standard deviation at n = 5

The two-dimensional version of UIF (i.e., UIF2D), is a pre-filter for raw GC×GC-MS data. Like UIF1D, its implementation provides a data set that is orders of magnitude smaller than the raw GC×GC-MS data, while preserving the multivariate information in the data. For each identified peak in the chromatogram, UIF2D retains a specified number of unique *m/z* for a specified number of sub-peaks in the ^{1}D and scans/spectra in the ^{2}D. Thus the peak is represented by the signal closest to the apex which is the purest section of the peak.

Figure 2.4.7 shows a series of TIC plots from a representative chromatogram before and after UIF2D application. Eight (8) peaks are identified in this region, numbered 1 to 8. Subscripts indicate sub-peaks in order of decreasing prominence, with the base peak marked s1. Peaks 2, 4, 6-8 were trace components with only one detectable modulated peak, while peaks 1, 3, and 5 had two detectable sub-peaks each. In Fig. 2.3.4a ($\mathbf{UIF2D}_{(1,\ 10,\ 1)}$) one spectrum is retained for the base peak for each identified peak while all other information about the peak is eliminated. When the number of spectra/width in ^{2}D is increased to 15 (i.e, $\mathbf{UIF2D}_{(1,\ 10,\ 15)}$), the number of spectra included by UIF2D is increased accordingly. Fig. 2.3.4a. Fig. 2.3.4d depicts the results of UIF2D with m = 3, p = 10 and w =1. When present, sub-peaks from compounds present in multiple modulations (peaks 1, 3, and 5) are included. Fig. 2.3.4e shows the results of $\mathbf{UIF2D}_{(3,\ 10,\ 15)}$.

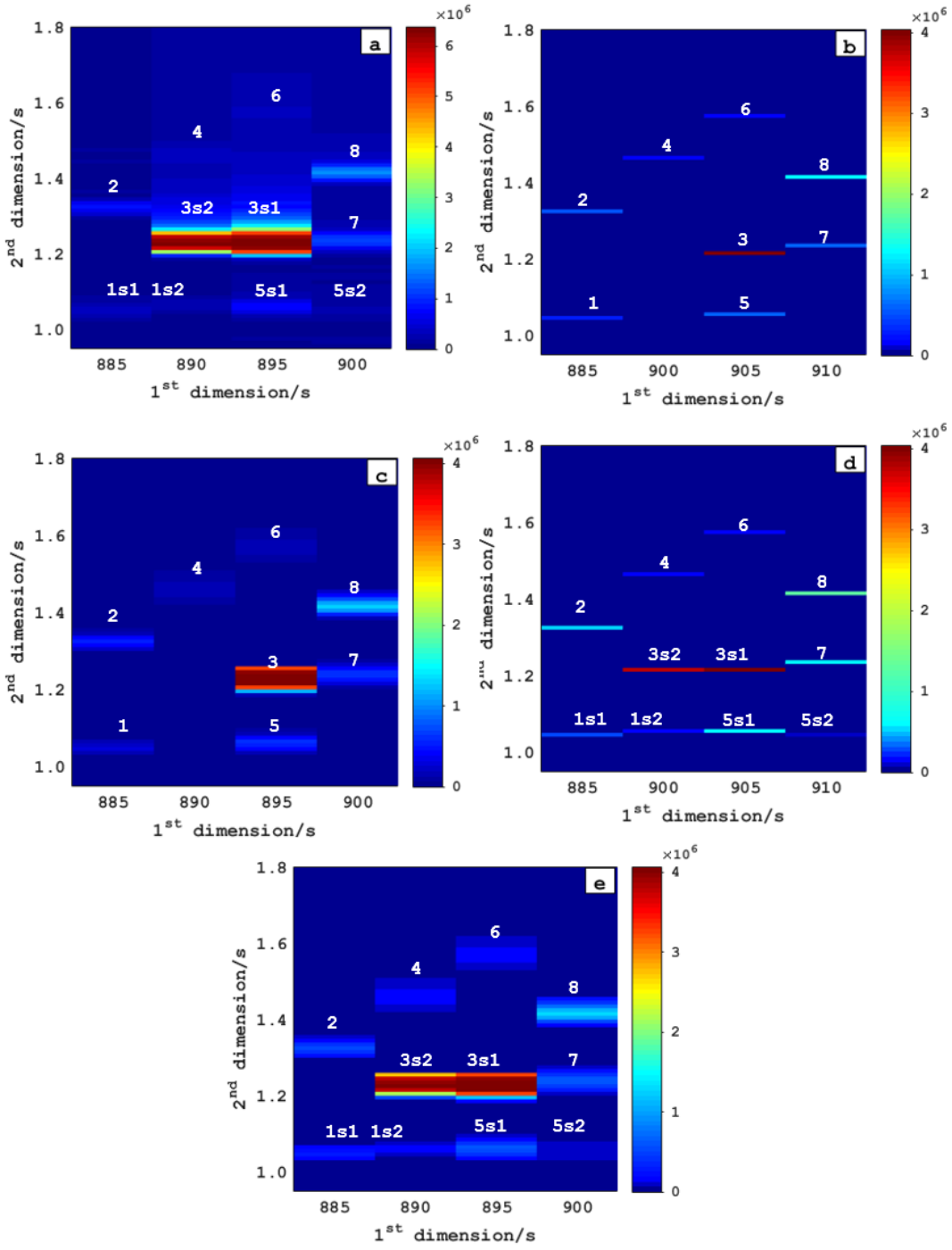


Fig. 2.4.7: Application of UIF2D to GC×GC-MS chromatogram. This shows the changes in the appearance of the chromatogram as the number of sub-peaks/modulations (p) and the number of scans/spectra (w) retained as these parameters are altered in the induction of UIF2D. a - raw data, b - $\mathbf{UIF2D}_{(1,\,10,\,1)}$, c - $\mathbf{UIF2D}_{(1,\,10,\,15)}$, d - $\mathbf{UIF2D}_{(3,\,10,\,1)}$ and e - $\mathbf{UIF2D}_{(3,\,10,\,15)}$.

Increasing m and w results in retention of a greater portion of the peak, and thus the appearance of the filtered data approaches that of the unfiltered data. Only *um/z* are retained in Fig. 2.3.4b-e. Even though the ${}^{2}t_r$ of peaks (1, 5) and (3, 7) are quite close, UIF2D is able to identify each of the peaks and assign them correctly. Application of UIF2D reduces the number of variables remaining for subsequent computations to between 6,518 ($\mathbf{UIF2D}_{(1,\ 1,\ 1)}$) and 418,086 ($\mathbf{UIF2D}_{(3,\ 10,\ 15)}$); representing reductions of 99.9 and 99.2%, respectively.

To investigate the effect of UIF2D data reduction on the feature selection process, a similar experimental workflow used for UIF1D(Fig. 2.4.1) was applied: the benchmark pathway (A) where the filter was not used and the UIF2D pathway (B). Performing feature selection on the entire raw data was not possible as this overwhelmed the computing power of the system used for this work. I resorted to reducing the number of variables by applying an ion count threshold of 200. Variables for which ion counts in all samples were less than this threshold were removed. The number of variables was thus reduced from 50.54×10^6 to 1.98×10^6 representing a 96.1 % reduction. The threshold of 200 was selected as that was the ion count threshold used in the peak finding algorithm. Increasing the ion count threshold would reduce the number of variables further; however, there is an increased risk of losing informative signals from analytes with weak signals as the threshold is raised.

Datasets were generated with UIF2D applied using a permutation of m = 1, 2, 3, p = 1, 2, 4, 10 and w = 1, 3, 5, 7, 9. The resulting data sets generated were subjected to feature selection. In the case of the dataset from the raw data, the number of top-ranked variables evaluated by CR-FS was varied from 500 to 20,000, while only 500 top-ranked variables were evaluated when UIF2D was applied.

Fig. 2.4.8 shows the model quality plots for the benchmark pathway. An increase in feature selection time commensurate with the number of top-ranked variables checked was observed, due to the added time required for checking each additional variable. A model prediction accuracy (red line) of 1.00 was achieved after 15,000 top-ranked variables were checked. This required 3.5 h. The number of variables included in this model was 1108, which represents 7.4 % of the total variables evaluated. This implies that the majority of the computer time was used to evaluate variables that were not helpful (over 90% of 3.5 h). When UIF2D was applied, the number of features retained was 199 (±47). Thus in addition to the shorter processing time, the computer is used more efficiently. Approximately 50% of the time is used to evaluate variables that are included in the model (50% of 6.8 min).

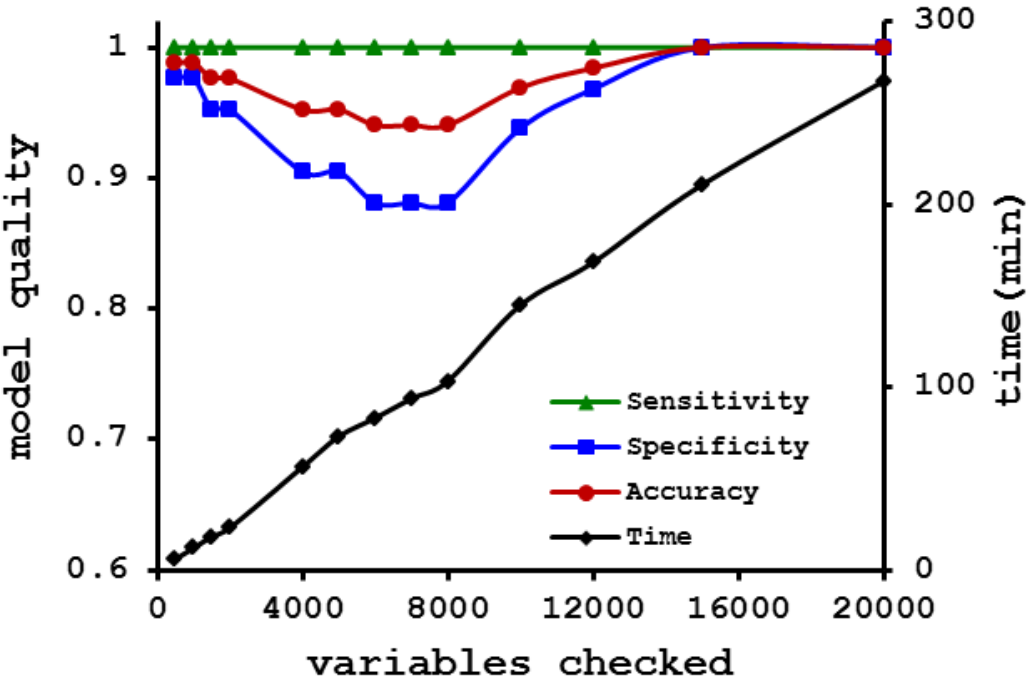


Fig. 2.4.8: Feature selection time and model quality plot for benchmark pathways for GC×GC-MS data. The number of variables evaluated during the application of CR-FS was increased until the PLS-DA prediction accuracy of the external validation set was 100%. The total number of variables evaluated until this was achieved was 15000.

Fig. 2.4.9 shows the model prediction accuracy for various values of m = 1, 2 and 3, p = 1, 2, 4 and 10 and w = 1, 3, 5, and 7 . In the region of p = 1, an increase in model accuracy occurs when w and/or p is increased. This is attributed to a modest increase in the number of variables retained per identified peak which contributes to the robustness of the model as more than one variable contributes information about a peak. Subtle variations in the order of *um/z* from one sample to the next, caused by differences in coelutions from one sample to another is the most likely reason why using only one *um/z* does not result in very robust models. When p is increased to 2, excellent model prediction accuracy was obtained for all values of m and w evaluated.

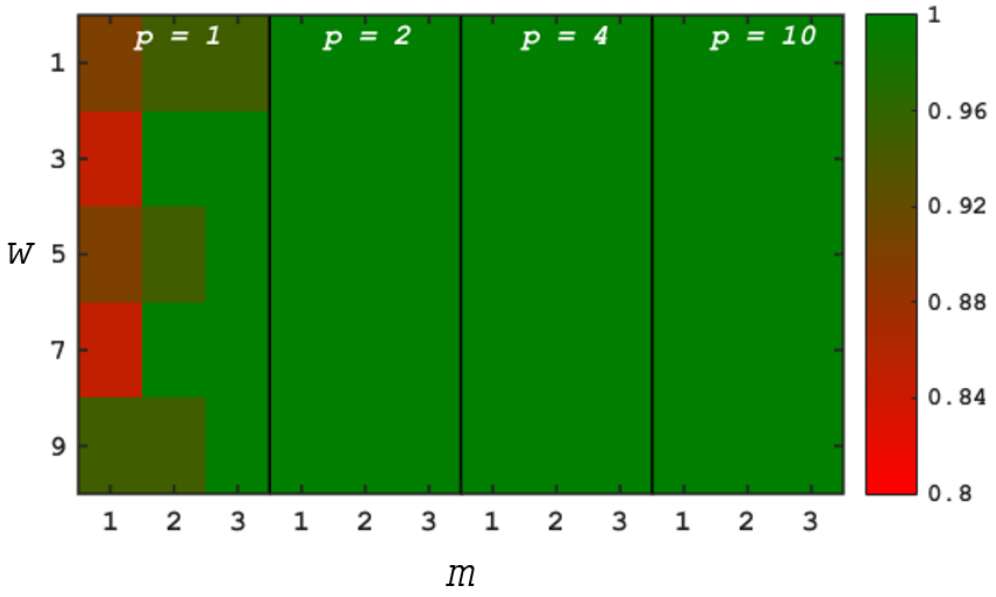


Fig. 2.4.9: PLSDA model prediction accuracy for UIF2D experiments for m = 1, 2 and 3, p = 1, 2, 4 and 10 and w = 1, 3, 5, and 7. CR-FS was performed on the variables obtained after UIF application. The variables obtained after CR-FS were used to construct PLS-DA models. Model accuracy were obtained from the PLS-DA model prediction of the external validation set.

Fig. 2.4.10 shows the class prediction plot for the PLS-DA model constructed using the 1108 variables that survived with the benchmark pathway on the left (Fig. 2.4.10 a-1 to a-3) and **UIF2D**$_{(1,\,2,\,5)}$ where 87 variables were selected on the right (Fig. 2.4.10 b-1 to b-3). Tea samples from India cluster closer to each other in **UIF2D**$_{(2,\,1,\,5)}$ relative to the optimum of the benchmark. This is due to the reduction in within-class variance, which is attributable to the extreme reduction and elimination of overly redundant variables in the data by the filter.

The location of the features that survived in the benchmark optimum and the UIF2D$_{(1,\,2,\,5)}$ are shown in Fig. 2.4.11a and Fig. 2.4.11b, respectively. Fig. 2.4.11 shows that all the features that were selected by UIF2D$_{(1,\,2,\,5)}$ in Fig. 2.4.11b are also included in the features that were selected in the benchmark pathway. This was confirmed by comparing the vector of features selected. Thus, in addition to the 87 variables that were passed, there are over a 1000 variables that are either irrelevant/noise or redundant to the models. Table 2.4.1 shows the comparison of UIF2D$_{(1,\,2,\,5)}$ and the benchmark optimum with respect to data unfolding and feature selection results.

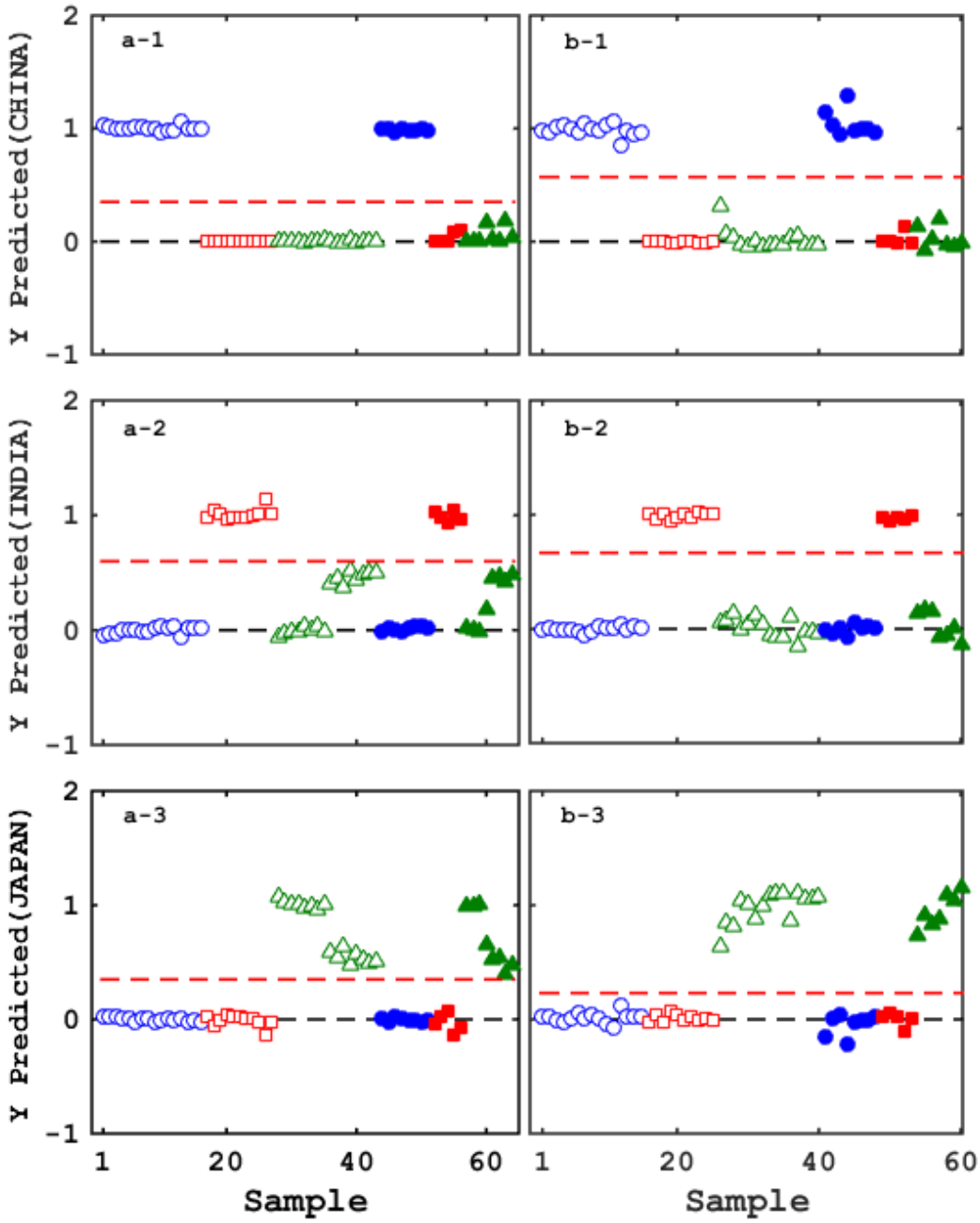


Fig. 2.4.10: PLS-DA y-prediction plot for optimum benchmark pathway and **UIF2D**$_{(1,\,2,\,5)}$. Models were generated using features obtained after feature selection. The number of variables retained for the the benchmark and **UIF2D**$_{(1,\,2,\,5)}$ were 1108 and 87, respectively. Blue circles, red squares and green triangles indicate teas samples from, China, India and Japan, respectively. Hollow markers indicate the training and optimization set while solid markers indicate the validation set.

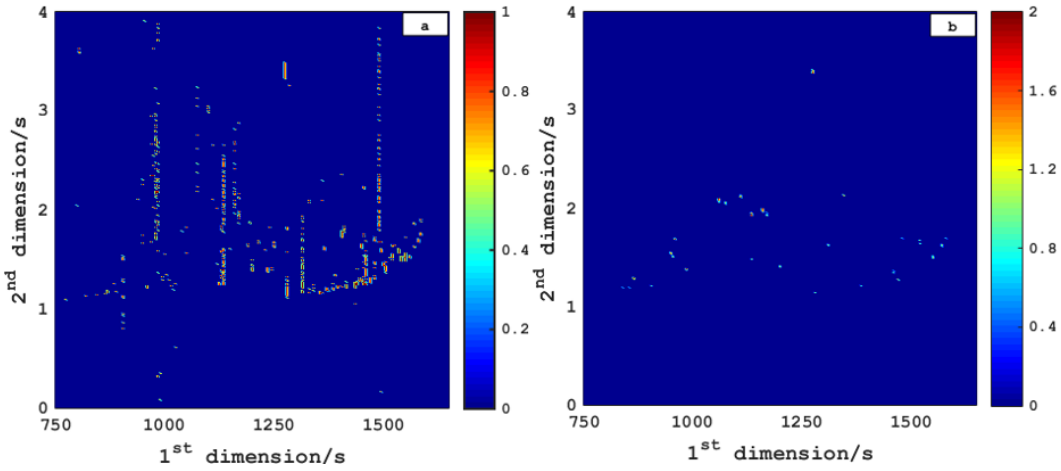


Fig. 2.4.11: Comparison of selected features for optimum benchmark pathway(a) and **UIF2D**$_{(1,\,2,\,5)}$ (b).

Table 2.4.2 shows the comparison of **UIF2D**$_{(1,\,2,\,5)}$ and the benchmark optimum with respect to data unfolding and feature selection results. As observed in the use of UIF1D, evaluating only the top 500 variables leads to a 100% model prediction accuracy when UIF2D is applied. However, this is not the case without the application of UIF. This indicates that the application of UIF2D eliminates noise as well as irrelevant variables, allowing the more useful and unique features of the peaks to be ranked higher.

Table 2.4.2: Results of feature selection and model quality for optimum benchmark and UIF2D

Data unfolding		Feature selection				Model Quality
Condition	Time/sample (s)	Total	Checked	Passed	Time/min	Accuracy
UIF2D$_{(1,\,2,\,5)}$	2.68 (0.02)[a]	0.02×10^6	500	87	6.8(0.5)[a]	1
NOUIF	27 (2)[a]	1.98×10^6	500	168	6.5(0.8)[a]	0.96
NOUIF	27 (2)[a]	1.98×10^6	15000	1108	229(13)[a]	1

[a] mean and standard deviation at n = 5

2.5 Conclusions

UIF is a novel feature reduction approach for preprocessing of multivariate data. The filter does not require *a priori* knowledge of the samples being analyzed. 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 and reduces the need to check a higher number of top-ranked variables.

Application of the UIF does not alter the fundamental chemical information in analytical data upon which models are ultimately based. With the increase in the use of 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 UIF1D and UIF2D were applied to GC-MS and GC×GC-MS data in this study, it can be adapted to other chromatographic data with a multivariate detector.

"Study without desire spoils the memory, and it retains nothing that it takes in"

Leonardo da Vinci

3

Comparison of Total Ion Spectra and Segmented Total Ion Spectra as Preprocessing Tools for Gas Chromatography-Mass Spectrometry Data for the Chemometric Analysis of Casework Fire Debris Samples

3.1 Introduction

Analysis of fire debris for the presence or absence of ignitable liquids is an essential step in the forensic investigation of fires. These analyses provide information that can aid fire investigators in ascertaining the causes of fires.

Adutwum L. A., Abel R. J. and Harynuk J. J., *J. Forensic Sci.* under review, reference number: JOFS-17-303

Where there is suspicion of malice, ascertaining the presence and identity of ILs in the debris is a critical step in the investigation. Ignitable liquids tend to be petroleum-based mixtures of volatile organic compounds (VOCs) which are easily collected via headspace sampling. The collected VOCs, comprised of pyrolysis products from the matrix as well as compounds from any ignitable liquids present, are subsequently analyzed by gas GC-MS.[147,162] Interpretation of the resulting data to determine the presence of ignitable liquids can be challenging due to the myriad of pyrolysis products generated from the various materials in the substrate and the highly variable nature of fire.[147,185,186] These frequently mask the presence of ignitable liquids, which are inherently complex and subject to alteration by the fire and fire suppression efforts.[147,185,186]

Conventional interpretation of fire debris data involves the visual inspection of various extracted ion chromatograms for the fingerprints/signatures of suspected ignitable liquids, examination of the mass spectrum for each questioned peak, and a peak-to-peak visual and mass spectral comparison with reference chromatograms.[147,185,187] This process relies heavily on the experience of the individual performing the interpretation and can introduce some degree of subjectivity. To reduce the risk of error, most labs employ a peer review system where two or more analysts will perform their own independent interpretations of the data to arrive at a consensus decision. All of these factors limit throughput and

increase the number of person-hours required for data interpretation.

The use of chemometric techniques (i.e., classification and pattern recognition) for identifying ignitable liquids in simulated[57,89,148,188,189] and casework[20,21] fire debris samples is an area of active research. The motivation is that a reliable chemometric approach could largely automate the interpretation process, saving time and reducing/eliminating subjective biases. A secondary benefit would be that the forensic scientists would be free to perform other tasks, thereby increasing sample throughput and decreasing sample turn-around time for multiple types of analyses throughout the laboratory.

One of the challenges for the application of chemometric methods to chromatographic data is the need for signal alignment, which aims to ensure as much as possible that the signal for each compound in the mixture is registered in the same location in the data matrix in every analysis. For gas chromatographic data, this is particularly challenging due to the narrowness of chromatographic peaks and the many factors that can introduce subtle shifts in retention times from one run to the next. While it is possible to use the entire raw data set, possibly with some feature selection step, to identify ignitable liquids in fire debris samples,[20,21] there are challenges that remain. The size of the data files may be a technical hurdle for some computers; however, the need to account for sample-to-sample shifts in retention times coupled with the highly unpredictable and inconsistent chromatographic patterns present in

the data present a more significant challenge. Due to a lack of reliable retention time markers, algorithms that require landmarks or anchors[73,103,104] struggle with fire debris chromatograms. Algorithms such as correlation optimized warping (COW),[98,100] interval correlated shifting (Icoshift),[101] and recursive alignment by fast Fourier transform (RAFFT)[102] are similarly defeated by the inherent lack of correlation between fire debris chromatograms. Finally, with any alignment algorithm, there is the potential for artefact generation, which could influence the resulting chemometric model.[190,191]

In an effort to avoid the need for strict chromatographic alignment, Sigman *et al.* introduced summed ion spectrum/total ion spectrum (TIS) as an alignment-free preprocessing step for raw GC-MS data.[142] A TIS is generated by summing the raw GC-MS data in the mass-to-charge ratio (m/z) dimension. The TIS data are analogous to those obtained from headspace - mass spectrometry (HS-MS), the direct injection of the headspace of a sample into a mass spectrometer as neither contains chromatographic information.[148] However, TIS has the advantage that the chromatographic separation limits the mass flux of material to the MS, unlike direct HS-MS. This minimizes the risk of saturating the MS detection system for mass channels of high abundance. Any saturation would result in skewed data and an increase in chemically meaningless variations that would degrade the results. TIS eliminates the need for chromatographic alignment, facilitating inter-laboratory transfer and utilization of data and chemometric models. In fact,

with TIS, successful comparison of data between laboratories is achievable even if they use vastly different separation conditions and/or completely different stationary phases. Furthermore, there is a huge reduction in the total number of variables and thus the redundancy in the data which could confuse the learning algorithm.[181] For example, a 20 min GC-MS experiment with a data rate of 20 Hz and a m/z range of 300 u would result in 7.2×10^6 variables per chromatogram. The TIS of these chromatograms would shrink each sample 99.99% to a mere 300 data points.

However, it may be advantageous to retain some chromatographic information in the data, which could aid in the identification of regions of the chromatogram from which important signals originate. This can aid in confirming or refuting the results of ambiguous samples, as the general separation region of features used for identifying different ignitable liquids are known.

Additionally, in the original implementation of TIS, there is no preprocessing step (i.e., any attempt to eliminate noise from the data).[142] In the case where an ignitable liquid is present at a very low concentration, the addition of background noise signals could mask or alter signals originating from the ignitable liquid. Furthermore, inclusion of empty chromatographic space will contribute meaningless ions to the TIS, which could negatively influence the chemometric model. I hypothesize that the elimination of inactive regions and noise from the raw GC-MS signal data prior to the

generation of TIS could improve the resulting models.

Herein I present the segmented total ion spectrum (STIS), which is similar to TIS, except it partially preserves the chromatographic separation while avoiding the need for strict chromatographic alignment. STIS is an alternative to both the TIS and the use of the entire raw data set. The performance of TIS and STIS for the classification of casework fire debris samples for the presence of gasoline is compared. I envisage that the partial preservation of the chromatographic separation by STIS will lead to improved chemometric models and facilitate the identification of regions of the chromatogram responsible for those signals important to the classification.

Further, I investigate the effect of a noise-reduction preprocessing step on both TIS and STIS. Regions of interest are identified via the application of a peak finding algorithm to each mass channel prior to TIS and STIS generation.[114] Using these sets of data, cluster resolution-feature selection (CR-FS) was used to identify relevant features.[18] Model prediction sensitivity, specificity and accuracy from partial least squares discriminant analysis (PLS-DA) were used to compare the results of TIS and STIS.

3.2 Experimental

Data used for this proof-of-concept work was part of a dataset from a previously published work. Details of the sample preparation and

analysis conditions are reported elsewhere.[20] Briefly, the data comprises of a series of GC-MS chromatograms (226 samples) obtained from the analysis of casework fire debris samples. Samples were analyzed by passive headspace sampling using activated carbon strips[192] which were subsequently eluted with CS_2 containing a perdeuterated alkane ladder of *n*-C_7 to *n*-C_{21} (odd alkanes only) and analyzed by GC-MS. Trained experts in the Trace Evidence Services Laboratory of the Royal Canadian Mounted Police National Centre for Forensic Services established the ground truth for each chromatogram based on forensic interpretation following established protocols. Sixty-four samples were confirmed to contain gasoline, while 162 were confirmed to not contain gasoline. Each sample chromatogram was imported into Matlab®2016a (The Mathworks, Natick, MA) using in-house written algorithms. TIS and STIS were generated from samples with and without the application of the noise reduction step. Each dataset was subjected to the feature selection and chemometric analysis. Feature selection was performed in Matlab®using in-house algorithms. The CR-FS algorithm described in Chapter 1 was used to identify relevant features. Chemometric models were constructed using PLS Toolbox 8.1.1 (Eigenvector Research Inc., Wenatchee, WA). All chemometric analyses were performed on a 64-bit Windows 7 Enterprise running on a core i7 - 4790K Intel processor and 32 GB RAM.

3.3 Generation of TIS and STIS

Each sample chromatogram was imported as a data matrix $\mathbf{D}^{m} \times^{n} (\mathbb{R})$, where m is the number of m/z and n is the number of spectra for the entire chromatogram. As a noise reduction step, a peak finding algorithm was applied to each mass channel. A null matrix of the same size as the sample chromatogram was generated. Using the peak apex locations and peak widths, the null matrix was modified by setting locations defined by the peak to a value of 1. This resulted in a binary masking matrix, $\mathbf{M}$. The Hadamard product of the mask, $\mathbf{M}$ and the original data matrix $\mathbf{D}$ (i.e. $\mathbf{M}$ o $\mathbf{D}$), yields a de-noised chromatogram.

TIS for the samples were generated as previously described by Sigman *et al.*[142,193] Two TIS datasets were generated for the samples, $\mathbf{X}_{\mathbf{TIS\text{-}A}}$ and $\mathbf{X}_{\mathbf{TIS\text{-}B}}$. In the case of $\mathbf{X}_{\mathbf{TIS\text{-}A}}$, the GC-MS data matrix for each sample was summed in m/z dimension as shown in Equation 3.1.

$$\mathbf{x}_{\mathbf{TIS\text{-}A}} = \sum_{i=1}^{I} \mathbf{D}_{(i,\, j)} \tag{3.1}$$

where $\mathbf{x}_{\mathbf{TIS\text{-}A}}$ (a vector of 271 elements) is the TIS of a sample, i is the m/z index and j is the scan number. For the generation of $\mathbf{x}_{\mathbf{TIS\text{-}B}}$, the sum is calculated over the Hadamard product of the sample chromatogram, $\mathbf{D}$, and the mask, $\mathbf{M}$, which was described earlier. The calculated TIS for every sample were compiled into $\mathbf{X}^{s \times v}_{\mathbf{TIS\text{-}A}}$ and $\mathbf{X}^{s \times v}_{\mathbf{TIS\text{-}B}}$, where s and v are the number of samples

(226) and variables (271), respectively.

Similarly, two STIS datasets were also generated, $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$. To generate $\mathbf{x}_{\text{STIS-A}}$, the perdeuterated alkane ladder compounds were located in the chromatogram as previously described Sinkov *et al.*[20] The chromatogram was subsequently segmented into nine (9) regions by the eight (8) deuterated *n*-alkanes as shown in Fig. 3.3.1. Each segment of the chromatogram was then summed in the *m/z* dimension, creating a set of nine (9) sequential TIS segments, each covering a portion of the chromatographic space. This results in a vector of 2,439 (i.e., 271×9) variables per sample. The overall data matrix $\mathbf{X}_{\text{STIS-A}}$ with 226 samples $\times$ 2,439 variables was obtained by collecting the calculated STIS vectors for each sample in the data set. As described earlier, in the case of $\mathbf{x}_{\text{STIS-B}}$, the STIS is calculated on the output of the Hadamard product of $\mathbf{D}$ and $\mathbf{M}$. $\mathbf{X}_{\text{STIS-B}}$ is also generated from the calculated $\boldsymbol{x}_{STIS-B}$ yielding a data matrix same size as $\mathbf{X}_{\text{STIS-A}}$.

3.4 Results and Discussion

In the chemometric analysis of GC-MS data, there are advantages to using the raw data, as opposed to a peak table that may have artefacts introduced from integration errors and/or the loss of small peaks/peak shoulders that are problematic for integration algorithms.[194] However, raw GC-MS data presents several challenges to conventional chemometric tools.

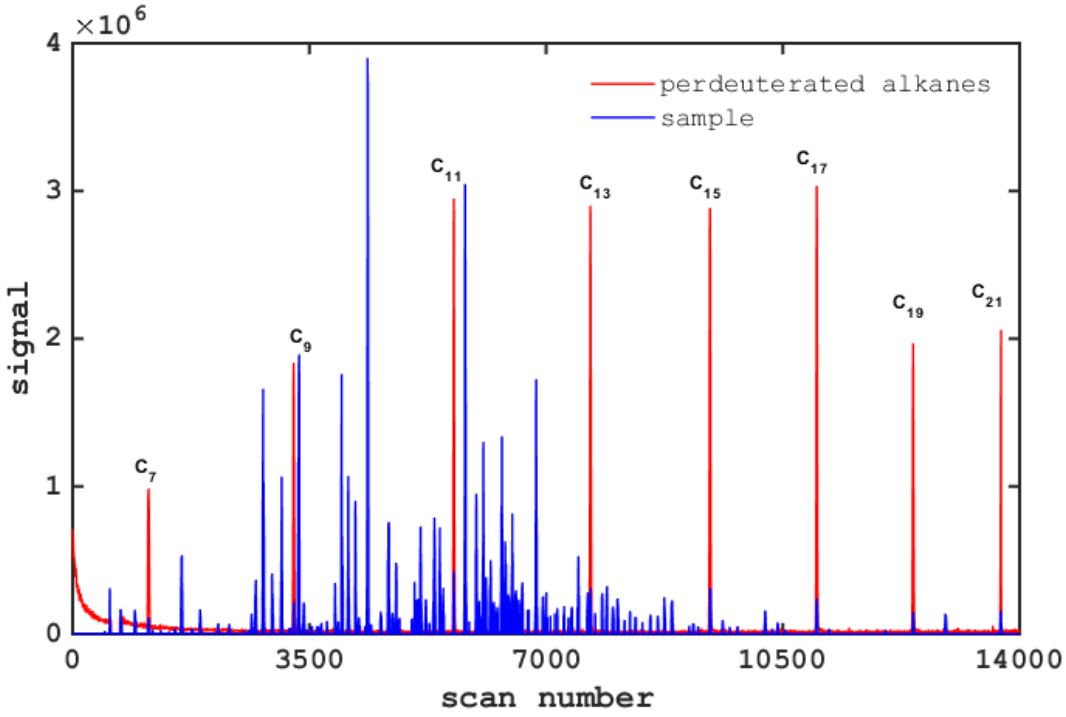


Fig. 3.3.1: A TIC of a typical fire debris sample chromatogram (blue) showing locations of perdeuterated *n*-alkane anchors (red). Perdeuterated alkanes were from *n*-C_7 to *n*-C_{21}

.

Chromatographic alignment of signals is one challenge, and several groups have come up with alignment algorithms, that are useful for some sample types.[20,98,100,101,194] The sheer number of variables in an entire raw GC-MS chromatogram necessitates data reduction strategies to remove those variables that are primarily noise (e.g., segments of empty baseline) or to reduce redundancy in the data.[15,140] This is critical, as without feature selection strategies, the noise in unimportant variables completely masks the information buried in the relatively few important variables.

The TIS algorithm, essentially presents the average signal for each mass channel over the course of the chromatogram. It is an aggressive, alignment-free data reduction strategy that also offers excellent intra-laboratory model portability; but at the cost of losing all chromatographic information. Hence, interpretation of the resulting chemometric models to identify specific sources of ions is a challenge, and overall model quality may suffer due to the loss of all separation information. Using a suite of clearly defined retention time marker, the chromatogram can be defined by segments. By generating TIS for each segment of the chromatogram (hence a segmented TIS, or STIS), the data size is reduced in addition to partial preservation of the retention information. The motivation being an improvement in model accuracy and easier interpretation of the chemical significance of the data.

Fig. 3.3.1, shown earlier, depicts a TIC of a fire debris sample (blue) showing the location of perdeuterated *n*-alkane anchors (red). For each

sample chromatogram, a TIS was generated by summing each ion along the time axis. A typical TIS for a fire debris sample is shown in Fig. 3.4.1. TIS reduces the number of variables in the entire chromatogram to the number of m/z monitored. In this data the number of variables for each sample chromatogram was reduced from 4.336×10^6 to a 271.

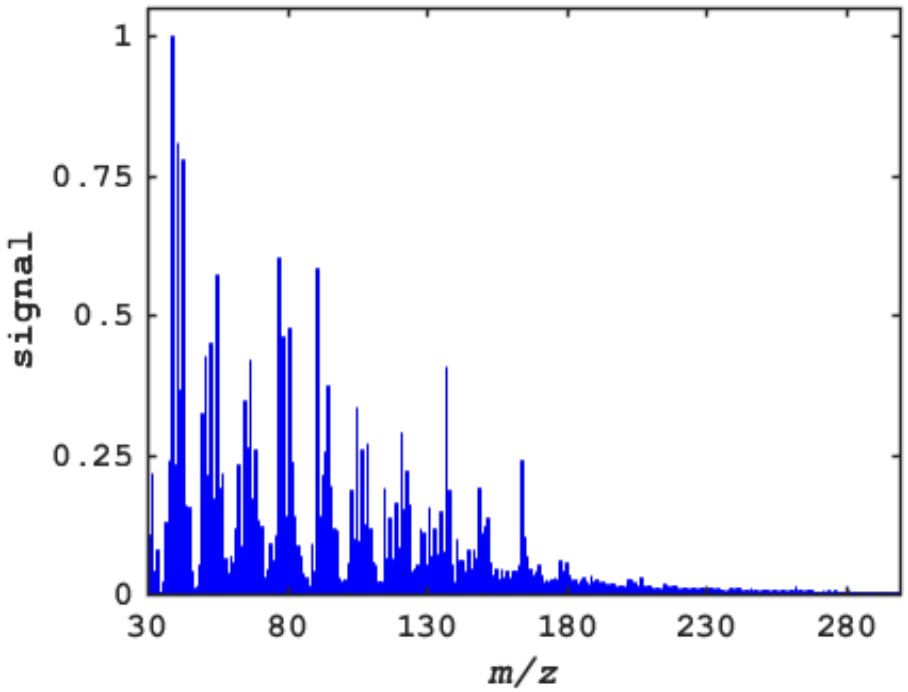


Fig. 3.4.1: A TIS of a typical fire debris sample. The TIS response is generated by summing the raw GC-MS data in the time dimension. The retention information in the data is eliminated.

The apexes of the eight (8) perdeuterated *n*-alkane signals divide the chromatogram into nine (9) sections. A TIS is generated for each section. Concatenation of the nine TIS segments yields the STIS for the sample with 2439 variables as shown in Figure 3.4.2.

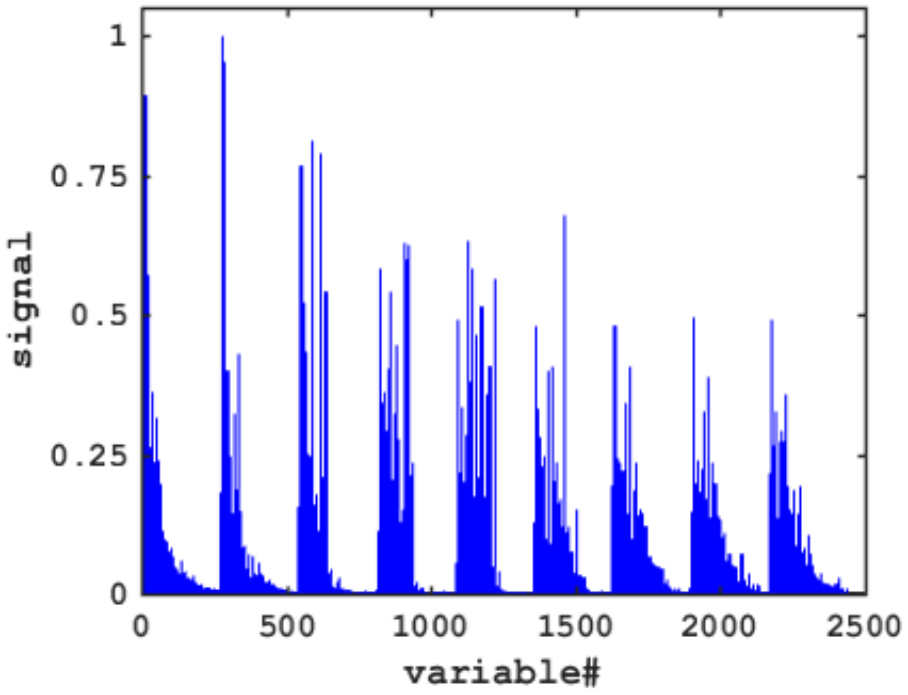


Fig. 3.4.2: An STIS of a typical fire debris sample. The STIS response is obtained by generating separate TIS for each of the nine segments and concantenating.

Each version of the dataset (i.e., $\mathbf{X}_{\text{TIS-A}}$, $\mathbf{X}_{\text{TIS-B}}$, $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$) was subjected to feature selection and the evaluation routine as described above. Each row was normalized to 1 and the columns scaled to unit variance. Each dataset was split, with two thirds randomly assigned to the training set (150 samples) and one third to the validation set (76 samples). Using a third of the training set data, variables were ranked based on their F-ratios.[18,167] The entire training set data was used for feature selection based on a cluster resolution-guided hybrid sequential backward-elimination/forward-selection algorithm.[18] The variables retained after feature selection were used to construct PLS-DA models. The models were evaluated using the prediction sensitivity, specificity and accuracy calculated for the validation set data. The

number of latent values (LVs) retained for the PLS-DA model was selected based on the minimum root mean square error for the cross validation. Feature selection and model validation steps were repeated ten times for each dataset. The data were randomly re-partitioned into training and validation groups for each iteration. The mean sensitivity, specificity and accuracy for $\mathbf{X}_{\text{TIS-A}}$, $\mathbf{X}_{\text{TIS-B}}$, $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$ across all iterations were then calculated. The results of these comparisons are shown in Fig. 3.4.3.

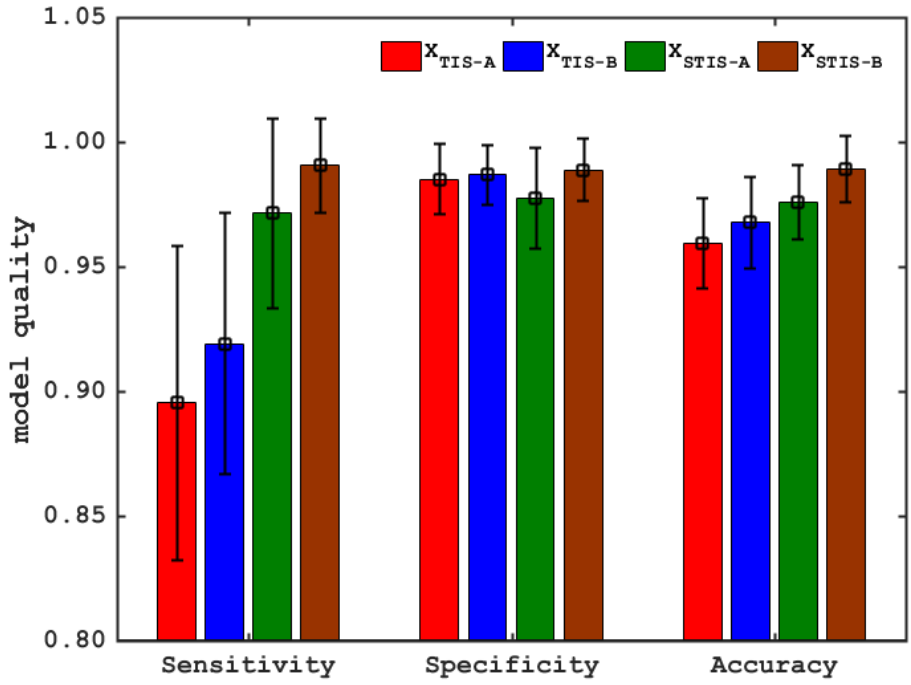


Fig. 3.4.3: Average model prediction sensitivity, specificity and accuracy for $\mathbf{X}_{\text{TIS-A}}$, $\mathbf{X}_{\text{TIS-B}}$, $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$ for the external validation set. The models were constructed using the variables selected after the implementation of CR-FS. These model quality and error bars indicate the averages and standard deviations for n = 10, respectively.

$\mathbf{X}_{\text{TIS-A}}$, $\mathbf{X}_{\text{TIS-B}}$, $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$ all perform quite well, with a prediction sensitivity, specificity and accuracy all above 0.85 (85%). With

the exception of prediction specificity, STIS provides better prediction sensitivity and accuracy relative to TIS. Even though both methods lead to data reduction, the extreme reduction in the data by TIS coupled with the loss of chromatographic information leads to a small number of variables that fails to accurately capture the variance in the casework fire debris samples. In general, the application of the noise elimination step in the form of eliminating regions where peaks are not detected tends to increase the overall model accuracy as $\mathbf{X_{TIS\text{-}B}} > \mathbf{X_{TIS\text{-}A}}$ and $\mathbf{X_{STIS\text{-}B}} > \mathbf{X_{STIS\text{-}A}}$. The order of performance for the comparison for sensitivity and accuracy were $\mathbf{X_{STIS\text{-}B}} > \mathbf{X_{STIS\text{-}A}} > \mathbf{X_{TIS\text{-}B}} > \mathbf{X_{TIS\text{-}A}}$ (Figure 3.4.3).

Since a new training and validation set were used for the ten permutations for feature selection, a different subset of features were retained each time for both TIS and STIS. To consolidate and compare all the features that survived for all iterations, the frequency of survival for the features were considered. Fig. 3.4.4. shows the feature survival frequency for TIS. Features which survived less than six permutations were eliminated. For $\mathbf{X_{TIS\text{-}A}}$ and $\mathbf{X_{TIS\text{-}B}}$ thirty-six and seventeen features, respectively, were above this threshold. This implies that the application of the noise reduction step led to approximately 53% reduction in the number of variables exceeding the threshold. TIS generation combines the noise as well as relevant information for any particular *m/z* into a single variable. Hence the presence of noise at any point in the chromatogram for a given mass channel will influence the

entire mass channel. For this reason, far more variables are included in the model during the training step to account for the higher level of noise.

The PLS-DA y-predicted plot for the $\mathbf{X_{TIS\text{-}A}}$ (with the 36 variables) and $\mathbf{X_{TIS\text{-}B}}$(with the 17 variables) are show in Fig. 3.4.5. Comparing Fig. 3.4.5a and 3.4.5b, the training set misclassification reduced from five to four when the noise reduction was applied. However, with the validation set the previous misclassification of two samples was corrected by the implementation of the noise reduction step. This demonstrates that the application of some noise reduction technique prior to the generation of TIS does improve the quality of the chemometric model.

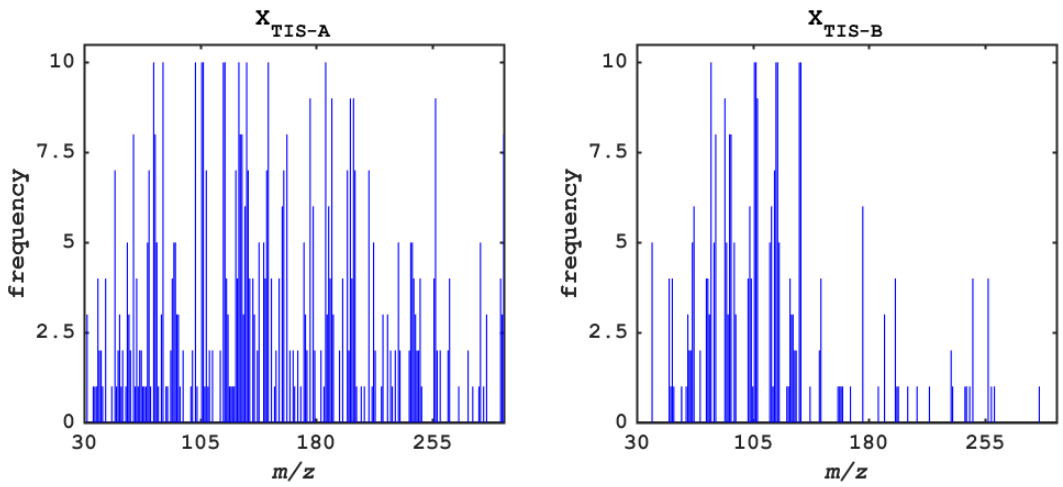
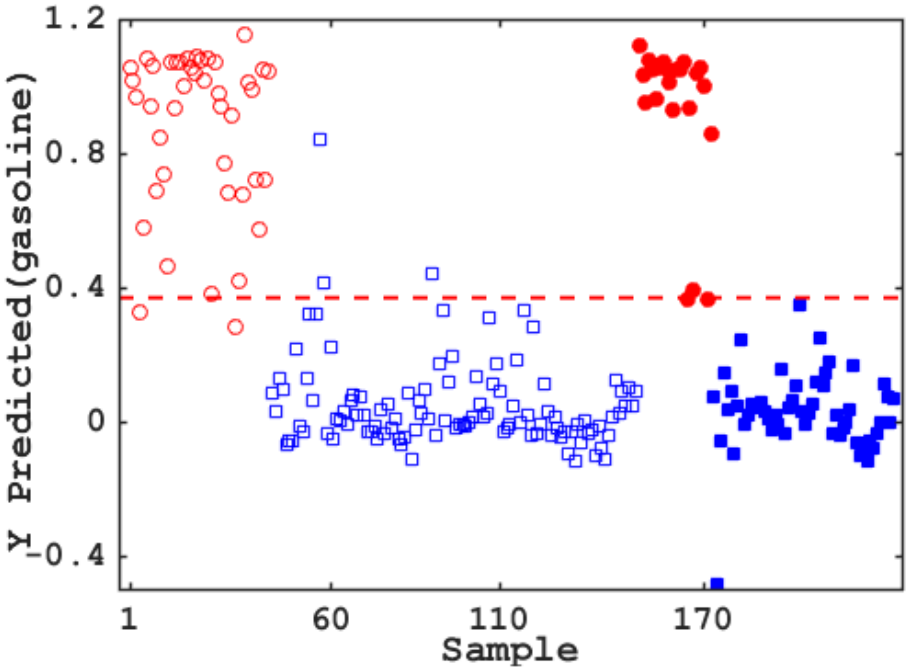
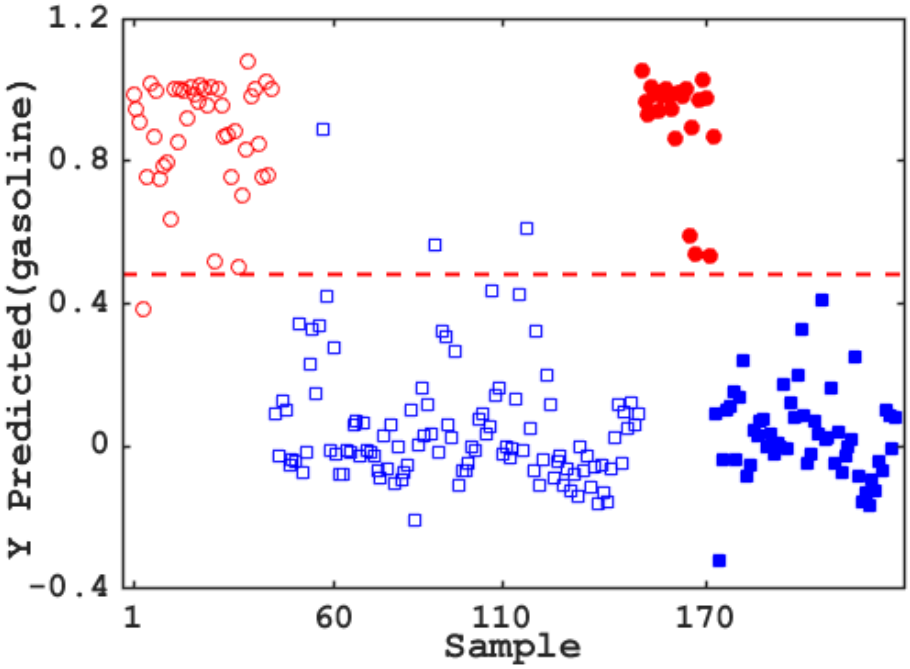


Fig. 3.4.4: Variable survival frequency for $\mathbf{X_{TIS\text{-}A}}$(a) and $\mathbf{X_{TIS\text{-}B}}$(b). Shows the number of times a variable survives in the ten CR-FS iterations.



(a) $\mathbf{X}_{\text{TIS-A}}$ (36 variables).



(b) $\mathbf{X}_{\text{TIS-B}}$ (17 variables).

Fig. 3.4.5: PLS-DA y-predicted (gasoline) for $\mathbf{X}_{\text{TIS-A}}$(a) and $\mathbf{X}_{\text{TIS-B}}$(b). The model was constructed using features that were retained in at least six of the ten iterations of CR-FS which were 36 and 17 for $\mathbf{X}_{\text{TIS-A}}$ and $\mathbf{X}_{\text{TIS-B}}$, respectively. Red circles and blue squares markers represents gasoline and no gasoline containing samples, respectively. Hollow markers and filled markers are for training and validation sets, respectively. The red line indicates the discrimination boundary.

The variable survival frequency plot for STIS is shown in Fig. 3.4.6. When a similar threshold was applied as in the TIS, the number of features retained were thirty four and thirty six for $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$, respectively.

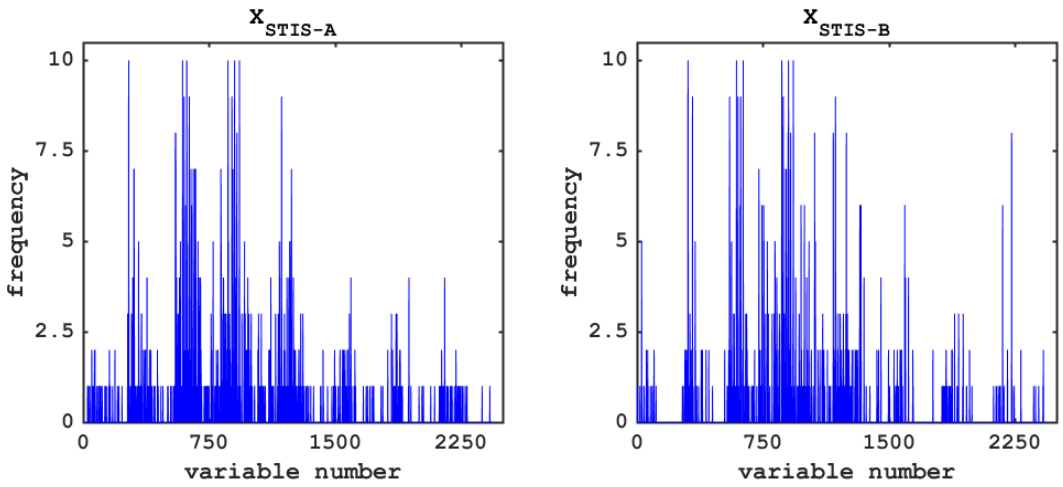
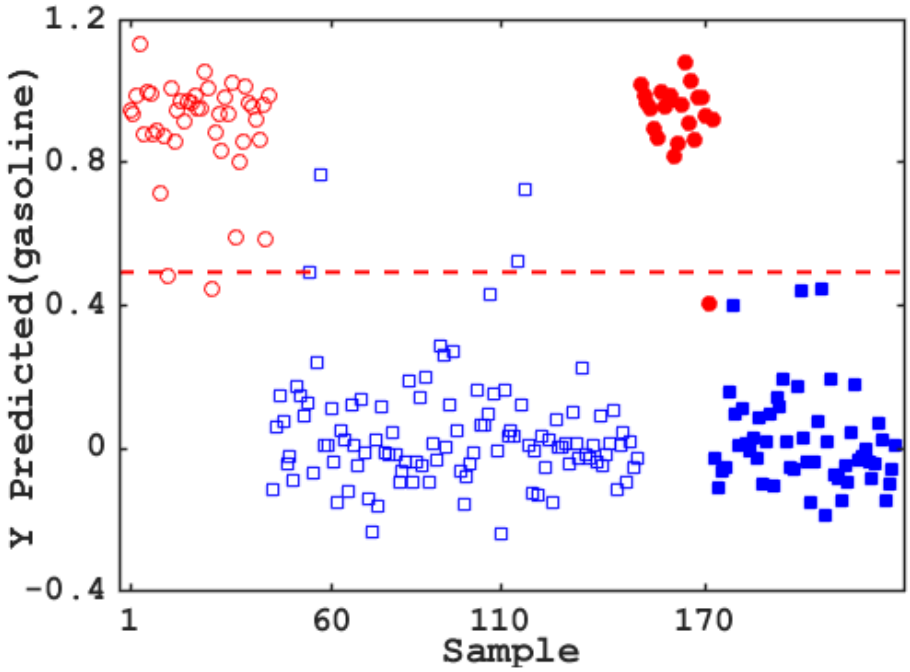
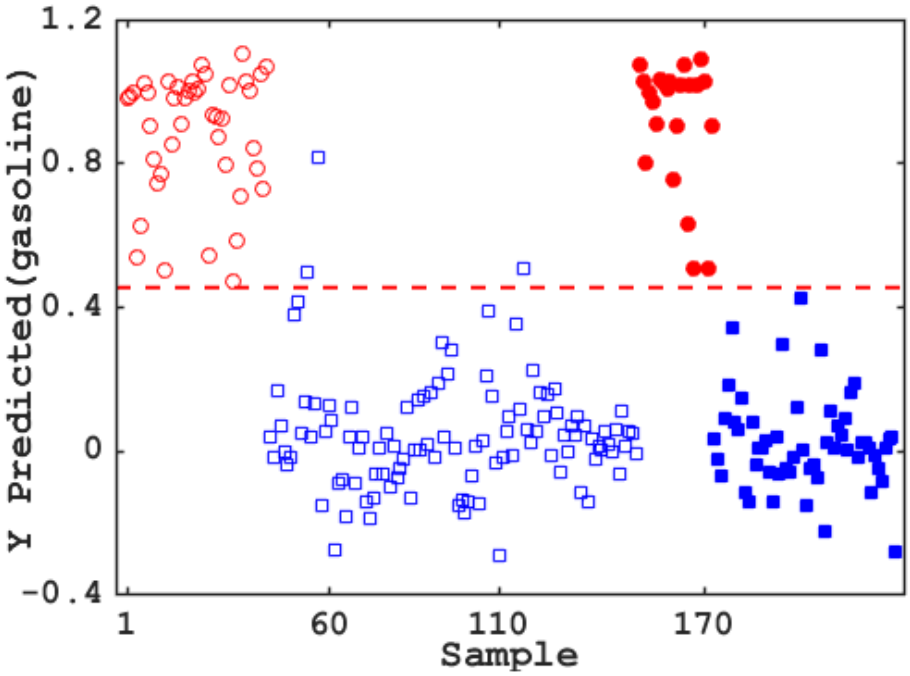


Fig. 3.4.6: Variable survival frequency for $\mathbf{X}_{\text{STIS-A}}$(a) and $\mathbf{X}_{\text{STIS-B}}$(b).

Comparing the difference in features that survived for TIS ($\mathbf{X}_{\text{TIS-A}}$ = 36 and $\mathbf{X}_{\text{TIS-B}}$ = 17) and that of STIS ($\mathbf{X}_{\text{STIS-A}}$ = 34 and $\mathbf{X}_{\text{STIS-B}}$ = 36, the impact of the noise reduction step was less for STIS than for TIS. In STIS, the signal for each ion is distributed across multiple segments of the chromatogram. Hence, each m/z is represented by nine (9) independent variables. For this reason, the overall impact of noise on a mass channel is reduced. PLS-DA y-predicted plots using the variables that were above the threshold are shown in Fig. 3.4.7. The number of misclassifications reduced from seven samples in $\mathbf{X}_{\text{STIS-A}}$ to three $\mathbf{X}_{\text{STIS-B}}$. This shows that STIS also benefits from noise reduction.



(a) $\mathbf{X}_{\text{STIS-A}}$ (34 variables).



(b) $\mathbf{X}_{\text{STIS-B}}$ (36 variables).

Fig. 3.4.7: PLS-DA y-predicted (gasoline) for $\mathbf{X}_{\text{STIS-A}}$(a) and $\mathbf{X}_{\text{STIS-B}}$(b). The model was constructed using features that were retained in at least six of the ten (10) iterations of CR-FS which were thirty-four and thirty-six for $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$, respectively. Red circles and blue squares markers represents gasoline and no gasoline containing samples, respectively. Hollow markers and filled markers are for training and validation sets, respectively. The red line indicates the discrimination boundary.

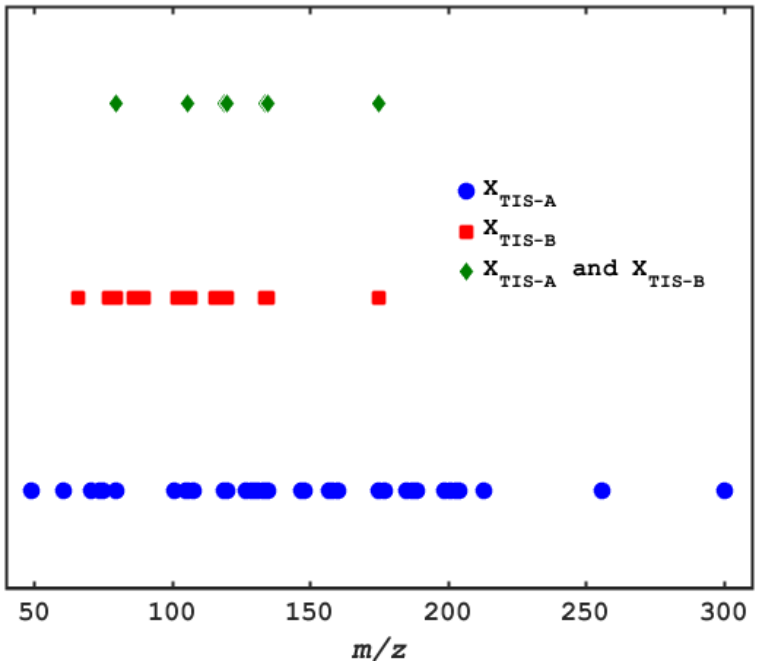


Fig. 3.4.8: A plot of features that survived more than five feature selection iterations ($\mathbf{X_{TIS\text{-}A}}$ - blue circles, $\mathbf{X_{TIS\text{-}B}}$-red squares and both $\mathbf{X_{TIS\text{-}A}}$ and $\mathbf{X_{TIS\text{-}B}}$-green diamonds).

Finally, we attempted to associate the selected variables (m/z)with the compounds from which they may have originated. This is important since the visual evaluation of fire debris chromatograms involves the identification of retention patterns. Fig. 3.4.8 shows a plot of selected feature for $\mathbf{X_{TIS\text{-}A}}$ (blue circles), $\mathbf{X_{TIS\text{-}B}}$ (red squares) and those common to both $\mathbf{X_{TIS\text{-}A}}$ and $\mathbf{X_{TIS\text{-}B}}$ (green diamonds). The eight ions common to both were m/z = 80, 105, 106, 119, 120, 134, 135, and 175. Amongst these, 105 and 106 can be attributed to the C_2 alkylbenzenes (*o*-, *m*- and *p*-xylene) and C_3 alkylbenzenes such as 3- and 4-ethyltoluene. m/z = 119, 120 and 134 and potentially its protonated product 135 could be attributed to C_4-alkyl benzenes such as 1, 2, 3, 5-tetramethylbenzene and 1, 2, 4, 5-tetramethylbenzene. The true sources

of these ions cannot be confirmed since there is no retention information.

With STIS, selected variables can be narrowed down to a specific segment straddled by *n*-perdeuterated alkanes. This permits interpretation of signals with retention index regions. The features that survived at least six iterations for $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$ are shown in the Appendix B. Twenty three of these features that survived in at least six iterations were common to both $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$ and are shown in Fig. 3.4.9.

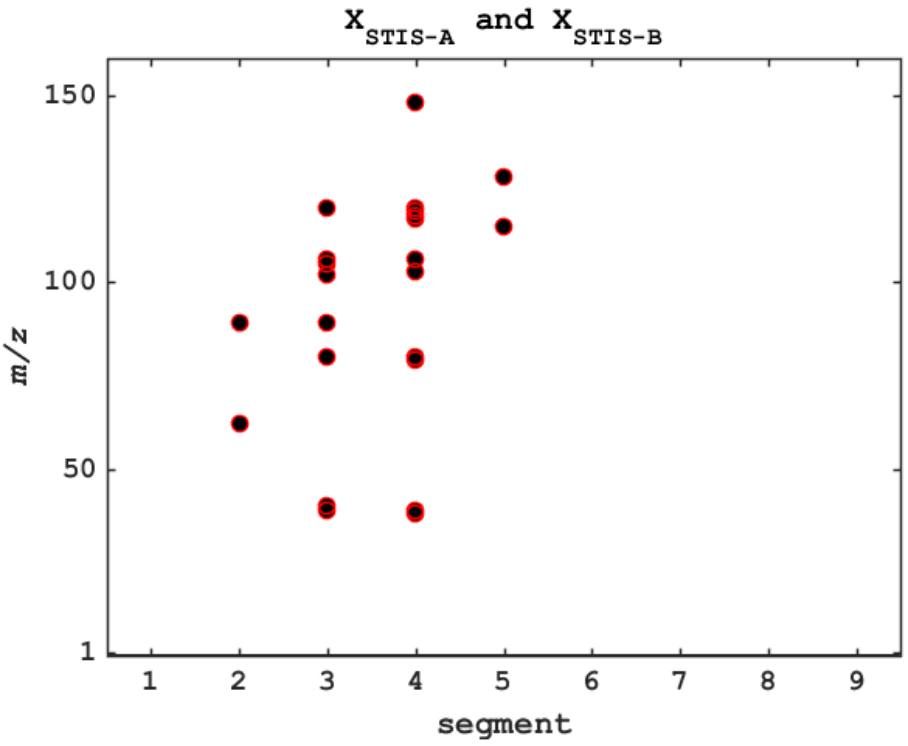


Fig. 3.4.9: A plot of features that survived in both $\mathbf{X}_{\text{STIS-A}}$ and $\mathbf{X}_{\text{STIS-B}}$. The segments indicates the region of the chromatogram from which these features originated from. The eight perdeuterated alkane anchors segments the chromatogram into nine segments.

In segment three, the ions retained included m/z = 105, 106 and 120. In TIS above these ions were attributed to the C_2 alkylbenzenes like xylenes, C_4 alkylbenzenes such as 3- or 4-ethyltoluene and 1, 3, 5-methylbenzene. Segment three is bordered by C_9 and C_{11}. Thus one can eliminate xylene as the potential source of these ions since the retention index of xylenes on HP-1MS columns is 878±1,[195] and thus elutes before C_9. This makes the 3-/4-ethyltoluene and 1, 3, 5-methylbenzene the more probable sources, as they have a retention index of 948±5.[195] In segment four, m/z = 39, 79, 103, 106, 118 and 120 were retained. This makes 1, 2, 3, 5-tetramethylbenzene, with retention index of 1096 - 1109 on HP-1MS columns[196] the most likely source of these ions since segment four is bordered by C_{11} and C_{13}. Thus, unlike TIS where only the m/z information is retained, the additional retention information aids model interpretation and ensuring the chemical logic of the models.

3.5 Conclusions

STIS was presented as an improved alignment-free preprocessing step for the chemometric analysis of GC-MS data from fire debris. STIS retains the advantages of TIS in addition to partial preservation of retention information. Although both STIS and TIS are capable of distinguishing gasoline-containing and gasoline-free fire debris samples, STIS outperforms TIS. Applying noise

reduction to TIS leads to a reduction in the number of features retained and reduces the number of misclassifications in validation data. After feature selection $\mathbf{X_{TIS\text{-}A}}$ and $\mathbf{X_{TIS\text{-}B}}$ retained thirty six and seventeen features, respectively. In STIS application of noise reduction did not reduce the number of features retained but reduced the number of misclassified samples. The partial preservation of retention information helps in identifying the potential source of selected variables.

"The significant problems we have cannot be solved at the same level of thinking with which we created them"

Albert Einstein

4

Estimation of Start and Stop Numbers for CR-FS Algorithm; An Empirical Approach using Null Distribution Analysis of Fisher Ratios

4.1 Introduction

Throughout this thesis, I have been relying on the CR-FS algorithm. Over the course of my research, I have made minor refinements to the code to improve performance. I have also made contributions towards the complete automation of CR-FS algorithm.

Adutwum L. A., de La Mata A. P., Bean H. D., Hill, E. J. and Harynuk J. J., *Anal. Bioanal. Chem.* under review, reference number: ABC-01259-2017

As has been previously explained in the earlier chapters, CR-FS is an automated hybrid feature selection algorithm based on a model parameter termed cluster resolution (CR).[18,19] CR measures the separations between clusters of samples in different classes in a reduced dimensionality space. Irrespective of the number of classes a single quality parameter bounded by 0 and 1 is computed to estimate the overall model quality. This makes it useful for the simultaneous optimization of multi class problems. The search of variables which improve the CR of a model is achieved through an initial sequential backward elimination (SBE) followed by a sequential forward selection (SFS). During the SBE, an initial population of highly ranked variables are used to generate a principal component analysis (PCA) model. The number of highly ranked variables used is termed the start number. SBE then proceeds, testing each of the initial variables in order from the lowest-ranked to the highest-ranked with CR being evaluated at each step. Variables whose elimination leads to a deterioration of the model are discarded. In the SFS, variables that were not tested in the SBE are evaluated in order of decreasing F-ratio (i.e., reducing relevance). Features whose inclusion do not improve the CR are eliminated. The total number of variables evaluated in both the SBE and SFS is termed the stop number. Algorithms based solely on SBE are known to be greedy while those based on SFS are likely to suffer a nesting problem.[16,122] Nesting implies that once a variable is added through SFS or removed through SBE, it is permanently included in (or excluded from)

the final model. This makes the start and stop number for SBE and SFS critical parameters for the CR-FS algorithm.

CR-FS has been successfully applied to various types of problems and has always led to improved model prediction sensitivity, specific and accuracy.[18–20,140,141,197] In all the previous studies, setting the start and stop numbers has been a matter of trial and error, relying on the experience of the user. This introduces subjectivity in the feature selection process, slows down the process, and prevents the true automation of CR-FS. When the total number of variables in the dataset is not very high (e.g, a few hundreds), the effect of each on CR can be evaluated. Evaluating all variables becomes prohibitive when several thousands or millions of variables exist for each sample, as is the case with raw chromatographic data. Previous experience with CR-FS demonstrates that the start and stop numbers can influence which features are retained. This makes it very important to find an objective and unbiased approach to the choice of the start and stop number for CR-FS.

When class labels are uncoupled from a data set and reassigned randomly, the F-ratios calculated from these misclassified datasets are termed null F-ratios. This is because they are generated under a distribution where the new class means vary from the true class means. Comparison of the distribution of null and true (obtained with correct class assignments) F-ratios can provide useful information in determining the limit below which a true F-ratio may not be as informative. This technique has been applied to the analysis

of comprehensive two-dimensional gas chromatographic data to reduce false positive rates.[177]

Overlapping coefficient (OVL), also known as Weitzman's measure, is a measure of similarity (and for that matter dissimilarity) between two probability distributions represented by continuous probability density functions (PDFs).[198,199] OVL was first used by Weitzman to determine the degree of overlap of income distributions between families in the United States.[198] Even though other similarity measures such as Matusita's and Morisita's are available, OVL is preferred due to its simplicity and naturalness.[199–201] OVL compares the density functions for two probability distributions and relates the similarity to the overlapping regions of the area under the two density functions.[198,199]

In this study, I drew inspiration from OVL to find the "dissimilarities" between two PDFs from the true and null F-ratios from a dataset. The degree of dissimilarity is used as a guide to determine the number of variables which have a higher probability of being from the true F-ratios. Parameters obtained from the two density functions are used to devise empirical equations to estimate the start and stop numbers for CR-FS. The empirical equations are then tested with real data with the aim of finding the optimal parameters. It is hoped that this will eliminate the subjective and trial and error approach to the implementation of CR-FS.

4.2 Theory

As indicated earlier, the start and stop numbers for CR-FS are critical parameters. The availability of an empirical equation to estimate these parameters will make CR-FS fully automated. In the context of categorizing variables into whether they are truly relevant or not, a comparison of the probability density functions of the true F-ratio and the null F-ratio can be very informative. A higher probability in the true distribution relative to the null distribution implies they are relatively more likely to come from the true distribution. Since the aim is to find F-ratios belonging to the true F-ratio, the focus is on the non-overlapping region of the two density functions.

4.2.1 True and Null F-ratios

The F-ratio ($\mathbf{f}$) is ratio of between class variance (σ_{bc}^2) to within class variance (σ_{wc}^2) and it is calculated as shown in Equations 4.1, 4.2 and 4.3.[167]

$$\mathbf{f} = \frac{\sigma_{bc}^2}{\sigma_{wc}^2} \tag{4.1}$$

$$\sigma_{bc}^2 = \frac{\sum n_i(\bar{x}_i - \bar{x})^2}{K-1} \tag{4.2}$$

$$\sigma_{wc}^2 = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2}{N-K} \tag{4.3}$$

where n_i is the number of variables in the i^{th} class/group, $\bar{x}_i$ is the mean of the *ith* class, $\bar{x}$ is the data mean, x_{ij} is the j^{th} observation in the i^{th} out of K class/group, and N is the sample size.

True F-ratios ($\mathbf{f_{TRUE}}$) are calculated from a subset of the training set data. Null F-ratios ($\mathbf{f_{NULL}}$) are calculated from the same subset after swapping the class assignments of approximately 10-15% of the data in each class.

4.2.2 Proposal of Empirical Equation for Estimating Start (n_{ST}) and Stop (n_{SP}) Numbers

$\mathbf{f_{TRUE}}$ and $\mathbf{f_{NULL}}$, are fitted to a selection of continuous PDFs. Density functions for continuous probability distributions were evaluated. These included Weibull, chi-square, inverse gaussian, log-normal, logistic, log-logistic, Gumbel and Frechet. To determine the optimum PDF, the Akaike Information Criteria (AIC) was used. After fitting the data, the AICs of all the PDFs were estimated. Based on the lowest AIC, the optimal PDF is selected.[202,203] If $f_{\mathbf{T}}$ and $f_{\mathbf{N}}$ are the optimal PDFs for $\mathbf{f_{TRUE}}$ and $\mathbf{f_{NULL}}$, respectively, then a simultaneous plot of these two PDFs provides very useful information (Fig. 4.2.1 - $f_{\mathbf{T}}$ - blue line, $f_{\mathbf{N}}$ - red line). The point of intersection (b) of the two PDFs can be determined by equating the two functions (i.e., $f_{\mathbf{T}}(b) = f_{\mathbf{N}}(b)$). The area under $f_{\mathbf{N}}(x)$, shaded in red, and $f_{\mathbf{T}}(x)$ where $x > b$ are estimated from Fig. 4.2.1. The area represented by the blue region is also determined. The area of the blue region relates to the cumulative density of

$f_{\mathbf{T}}(x)$, where $b \leq x \leq k$, and k is the maximum $\mathbf{f_{TRUE}}$.

From the analysis of Fig. 4.2.1 and based on my experience with CR-FS, I proposed two empirical equations (4.4 and 4.5) to estimate $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$. All but one parameter in the proposed equations can can be obtained from Fig. 4.2.1.

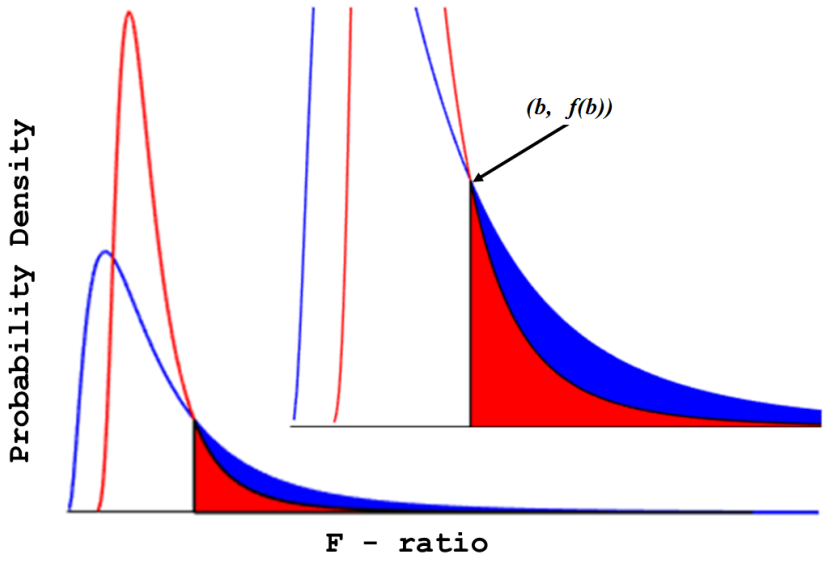


Fig. 4.2.1: Simultaneous plot of optimum PDFs for $\mathbf{f_{TRUE}}(f_{\mathbf{T}})$ and $\mathbf{f_{NULL}}(f_{\mathbf{N}})$. The optimum PDF is selected based on the AIC. Red region and blue region are area under the optimum PDF of $\mathbf{f_{NULL}}$, $(f_{\mathbf{N}})$ and $\mathbf{f_{TRUE}}$, $(f_{\mathbf{T}})$ with F-ratios $\geq b$ and b is the F-ratio value where $f_{\mathbf{T}} = f_{\mathbf{N}}$ intersect.

$$n_{\mathbf{ST}} = n_{\mathbf{SP}}^{d} + (b \times f_{\mathbf{T}}(b)) \tag{4.4}$$

$$n_{\mathbf{SP}} = \frac{\int_b^k f_{\mathbf{T}}(x)dx - \int_b^m f_{\mathbf{N}}(x)dx}{\int_b^k f_{\mathbf{T}}(x)dx} \times \mathbf{C} \tag{4.5}$$

where $n_{\mathbf{ST}}$ is the start number for SBE; $n_{\mathbf{SP}}$ is the stop number for SFS; $f_{\mathbf{T}}$ is the optimal PDF for $\mathbf{f_{TRUE}}$; $f_{\mathbf{N}}$ is the optimal PDF for $\mathbf{f_{NULL}}$; k is the maximum value of $\mathbf{f_{TRUE}}$; m is the maximum $\mathbf{f_{NULL}}$; $\mathbf{C}$ is the number of variables in $\mathbf{f_{NULL}}$ > b; and d is a constant.

4.3 Chemometric Analysis

4.3.1 Datasets

Five different data sets were used for this study. Dataset 1 (bac - bacteria) was obtained from the GC×GC-TOF MS analysis of the volatilome (i.e., volatile metabolites) of a suite of bacterial samples. The data consisted of 63 samples and 1673 variables to be classified as Type 1 vs. Type 2, having 35 and 28 samples, respectively. Dataset 2 (ucp-unwashed cotton polyester) was obtained from GC×GC-TOF MS analysis of volatile compounds extracted from worn cotton and polyester fabrics which have not been washed. The data consisted of 80 samples and 2781 variables and was to be classified as unwashed cotton vs. unwashed polyester. Dataset 3 (wcp-washed cotton polyester) was obtained from GC×GC-TOF MS analysis of volatile compounds from worn cotton and polyester fabrics after they have been washed. This data consisted of 80 samples and 2781 variables and was to be classified as unwashed cotton vs. unwashed polyester. Dataset 4 (coff - coffee) was a peak table obtained from the LC-MS analysis of coffee which consisted

of 78 samples and 701 variables. The coffee data was to be classified as Arabica (Ara.) vs. a mixture of Arabica and Robusta (Ara. + Rob.). Dataset 5 (cvp-cotton vs. polyester) was obtained from GC×GC-TOF MS analysis of volatile compounds obtained from worn fabric and was made up of 168 samples and 2781 variables to be classified as cotton vs. polyester. Detailed experimental conditions about datasets 2, 3 and 5 can be obtained from an earlier published work.[197] Datasets 1 and 4 were obtained from collaborators and were used as received.[204] Datasets 1, 2 and 3 were used to find the optimum value for the constant, *d*, in Equation 4.4. Using the optimum value of *d*, the validity of the approach was tested with datasets 3 and 4. Data importation and all computations were performed in Matlab® 2016b using in-house written algorithms. Chemometric models were constructed using PLS Toolbox 8.2.1 (Eigenvector Research Inc., Wenatchee, WA). All chemometric analyses were performed on 64-bit Windows 7 Enterprise running on a core i7 - 4790K Intel processor and 32 GB RAM.

4.3.2 Estimation of the constant *d* and n_{ST}

Each of the three datasets (i.e., 1, 2 and 3) were split into two-thirds for training and one-third for external validation sets. Using half of the training set data, F-ratio analysis was performed as previously described. n_{ST} and n_{SP} were estimated from the optimum PDF. n_{ST} is determined for a set of *d* values, as shown in Equation 4.4, such that $0.05 \leq d \leq 0.95$. Using these n_{ST} and n_{SP}

values, CR-FS was implemented on the entire training set data. This step was repeated ten times. During each iteration, a different subset of the training set data was used for F-ratio analysis and model optimization. Variables that were selected in at least six iterations were used for model evaluation. PCA and PLS-DA models were constructed with the training set data using only the selected variables. The validation set data was projected into the PCA model and the validation set CR ($cr_{\mathbf{max}}$) determined. The PLS-DA model prediction accuracy of the validation set was also determined. The product of the validation set CR ($cr_{\mathbf{max}}$) and PLS-DA prediction accuracy (acc) was used as the objective parameter in determining the best value for d.

4.4 Results and Discussion

CR-FS is a hybrid (filter and wrapper) feature selection algorithm which has been useful for improving classification accuracies of chemometric models. The two main parameters required by the algorithm are the $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$, for the SBE and SFS, respectively. The lack of a guidance as to the choice of these parameters introduces subjectivity and increases the feature selection time due to the trial-and-error nature of the optimization of these parameters. The aim of this study was to devise an empirical approach to the determination of $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$. This would eliminate subjectivity and allow for the true automation of the entire feature selection process.

The $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$ in an optimization with CR-FS varies with the data set. It was also observed that the probability density of the F-ratio also varies with the data set. Hence, it is possible to generalize the $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$ by connecting it to the PDFs of the F-ratios. Comparison of the PDFs obtained from $\mathbf{f_{TRUE}}$ and $\mathbf{f_{NULL}}$ was made using concept of OVL to guide the proposal of two empirical equations for the determination of $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$ (Equations 4.4 and 4.5).

The optimum density function, i.e, $f_{\mathbf{T}}$ and $f_{\mathbf{N}}$, varies with the dataset. Hence to determine the optimum PDF for $\mathbf{f_{TRUE}}$ and $\mathbf{f_{NULL}}$ for a dataset, several continuous density functions were tested. Amongst the distributions evaluated were Weibull, chi-square, inverse gaussian, log-normal, logistic, log-logistic, Gumbel and Frechet. Since the optimum PDF is unknown, determination of the AIC provided a means to evaluate the PDFs. AIC is a measure of relative quality of statistical models used to fit the same data.[202,203,205–207] The use of AIC to determine the optimum PDF eliminates the risks associated with overfitting or underfitting the data. The two PDFs i.e, $f_{\mathbf{T}}$ and $f_{\mathbf{N}}$, do not necessarily have to be the same. Thus irrespective of the density functions, a figure similar to Figure 4.2.1 results.

All but one parameter (i.e., d), in the empirical equations can be obtained from the analysis of the density functions obtained from $\mathbf{f_{TRUE}}$ and $\mathbf{f_{NULL}}$. It can be deduced from Equation 4.4 that a lower d value yields a smaller $n_{\mathbf{ST}}$ and the feature selection is dominated by SFS. A higher d value on the other hand yields a higher $n_{\mathbf{SP}}$ which makes CR-FS SBE-dominated. Three of the

datasets were used to determine the optimum value of d in Equation 4.4. To simultaneously compare the results of PCA and PLS-DA models, the product of the validation from PCA and prediction accuracy of the PLS-DA (i.e., $cr_{\mathbf{max}} \times$ acc), was used as the objective model quality parameter. A plot of the results is shown in Fig. 4.4.1a and 4.4.1b.

From Fig. 4.4.1a, lower d values seem to lead to better models; however, Fig. 4.4.1b shows the standard deviation of the $cr_{\mathbf{max}} \times$ acc is higher at lower d values. During each iteration, a different subset of the training data was used for training and optimization. If $n_{\mathbf{ST}}$ is too low, retained features tend to overfit the model to that specific subset. Hence when applied to the external validation set, high variability in prediction accuracies occurs. The standard deviation decreases as the $n_{\mathbf{ST}}$ is increased (Fig. 4.4.1b). As d increases beyond 0.65, the model prediction capability for all datasets starts to deteriorate. This is because at high d values, $n_{\mathbf{ST}}$ tends to be high (Equation 4.4). CR-FS performed with $n_{\mathbf{ST}}$ are dominated by SBE. Since SBE is greedy, several variables that may not be highly relevant end up in the model and lead to poor prediction accuracies. Fig. 4.4.2 shows a z-score (mean/σ) plot which shows a region of good model prediction accuracies with lower standard deviations. The region for d such that $0.48 \leq d \leq 0.57$, tend to have a higher model prediction quality with an accompanied lower standard deviation.

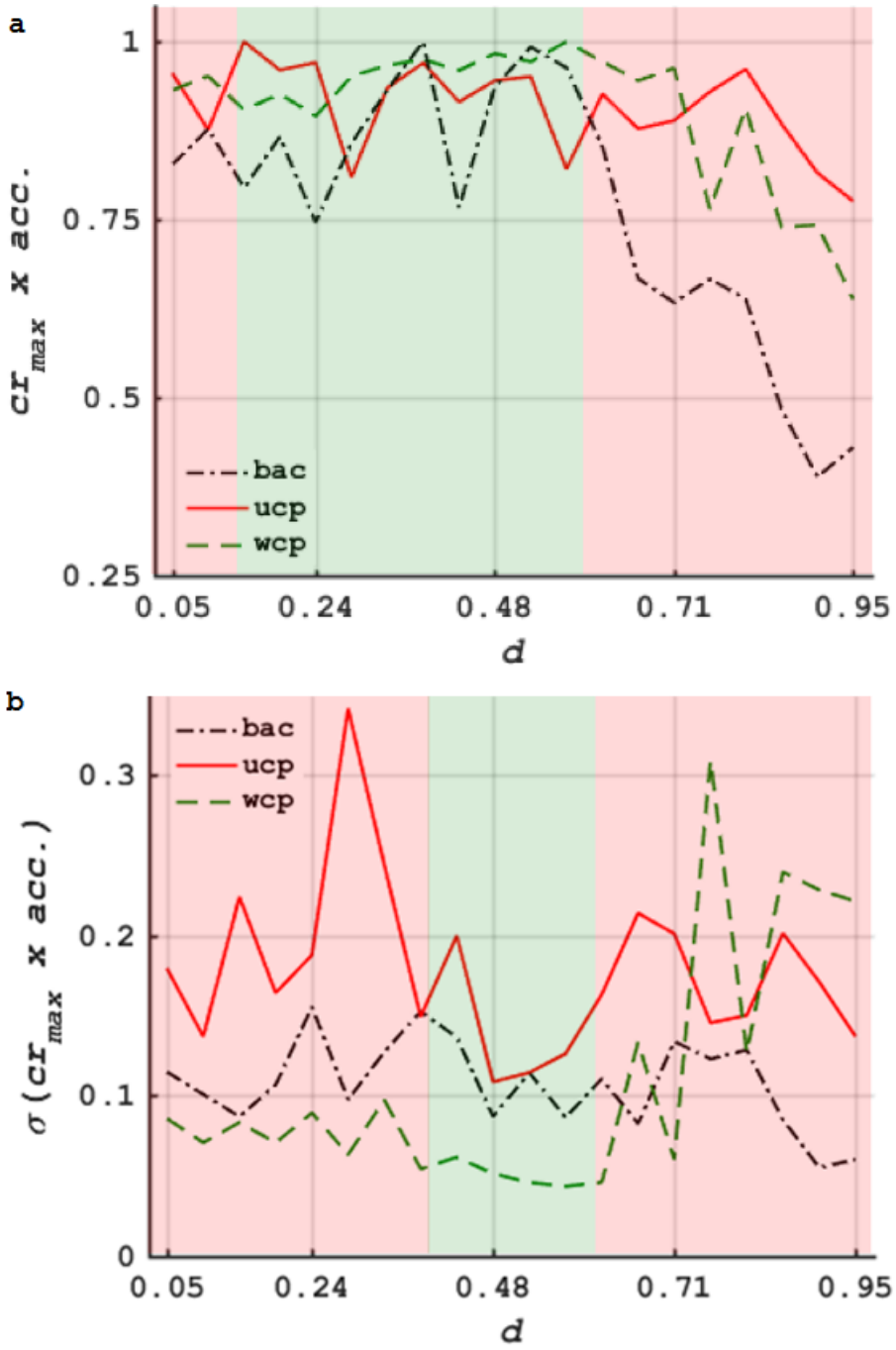


Fig. 4.4.1: A plot of overall model quality (a) and standard deviation (b) as a function of d ($0.05 \leq d \leq 0.95$). Model quality $cr_{\mathbf{max}} \times$acc vs. d. Variation in d influences the $n_{\mathbf{ST}}$ for SBE according to Equation 4.4. $n_{\mathbf{SP}}$ was determined by Equation 4.5. The optimum d region are shaded in green.

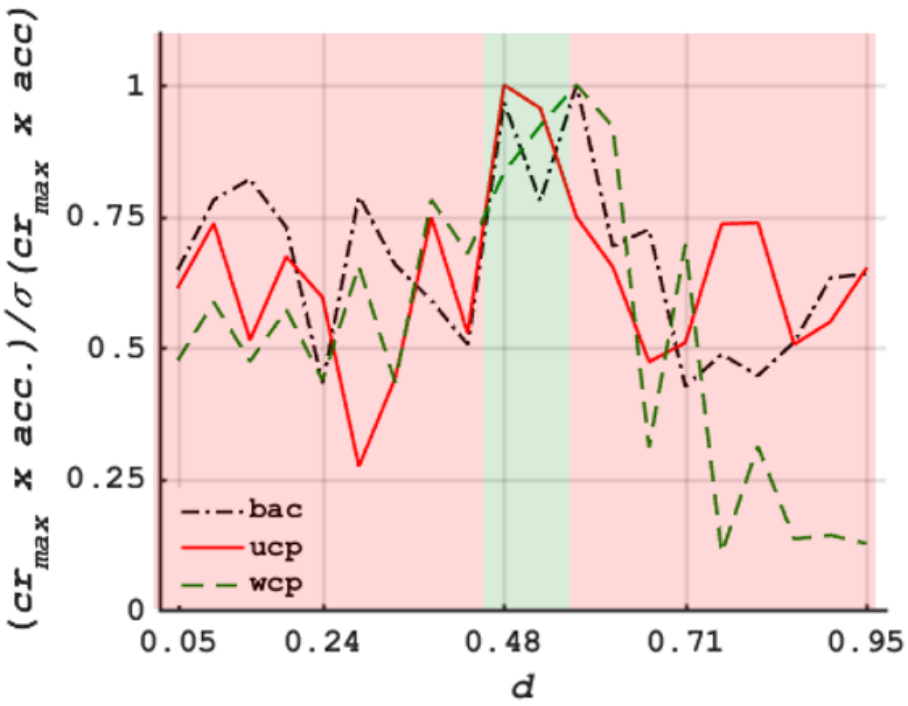


Fig. 4.4.2: A z score plot for the determination of optimum value of d ($0.05 \leq d \leq 0.95$) The optimum d region are shaded in green.

Thus if CR-FS implemented with $n_{\mathbf{ST}}$ estimated from d is between 0.48 and 0.57, a core number of features are retained which leads to good predictions irrespective of the subsets of the training data used for optimization.

Since $n_{\mathbf{SP}}$ was estimated from the empirical equation, it was also important to check if the estimated values are below the optimum (i.e., was SFS being stopped too early). To check this, CR-FS was implemented using $n_{\mathbf{ST}}$ estimated for five d values from 0.48 to 0.57 with $n_{\mathbf{SP}}$ set to be equal to the total number of variables. For each of the five values of d, F-ratio analysis and

the feature selection with CR-FS was performed ten times. During each of the ten iterations, a different subsets of training data was used for F-ratio analysis and model optimization. Feature survival rate was calculated as the number of times a variable was retained after CR-FS. A feature with a survival rate of 100% indicates it was selected in all the ten iterations performed for each of the five d values (i.e., 50 times). Fig. 4.4.3 a - c shows the overall feature survival rate for bac, ucp and wcp datasets. The $n_{\mathbf{SP}}$ estimated from F-ratio analysis of each dataset is indicated by the red vertical line. A sharp drop in feature survival is observed right after the where the search should have been stopped (red line) can be seen.

If each of the ten replicates for the values of d, is treated as an independent feature selection and a survival threshold of six as used earlier for PLS-DA and PCA plots, then the feature survival plot is as shown in Fig. 4.4.4. In Fig. 4.4.4 a-c, in all the three datasets (i.e., bac, ucp and wcp), beyond $n_{\mathbf{SP}}$ (i.e., red vertical line), no variable survives at more than one d value. This indicates that the $n_{\mathbf{SP}}$ values obtained from F-ratio analysis are very good estimates.

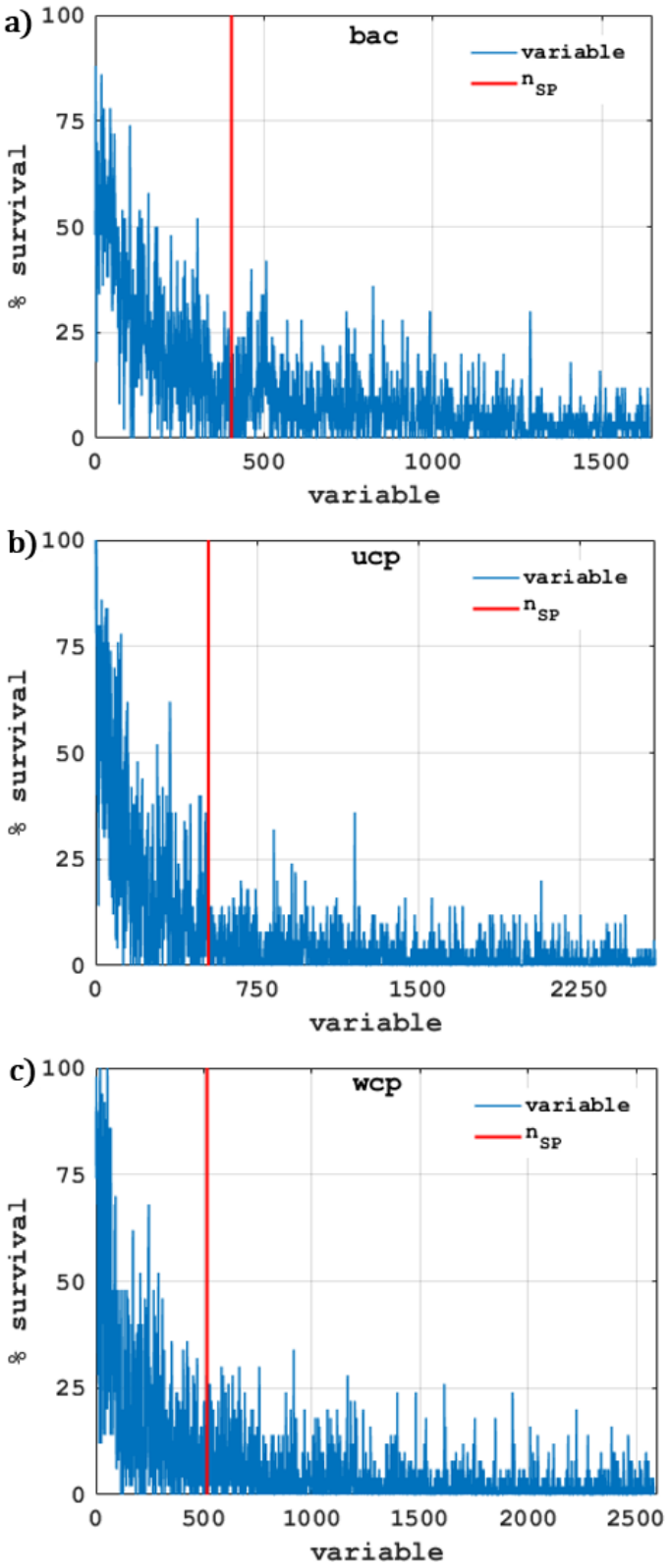


Fig. 4.4.3: Feature survival rate for all variables for (a) bac, (b) ucp and (c) wcp. n_{SP} was estimated from the optimum values of d, all variables were evaluated.

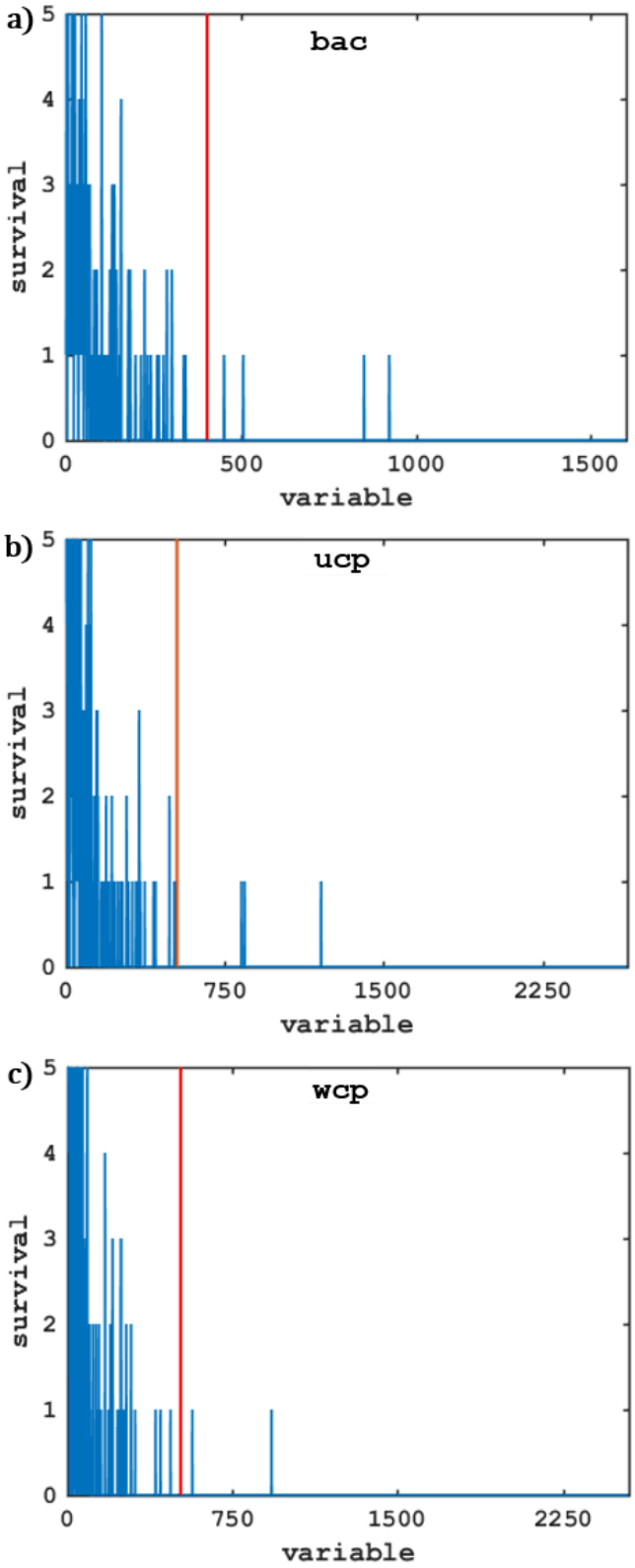


Fig. 4.4.4: Feature survival rate for all variables for (a) bac, (b) ucp and (c) wcp. n_{SP} was estimated from the optimum values of d, all variables were evaluated. Only features that survived at least six (6) iteration are shown.

Finally, CR-FS was implemented on two new datasets, 4 (cof) and 5 (cvp). Each of the datasets was split into two-thirds for training and one-third for external validation sets. Using the training set, true and null ratio analysis were performed using half of the training set data. $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$ were estimated from the F-ratio analysis using $d = 0.48$. This was followed by feature selection with the CR-FS algorithm. The process was repeated ten times using a different subset of the training set data for the F-ratio analysis. For the coffee data, $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$ were 17 and 160, respectively, while for the cvp data the values were 28 and 580, respectively. Features that survived at least six times were used to construct final PCA and PLS-DA models. For the coffee data, 13 out of 701 features met this criteria. Fig. 4.4.5 compares the PCA and PLS-DA results for coffee data before (701 features) and after feature selection (13 features). In the PCA model, the explained variance in the first and second principal components for before and after feature selection were 35% and 76%, respectively. The prediction accuracy for the PLS-DA model improved from 96.3% to 100% after feature selection. For the cvp data, 35 out of 580 features survived. The PCA and PLSDA models for the cvp data before (2781 features) and after feature selection (35 features) are shown in Fig. 4.4.6. The PCA model shows an increase in the explained variance for the first two principal components from 14% to 28%. The PLS-DA prediction accuracy for before and after feature selection were, 90% and 100%, respectively.

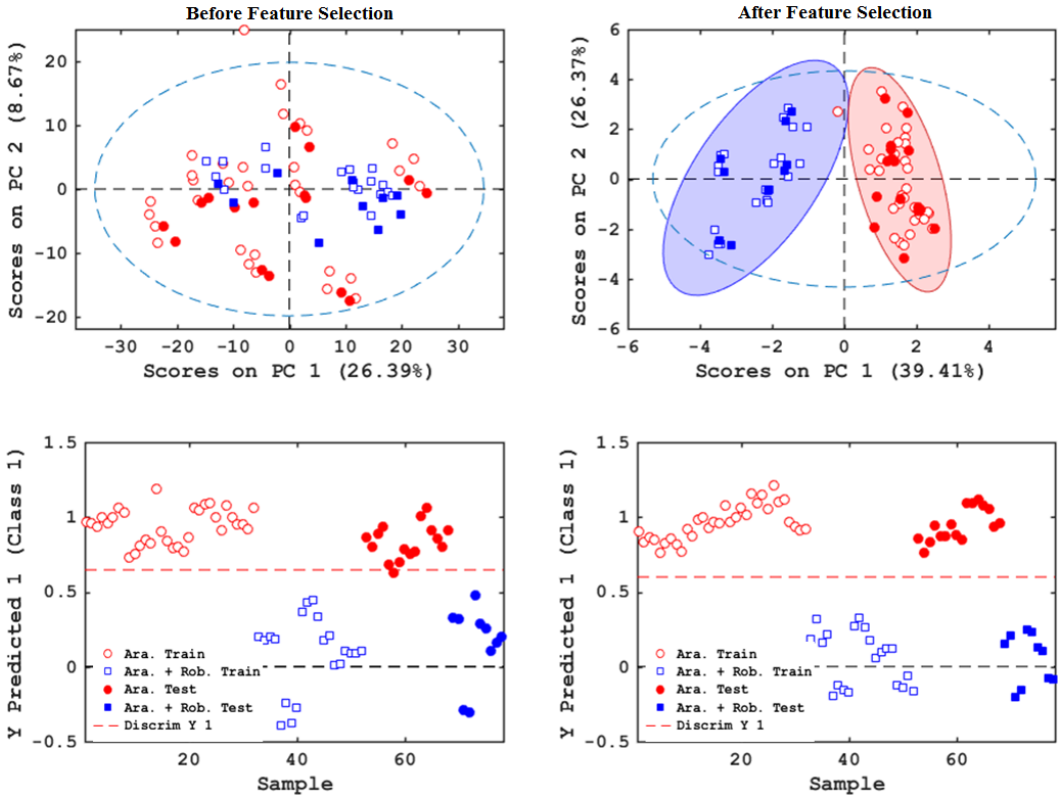


Fig. 4.4.5: PCA and PLS-DA models of Coffee data to be classified as Arabica vs. mixture of Arabica and Robusta. Depicts PCA and PLS-DA models before and after feature selection using 701 and 13 variables, respectively. Feature selection were performed using $n_{\mathbf{ST}}$ and $n_{\mathbf{SP}}$ estimated from Equations 4.4 and 4.5, respectively. Red markers represents Pure Arabica whiles blue markers represent a mixture of Arabica and Robusta coffee samples. Hollow and filled markers represents training and validation set data, respectively.

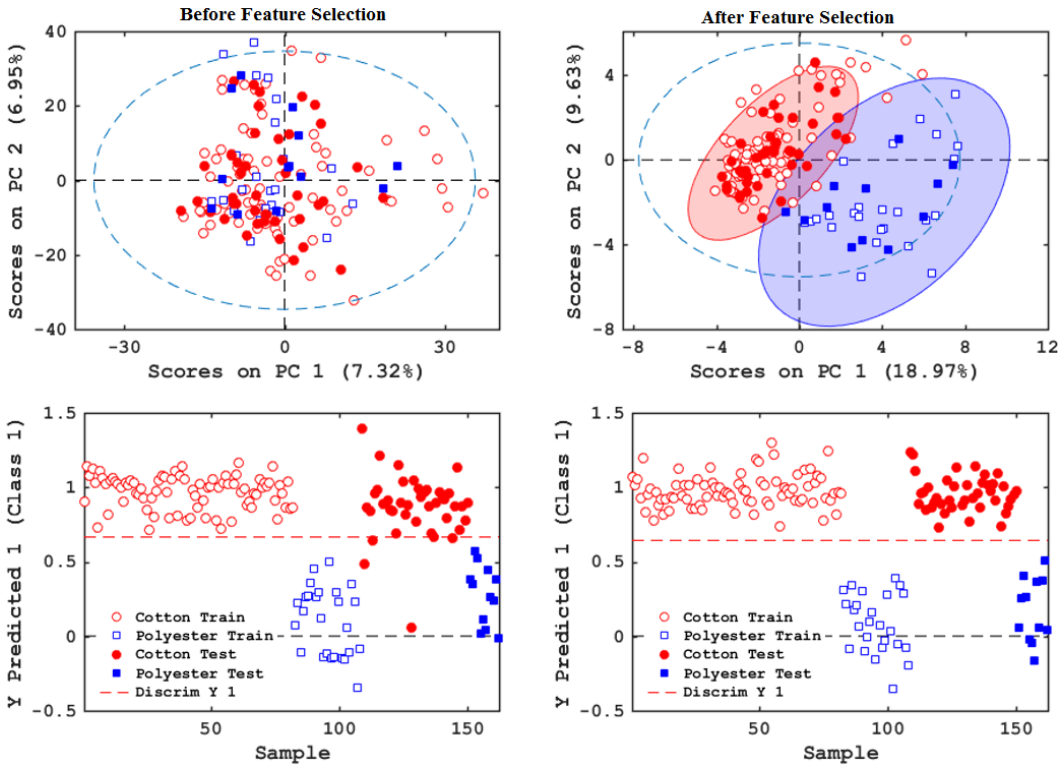


Fig. 4.4.6: PCA and PLS-DA models for fabric data to be classified as cotton vs. polyester.
This figure depicts PCA and PLS-DA models before and after feature selection using 2781 and 35 variables, respectively. Feature selection were performed using n_{ST} and n_{SP} estimated from Equations 4.4 and 4.5, respectively. Red markers represents cotton whiles blue markers represent polyester samples. Hollow and filled markers represents training and validation set data, respectively.

4.5 Conclusions

Through the analysis of true and null F-ratios obtained from a dataset for classification models, an empirical equation was developed to estimate the start and stop number for CR-FS. All but one of the parameters in this equation are obtained by comparing the probability density functions of the true and null F-ratios. The constant to be determined was estimated to be in the range of $0.48 \leq d \leq 0.57$. The validity of this empirical equation was confirmed by testing two new data sets. Using start and stop numbers obtained from the empirical equations, excellent model prediction accuracies were achieved with variables obtained after implementation of CR-FS. The use of this empirical equation can now be used as a guidance in setting the start and stop number for CR-FS, enabling a true automation of the feature selection process.

"A teacher is one who makes himself progressively unnecessary"

Thomas Carruthers

5

Applications of Cluster Resolution Feature Selection

Since its invention, the CR-FS algorithm has been applied largely to the raw or peak tables data obtained from GC-MS or GC×GC-MS analysis. CR-FS however has tremendous potential for applications in other fields of scientific research where classification is the ultimate goal. This chapter explores other classification problems where CR-FS is applicable. My role in the various research work presented in this chapter is the chemometric analysis and machine learning performed on the data from collaborators. In the first two sections of this chapter, synthesis and characterization of inorganic compounds were done by collaborators from the Departments of Chemistry of the following institutions: University of Alberta, University of Houston and Technical University of Munchen. In the final chapter, the strains

Oliynyk A. O., Adutwum L. A., Harynuk J. J. and Mar A., *Chem. Mater.*, 2016, 28(18), 6672-6681. Oliynyk A. O., Adutwum L. A., Rudyk B. W., Pisavadia H., Tehrani M. A., Hukhyy V., Harynuk J. J., Mar A. and Brgoch J., *J. Am. Chem. Soc.*, under review, reference number: ja-2017-08460p).

of *Lactobacillus reuteri* were cultured by collaborators at the Department of Agricultural Food and Nutritional Science at the University of Alberta.

The aim of the first project was to develop chemometric models that would predict the crystal structures from elemental properties. Since not all the properties may be relevant to the prediction, CR-FS was employed to identify relevant features. The initial study involved the prediction of the crystal structures of binary (AB) compounds. The model obtained was used to predict the structure of an entirely new compound. Subsequently, the new compound was synthesized by our collaborators to determine the accuracy of the prediction made from my analysis of the elemental properties. A summary of this study is presented in the first section of this chapter. Due to the success of the structural prediction of binary (AB) compounds, a second study was commissioned to extend the approach to the much more complex ternary (ABC) compounds. In the second section of this chapter, I present the findings of the structural predictions of ternary (ABC) compounds.

Matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) is rapidly becoming the instrument of choice for the identification of bacteria. Automation of bacterial identification protocols is also desirable. However, not all the variables in the entire spectrum obtained from MALDI-TOF MS analysis may be relevant to the correct identification of the bacteria.

Using the concept of exploratory data analysis (EDA), strain level identification/classification of 12 strains of of *Lactobacillus ruteri* was explored. The results of this study are reported in the third section of this chapter.

5.1 Classifying Crystal Structures of Binary Compounds AB through CR-FS and SVM

For any combination of elements, the prediction of the resulting compounds and their crystal structures are important goals of X-ray crystallographers. For the simplest case of equiatomic binary compounds AB, where A and B are any elements in the periodic table, these predictions are still not simple. This is because of the many factors that influence structure formation. By correlating atomic properties and systematizing empirical structural information, it is hoped that crystal structures can be accurately predicted without X-ray diffraction patterns.[208] Earlier attempts were made to relate atomic size factors to rationalize the structures of ionic solids AB and their preferred coordination geometries. This could not be generalized as it failed to account for NaCl-type structures which tend to be far more prevalent than predicted.[209] Atomic properties such as electronegativities and valence electron numbers gave a more favorable picture in generating structure maps (e.g. Mooser-Pearson,[210] Phillips-van Vechten,[211] Pettifor,[212]

Zunger,[213] Villars[214,215]). These structural maps succeeded in segregating structure types. Empirical equations have been used to compute descriptors with the aim of finding patterns in binary AB compounds. These patterns can be used as guides to make predictions for new compounds. Focusing on intermetallic compounds AB, Villars considered 182 descriptors and tested their mathematical combinations to identify three expressions; namely the difference in Zunger pseudopotential radii sums, the difference in Martynov-Batsanov electronegativity, and the sum of valence electrons.[214,216] This separated 988 compounds into 20 structure types with <3% violations.[214] The availability of atomic properties and other empirical descriptors presents an opportunity for chemometric techniques to be used to predict crystal structures.

CR-FS was used to identify descriptors/variables which can be used to predict crystal structures. Using the selected descriptors, support vector machine (SVM) and partial least squares discriminant analysis (PLS-DA) models were constructed and validated. The validated models were used to predict a completely new compound. The new compound was then synthesized by collaborators and characterized to validate the structural prediction by SVM and PLS-DA.

5.1.1 Data Extraction and Organization

Crystallographic data for AB compounds were extracted from Pearson's Crystal Database, ASM Alloy Phase Diagram Database and SciFinder.[217,218] AB compounds that met the following criteria were included in the study:

1. They did not contain hydrogen, a noble gas or elements with Z > 83 (radioactive elements and actinides).

2. They must exhibit exact 1:1 stoichiometry.

Out of 107 classes that met these criteria, only those with at least 30 compounds in a class were chosen for subsequent analysis. This final dataset consisted of 706 AB compounds crystallizing in seven crystal types and 56 variables, and is shown in Table 5.1.1 below. The list of the variables are shown in Appendix C.

Table 5.1.1: Structure types and number of samples in each class for AB compounds

Class	Structure Types	Number of Compounds
1	CsCl	257
2	NaCl	205
3	TlI	102
4	β-FeB	42
5	NiAs	36
6	ZnS	33
7	CuAu	31
	Total	706

5.1.2 Chemometric Analysis

The dataset was split into two parts: two-thirds (470) for training and one-third (236) for external validation. Using half of the training set data (235), variables were ranked according to their F-ratio scores from ANOVA.[167] The CR-FS algorithm was implemented in a three-dimensional PCA score space (PC1 vs. PC2 vs. PC3).[18,19] A start number of 20 was used for the SBE stage and the rest of the variables evaluated during the SFS. PLS-DA and SVM models were constructed with samples from the training set using variables selected by CR-FS. The SVM classification was performed with a radial basis function. A venetian blind cross-validation with 10-fold data split was used to optimize the model. The SVM and PLS-DA models were validated with the external validation set data. The validated models were then used to predict the crystal structure of a completely unknown compound, RhCd.

5.1.3 Synthesis of RhCd and X-ray Diffraction Analysis

A pressed pellet of Rh powder (99.95%, Alfa-Aesar) and filed Cd pieces (99.95%, Alfa-Aesar) in a 1:1 molar ratio with a total mass of 0.2 g was placed in a fused-silica tube, which was evacuated and sealed. The tube was heated to 800 °C. It was kept at that temperature for a week and quenched in cold water. The product was examined by powder X-ray diffraction (XRD) performed on an Inel diffractometer equipped with a curved position-sensitive detector and by energy-dispersive X-ray (EDX) analysis on a JEOL JSM-6010LA scanning

electron microscope.

5.1.4 Results and Discussion

A chemometric approach to the prediction of the structure of binary AB compounds with experimental validation was investigated. CR-FS algorithm is well suited for the simultaneous optimization of multiple-class problems. CR-FS identifies relevant variables by determining their contribution to the separations of clusters in PCA score space.[18–20] The optimization was for a seven-class problem, with each class representing one of the seven common structure types adopted by binary compounds AB which are, CsCl, NaCl, ZnS, CuAu, TlI, β-FeB, and NiAs. After feature selection with CR-FS, thirty-one out of fifty-six variables were retained. This included highly ranked variables such as average Martynov-Batsanov or Mulliken electronegativities, Pauling electronegativities (and expressions derived from them), interatomic distances, and differences of Zunger radii sums ($r_s + r_p$). Other low-ranked variables such as average number of valence electrons and some expressions derived from Zunger radii sums were also retained. Thus variables included in the final model after feature selection consist of those retained in the backward elimination step and those added in the forward selection step (blue circles) (Fig. 5.1.1). A complete list of all the variables for the study of AB compounds can be found in the Appendix C. PLS-DA and SVM models were generated with the training set using the thirty-one variables retained. In both

models, internal cross validation was performed using venetian blinds with a ten-fold data split. Both SVM and PLS-DA models were validated with the validation set data. The model was then used to predict the crystal structure of a completely unknown compound, RhCd.

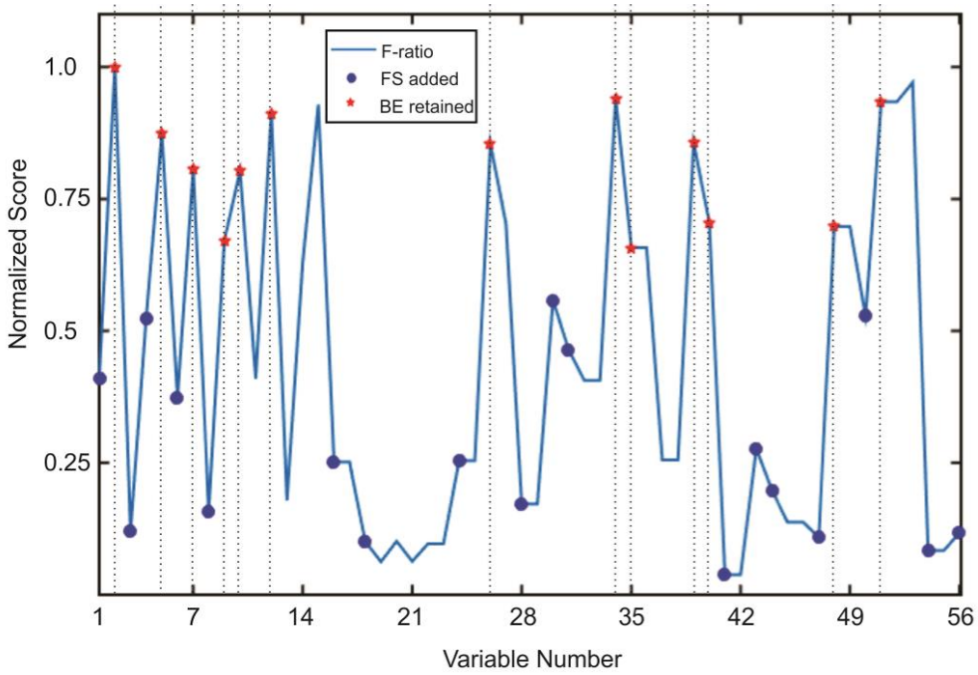


Fig. 5.1.1: Fisher ratio scores for all variables (identified in the legend) selected during backward elimination (red stars) and forward selection (blue circles).

Class predicted probability for PLS-DA is shown in Figure 5.1.2. The PLS-DA model predicted the training set data with sensitivity of 95.9% and specificity of 66.6%. Although the model predicts the CsCl-type structure correctly, the false positive rate is high with an overall accuracy of 77.2%. When the model was applied to the validation set, the sensitivity and specificity were 965% and 66.0%, respectively, and an accuracy of 77.1%.

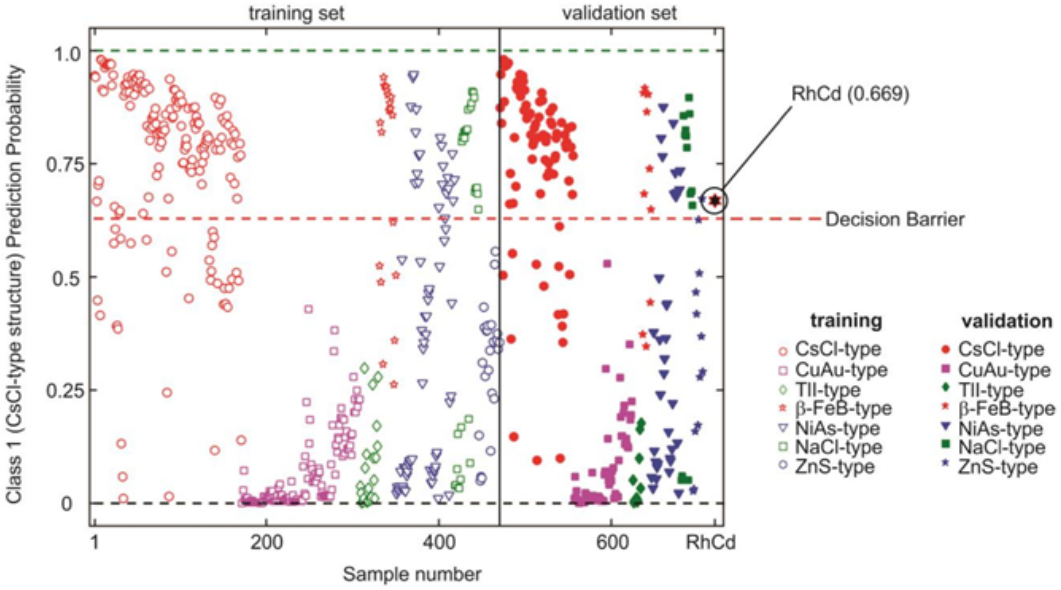


Fig. 5.1.2: PLS-DA class predicted probability for CsCl-type. Hollow and filled markers indicates training and validation sets, respectively. The class predicted probability for RhCd is circled.

The prediction probability for the test compound RhCd was 0.669, which is only slightly higher than the decision boundary as shown in Fig. 5.1.2. The SVM classification model was generated to predict various structure types. The prediction probabilities for the CsCl-type structure were much stronger (Fig. 5.1.3). For the training set data, the sensitivity was 100%, the specificity was 99.3%, and the accuracy was 99.6%; for the validation set data, the sensitivity was 94.2%, the specificity was 93.2%, and the accuracy was 93%. Thus, the model performance was significantly better with SVM than with PLS-DA methods. RhCd was predicted to be a CsCl-type with a predicted probability of 0.918 by SVM (Fig. 5.1.3). The refined crystallographic data for RhCd can be found in Appendix C.

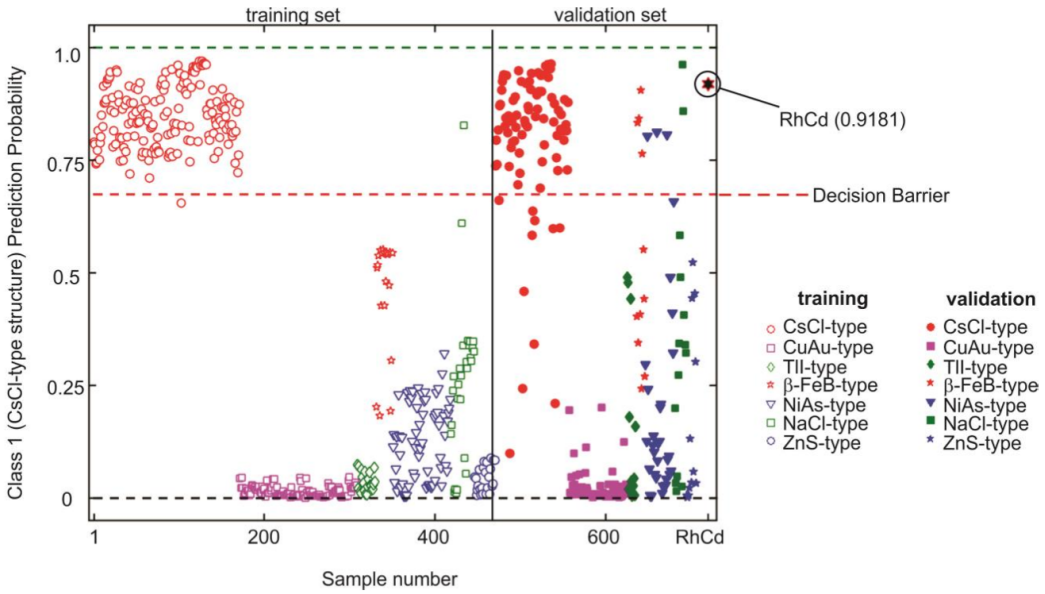


Fig. 5.1.3: SVM class predicted probability for CsCl-type. Hollow and filled markers indicates training and validation sets, respectively. The class predicted probability for RhCd is circled.

RhCd was synthesized by collaborators in Dr. A. Mar's group at the University of Alberta. The synthesized RhCd product was examined by SEM, EDX, and powder XRD (Fig. 5.1.5). Small single crystals, < 50 μm in their longest dimension, were obtained. Their average composition was 47(2)% Rh and 53(2)% Cd, in excellent agreement with the formula RhCd. The powder XRD pattern confirms that RhCd adopts the CsCl-type structure.

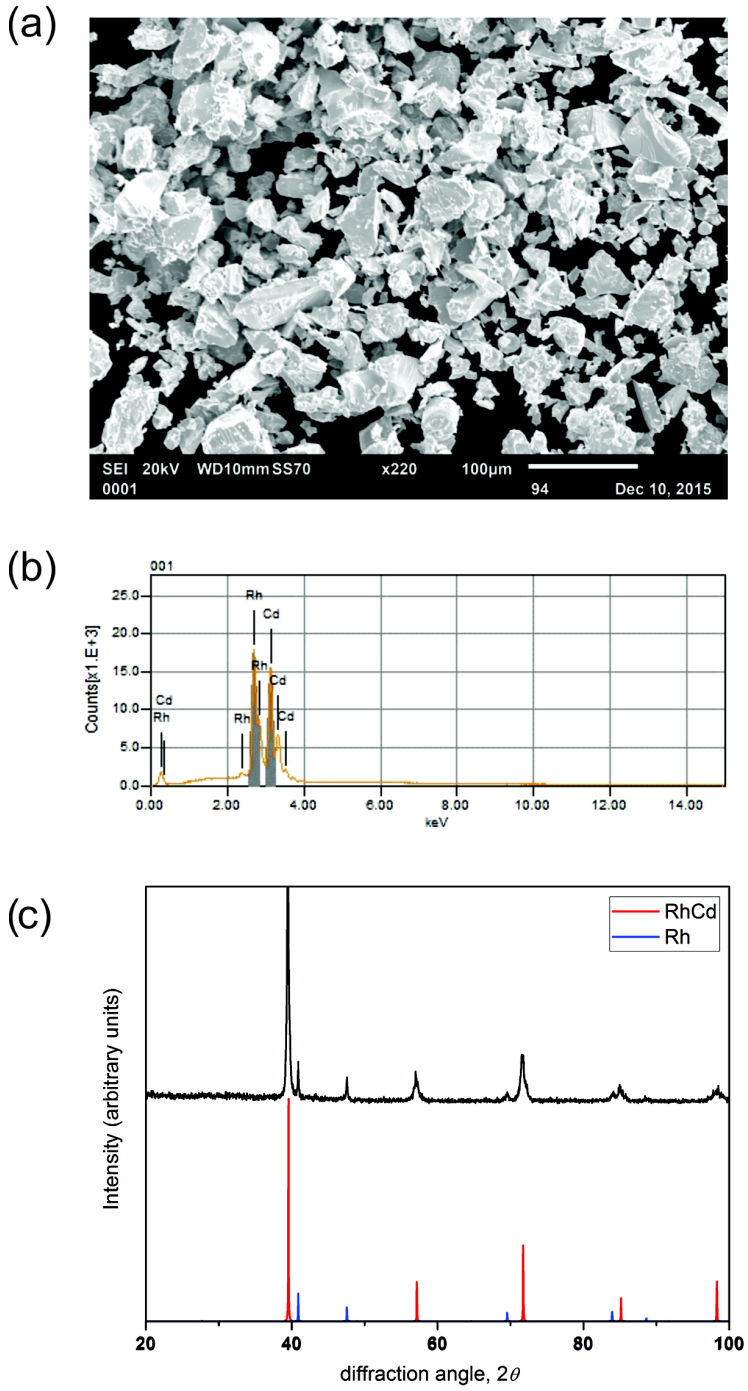


Fig. 5.1.4: New binary compound RhCd. (a) SEM image of crystals, (b) EDX spectrum indicating presence of equal ratios of Rh and Cd in crystals, and (c) powder XRD pattern confirming CsCl-type structure.

5.1.5 Conclusion

Using supervised learning methods PLS-DA and SVM, models were obtained which could predict the structural types from elemental properties selected by CR-FS with a high degree of accuracy. In general, SVM performs better than PLS-DA. Using the same approach, a new compound RhCd was predicted to adopt a CsCl-type structure. This was further confirmed by analyzing the synthesized compound.

5.2 Machine-learning structural characterization of ABC ternary equiatomic compounds and their polymorphs

Metallic phosphides that contain transition metals are known to possess superior mechanical properties. These include wear resistance and hardness in corrosion-resistant films which correlate with the phosphorus content.[219,220] Transition metal phosphides also exhibit wide ranges of physical properties that are interesting for magnetic and electronic applications.[221–223] In addition, they have catalytic applications in hydroprocessing, with higher catalytic activity for hydro-denitrogenation and hydro-desulfurization reactions relative to those reported with sulfides.[224] Ternary phosphides exists in various structural types. A large number of them exist as ZrNiAl-type or TiNiSi-type. TiFeP equiatomic phosphide has both ZrNiAl-type and TiNiSi-type even under the same synthetic conditions.[225–228] Several other ternary phosphides exist as polymorphs under various synthetic conditions.[229–233] The prediction of the structure type adopted by these compounds from a chemometric perspective has not been explored.

As presented in the earlier section 5.1, CR-FS in combination with SVM classification proved to be a very useful in predicting the structure type of binary AB equiatomic compounds.[141] Synthesis of a new compound that was predicted further confirmed the reliability of the prediction. This

section illustrates a similar approach towards the prediction of the structures of ternary equiatomic ABC equiatomic compounds, where A, B and C are the constituent elements, with emphasis on ternary phosphides. Descriptors obtained from atomic properties and mathematical transformations thereof were subjected to CR-FS. Features retained by CR-FS were used for classification models using SVM.

5.2.1 Data Extraction and Organization

Crystallographic data of equiatomic ABC compounds were extracted from Pearson's Crystal Data (2014 and 2015 editions) and ASM Alloy Phase Diagram Database.[234,235] Compounds were included in the study if they satisfy all of the following:

1. A, B, and C are not hydrogen, a noble gases, or elements with Z > 83 (radioactive elements and actinides).

2. The compound exhibit exact 1:1:1 stoichiometry, without crystallographic site mixing.

3. ABC structure has been confirmed experimentally.

In total, 1556 unique individual compounds belonging to seven structure types were selected. Table 5.2.1 shows the number of compounds in each structure type used for this study. Variables representing atomic properties

and mathematical transformations thereof for the selected compounds were obtained from open sources.

Table 5.2.1: Structure types and number of samples in each class of ABC compounds

Class	Structure Types	Number of Compounds
1	TiNiSi	670
2	ZrNiAl	502
3	PbFCl	154
4	LiGaGe	74
5	YPtAs	69
6	UGeTe	49
7	LaPtSi	38
	Total	1556

5.2.2 Chemometric Analysis

The data was presented in a 1556×990 matrix, where 1556 is the number of compounds and 990 is the number of variables. The 990 variables were as a result of 33 elemental properties combined with 30 formulae each (Appendix C). The data was split into two parts: two-thirds (1037) for training set and one-third (519) for external validation set. Feature selection was performed as described section 5.1.2. Optimization was performed in three dimensions. In this case however, a start number 100 was used for the SBE stage and the rest of the variables evaluated during the SFS. To avoid overfitting the training data, the feature selection was repeated twenty times. In each iteration, a different subset of the training set data was used for variable ranking and optimization.

Variables that were selected at least 11 out of the 20 iterations (i.e., >55%) were used for SVM models. The SVM classification was performed with a radial basis function. A venetian blind with 10-fold data split cross-validation was used. The model was evaluated with the external validation set. The processing method employed was row normalization and autoscale (mean center and scale to unit variance).

5.2.3 Results and Discussion

Crystal structure formation is a complex process, which is influenced by several factors. Ultimately, the structure type that any crystalline compound prefers is dictated largely by its constituent elements. The knowledge of compounds with known structure types can aid future predictions. Feature selection offers a path to identify which atomic properties (or combinations thereof) are relevant in determining preferable structure types. CR-FS and SVM was successfully employed to predict the structure type of binary AB compounds with experimental validation.[141] Prediction of structure type of ternary equiatomic ABC compounds were explored using a similar approach. The variables was initially ranked according to their F-ratio from ANOVA. Fig. 5.2.1 shows the results of the F-ratio scores. After feature selection with the CR-FS algorithm, 113 variables out 990 were selected. The selected features are shown in Fig. 5.2.1 (red).

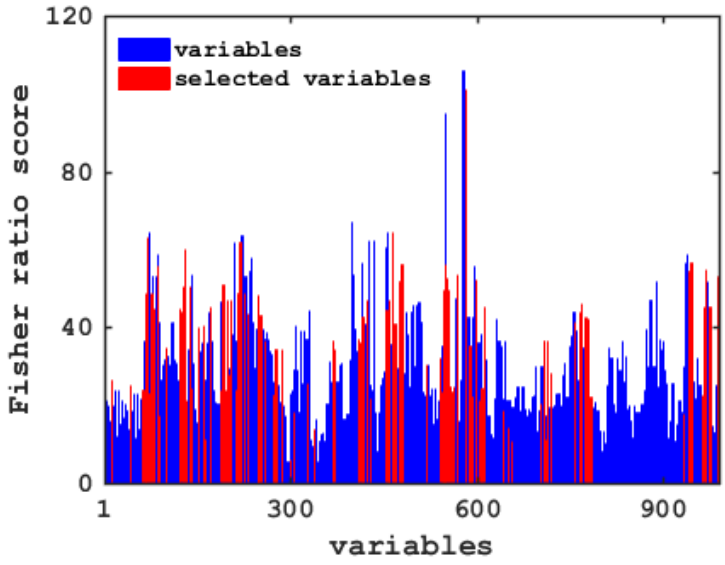


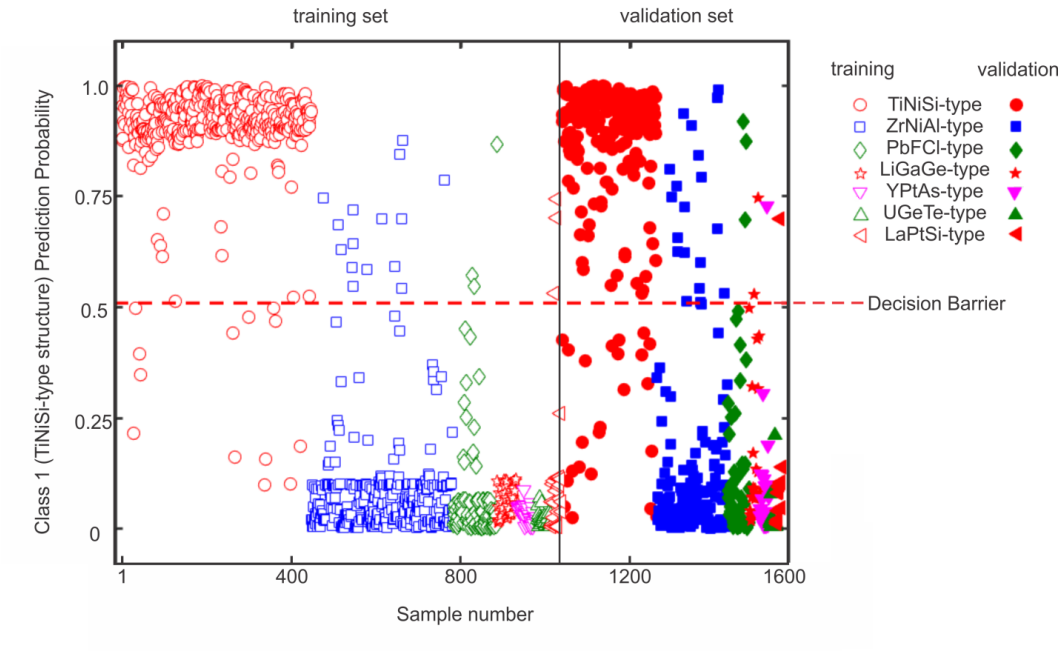
Fig. 5.2.1: F-ratio scores of all variables. Variables that were retained after feature selection are shown in red.

Fig. 5.2.1 shows that not all highly ranked features were retained. Conversely, not all variables with low rank were ignored. This buttresses the point that F-ratios and for that matter, ranking methods are indicative of potential relevance and do not guarantee that the variable will actually be important for the question at hand. Forty variables (over one third of all variables selected) are related to atomic size descriptors. The variables obtained after feature selection (i.e., 113) were used to generate SVM model for prediction. The validation set prediction sensitivity and specificity were 97.3% and 96.9%, respectively, with an error rate of 3.10%. Table 5.2.2 shows the details of the predictions are shown for all the seven structure types studied. TiNiSi and ZrNiAl are predicted to a higher degree of specificity relative to the other structure types.

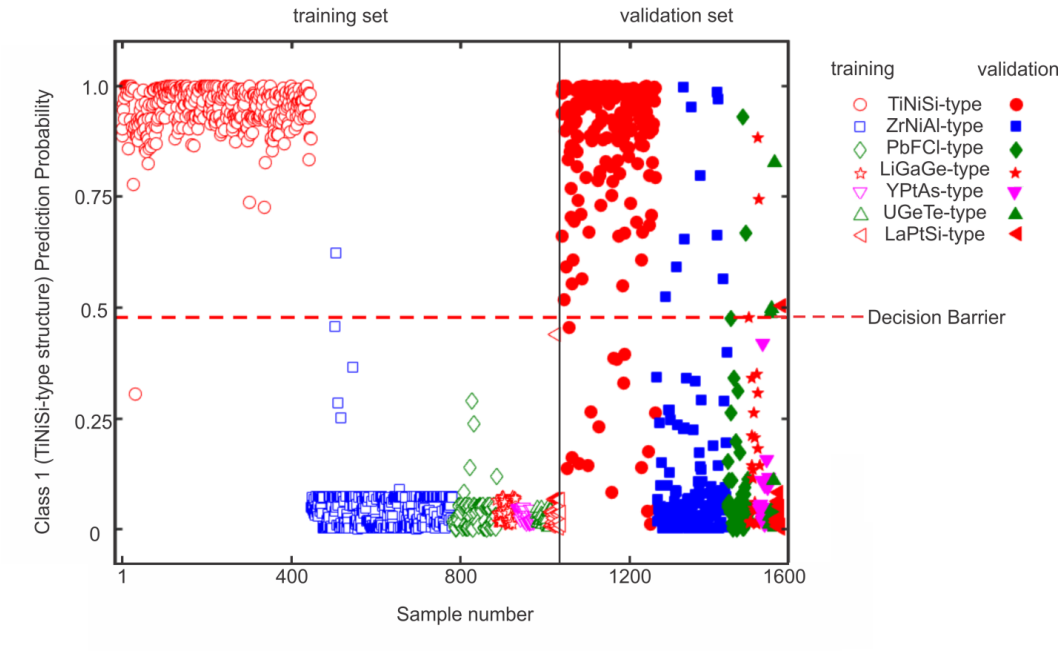
Table 5.2.2: SVM model sensitivity, specificity and accuracy for structure types

Structure type	TiNiSi	ZrNiAl	PbFCl	LiGaGe	YPtAs	UGeTe	LaPtSi	mean
Training set								
Sensitivity	100	100	100	100	100	100	100	100
Specificity	100	99.1	99.0	96.1	100.00	100	100	99.1
Accuracy	99.8	99.6	99.9	99.8	100.0	100	100	99.9
Error Rate	0.20	0.40	0.10	0.20	0.00	0.00	0.00	0.12
Validation Set								
Sensitivity	94.0	96.9	99.6	99.4	100	99.6	100	98.5
Specificity	91.96	93.41	96.1	88.0	100	87.5	100	93.9
Accuracy	93.1	95.0	96.7	97.3	99.0	99.2	98.3	96.9
Error Rate	6.9	5.00	3.30	2.70	1.00	0.80	1.70	3.10

The SVM prediction probabilities for TiNiSi-types before feature selection (990 variables) and after feature selection (113 variables) are shown in Fig. 5.2.2. The SVM prediction probabilities for ZrNiAl-types before feature selection (990 variables) and after feature selection (113 variables) are shown in Fig. 5.2.3. Some ternary phosphides can exist as polymorphs with both TiNiSi- and ZrNiAl-types. The predicted probabilities of known polymorphs using thirty-five known polymorphs are shown in Fig. 5.2.4. Compounds that exist as polymorphs under the same synthetic conditions were predicted with lower probabilities (0.7). Compounds whose preferred structural type varies with synthetic method were predicted with a higher degree of confidence (> 0.8). This indicates that for co-existing polymorphs, preferred structure types are determined largely by their elemental composition.

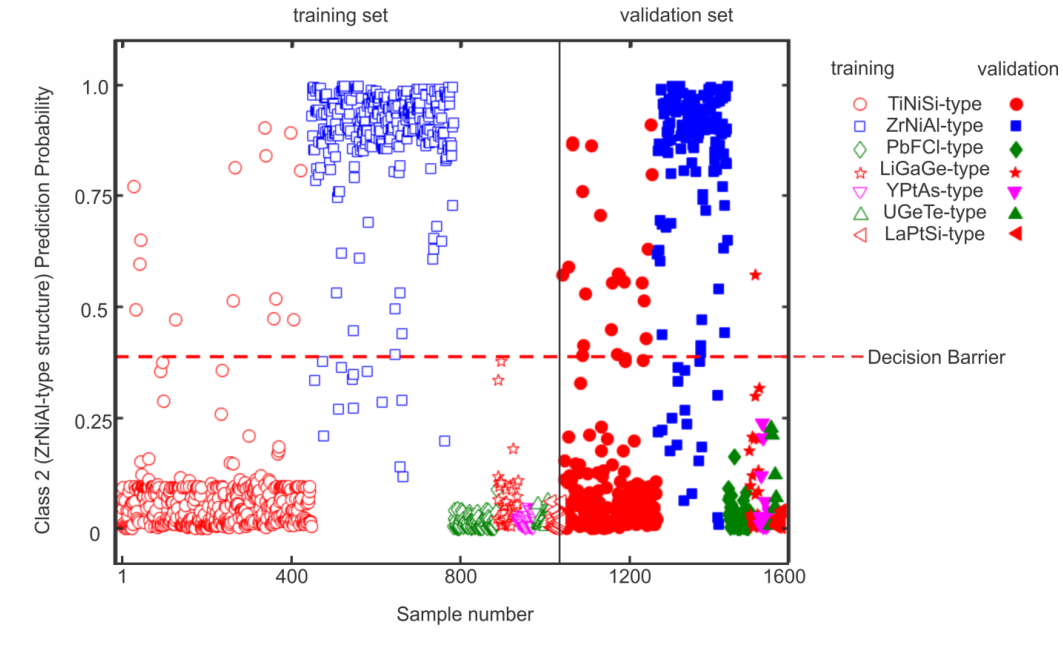


(a) Before feature selection with CR-FS (990 variables).

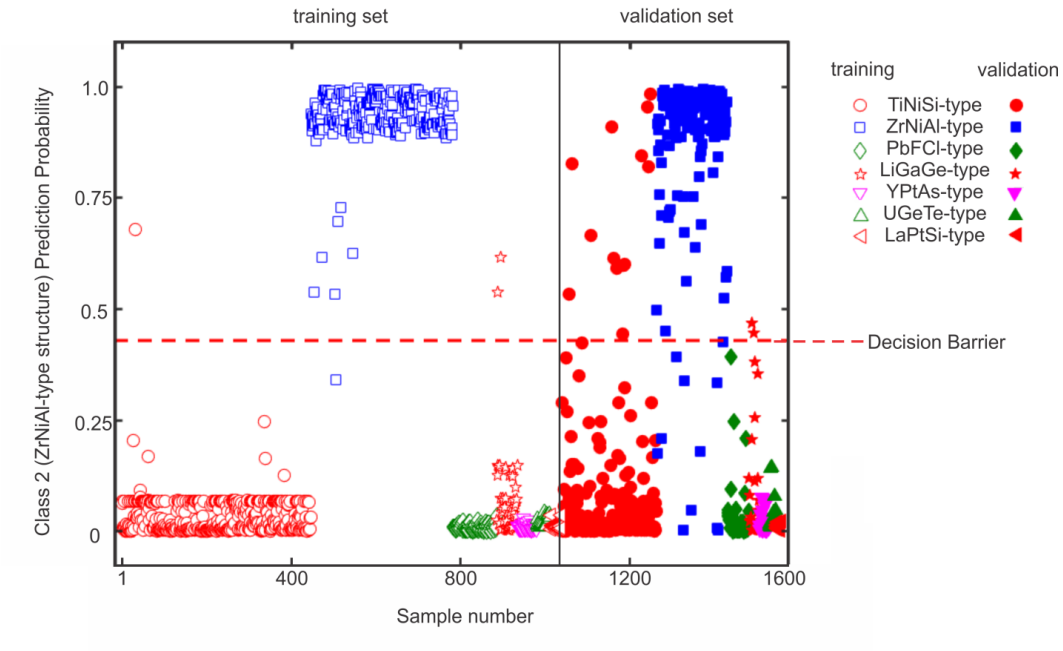


(b) After feature selection with CR-FS (113 variables).

Fig. 5.2.2: SVM predicted probabilities for TiNiSi-type structure. Models were constructed using 113 features retained after feature selection with CR-FS. The hollow and filled markers represent training and validation sets, respectively.



(a) Before feature selection with CR-FS (990 variables).



(b) After feature selection with CR-FS (113 variables).

Fig. 5.2.3: SVM predicted probabilities for ZrNiAl-type structure. Models were constructed using 113 features retained after feature selection with CR-FS. The hollow and filled markers represent training and validation sets, respectively.

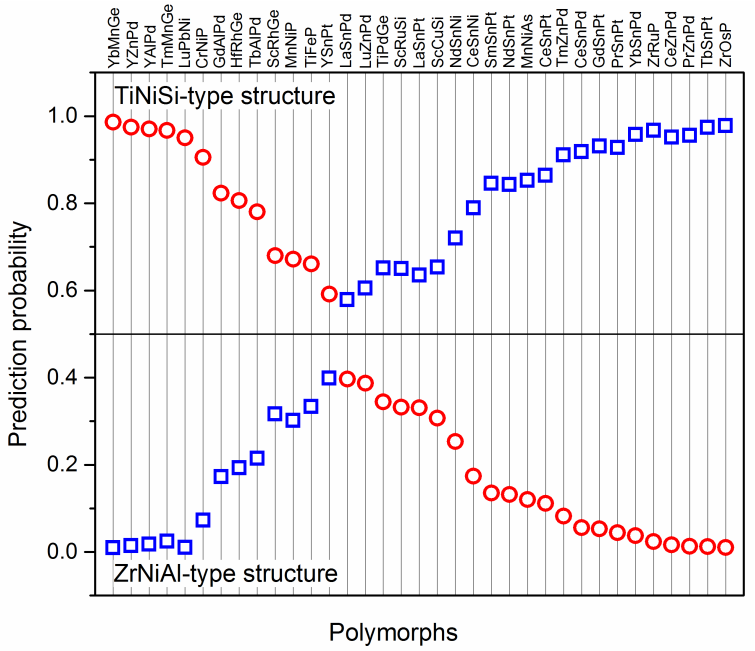


Fig. 5.2.4: Prediction probability to belong to a certain structure type for 35 experimentally confirmed polymorphs that adopt either TiNiSi- or ZrNiAl-type structure.

5.2.4 Conclusion

A machine-learning approach has been used to classify ternary equiatomic structure types based on parameters obtained from the elemental composition of structure type representatives. Variables important for the classification model were selected with the CR-FS algorithm. Out of 990 initially proposed features, 113 were selected. Compounds were classified using SVM with two-thirds and one-third for training and validation, respectively. The validated accuracy, sensitivity, and specificity is 96.9%, 97.3%, and 93.9% respectively. The variables important for segregation of ternary ABC structure types are mainly associated with A and C elements.

5.3 Strain Level Distinction of *Lactobacillus reuteri* through successive feature selection and principal component analysis

Arguably, the most important task in the study of bacteria is the classification and identification.[236] Several approaches to the identification of bacteria have been reported over the years.[236] In a clinical setting, the accurate and rapid identification of bacteria is crucial to diagnosis and management of bacterial infections. Earlier methods for classification and identification of bacteria have been based on morphology, substrate utilization and staining characteristics.[236] Substrate utilizations for fastidious bacteria with specific growth requirements are used to isolate and identify some types of bacteria. Gram staining, developed by Hans C. Gram, determines the presence or absence of peptidoglycan in the cell wall, and is the most popular staining method for bacterial classification.[237] Analysis of volatile organic compounds (VOCs) (e.g., fatty acid) profiles towards the identification and classification of bacterial have been reported.[238–240] Sequence analysis of the ribosomal RNA genes is the most acceptable method for bacterial species identification.[241–243] Matrix assisted laser desorption ionization-Time of Flight Mass Spectrometry (MALDI-TOFMS) has become a standard tool for protein and lipid analysis.[244] Analysis of proteins / lipids provide unique patterns that are specific to some organisms. These patterns can serve

as fingerprints for the identification of bacteria.[243–248] MALDI-TOFMS is a high-throughput technique with ease of automation, and hence presents as a simple, convenient and reliable approach to bacteria identification. Comparison of the protein or lipid profile/spectrum from a sample to that of a database is the rapid way of automating the identification with MALDI-TOFMS data. Rapid and automated application of MALDI-TOFMS spectra for bacterial identification require the use of some computational and statistical techniques. This presents an opportunity for chemometric techniques to be employed. However, the lack of reproducibility of spectra from run-to-run may contribute to inconclusive results.[245] In addition, it may be difficult to distinguish between closely related strains using the entire spectrum. Under such conditions, it is relevant to identify regions of the mass spectra that are more useful for the intended identification. After all, the use of MS to identify bacteria relies on the presence of some particular compound(s) in the spectrum.

Distinguishing between organisms and for that matter bacteria, becomes increasingly difficult as one moves down the taxonomic rank. MALDI-TOF MS has a reputation for typing/identification of bacteria at the species level but has not been widely tested for phylogenetic groups below the species level (i.e., strain). The aim of this study was to explore the capabilities of MALDI-TOF MS to distinguish between twelve closely related strains of *Lactobacillus reuteri*.

Spectra of samples are obtained via MALDI-TOFMS analysis of whole cell extracts. The analysis involved the use of exploratory data analysis in combination with CR-FS for the identification of relevant variables. *Lactobacillus reuteri* was chosen since its genome has been sequenced and characterized with respect to their taxonomic position below the species level.[249–251] The strains of *Lactobacillus reuteri* studied were 100-23, TMW1.112, TMW1.656, LTH2584, FUA3108, FUA3168, FUA3324, FUA3400, FUA3401, LTH5448, lpuph and mlc3.

5.3.1 Bacterial Culture and Sample Preparation

The strains of bacteria were grown separately on MRS agar at 37 °C under anaerobic conditions for two days prior to extraction. A 300 µL portion of water was pipetted into a 1.5 mL Eppendorf tube and a colony of micro-organisms were added. A 900 µL portion of ethanol was added and the tube vortexed thoroughly. The mixture was centrifuged at 15000 rpm for 2 min. The ethanol layer was decanted and the residue centrifuged an additional 2 min. All excess ethanol was removed with the aid of Eppendorf pipette. FiftyµL of 70% formic acid was added, vortexed thoroughly and allowed to stand for approximately 5 min. A 50 µL portion of acetonitrile was added and vortexed thoroughly. The mixture was centrifuged at 15000 rpm for 2 min. The supernatant was isolated for MALDI-TOF MS analysis.

5.3.2 MALDI-TOF MS Analysis

A 1 µL aliquot from the samples was spotted on a Bruker Daltonics M^{TM} AC800 Anchorchip^{TM} target plate and air-dried. A 1 µL portion of α-cyano-4-hydroxycinnamic acid (5 mg/mL in 50% H_2O) and 50% acetonitrile (containing 2.5% trifluoroacetic acid) was spotted on top and allowed to dry. Mass spectra were obtained in the positive linear mode of ionization using a Bruker Daltonics (Bremen, GmbH) UltrafleXtreme MALDI-TOF/TOF MS with m/z range of 2000 to 20000. Each sample was spotted five times on the target plate and five spectra were acquired for each spot (25 acquisitions / sample). Each spectra was an average of 1000 shots. Acquisition was done in automated mode using the Bruker Flex Control (Ver. 3.4 Build135) and Bruker WARP-LC (Ver. 1.3 Build 136.138) software packages. Data was exported using the Bruker FlexAnalysis software (Ver. 3.4 Build 76).

5.3.3 Chemometric Analysis

Each spectrum was imported into Matlab®as a vector of 84992 elements. All the spectra were compiled into a matrix of 1050 × 84992 (sample × variables). The baseline drifts in the data were corrected using the airPLS algorithm.[79] Noise in the spectra were minimized by applying a Savitsky-Golay smoothing (polynomial order = 2, window = 33).[76] During the analysis, the dataset (or subsets thereof) was split into two-thirds for training and one-third for validation.

Variables were ranked using the F-ratio scores obtained from the training set data. The preprocessing method used was normalization to unity and autoscale (mean center and scale to unit variance). The entire data was projected into a PCA score space to identify clusters. Cluster separations were optimized to increase model parsimony. Where clusters were well separated, an F-ratio threshold was used to reduce the number of variables in the model. Otherwise, feature selection was performed using the CR-FS algorithm.[18,19] Where CR-FS was used, the optimization was done in three-dimensional score space (PC1 vs. PC2 vs. PC3) using start and stop numbers of 2000 and 10000, respectively.

5.3.4 Results and Discussion

Identification of bacteria towards diagnosis and management of infectious diseases is a crucial clinical goal. Rapid and accurate identification is needed to determine the ideal antibacterial agent to administer. In this regard, the need for a fast and efficient technique cannot be over emphasized. MALDI-TOFMS is becoming an accepted tool for bacteria identification and strain typing. In combination with chemometric analysis, a fast, objective and automated workflow can be realized. This study explored the potential of MALDI-TOFMS for bacteria typing/identification below the speices level. This initial study focused on strain level identification of twelve strains of *Lactobacillus reuteri*, a well-characterized bacterium.[249–251]

Relevant variables are selected with either an F-ratio threshold (filter method) or CR-FS method based on the initial separation of the clusters in principal component analysis (PCA) score space. The use of a metric to rank variables in order of relevance without an induction algorithm falls under the filter methods of feature selection described earlier. Where the filter method is used, variables with F-ratio values greater than a factor (a) of standard deviation (σ) of all F-ratios were kept (i.e., F-ratio > $\sigma \times a$). The factor (a) is set at a value above which an increase brings no appreciable increase in the explained variance captured in the first two principal components (<5%). Otherwise, CR-FS is employed on the training set data. Variables retained by either F-ratio threshold or CR-FS on the training set are used to construct PLS-DA models.

A PCA score plot for all the samples in the twelve classes showed three distinct clusters as shown in Fig. 5.3.1. Each cluster contained at least two strains and were denoted Class A (two strains: FUA3108 & FUA3401), Class B (two strains: TMW1656 & mlc3) and Class C (eight strains: 100-23, TMW1112, LTH2584, FUA3168, FUA3324, FUA3400, LTH5448 & Lpuph). A PLS-DA model was constructed to classify the samples into either Class A, B or C. Fig. 5.3.2 shows the results of the PLS-DA model using all the variables. A class predicted sensitivity, specificity as well as accuracy were 100%. Samples in Class A have unusually high hotelling T^2 and their y -

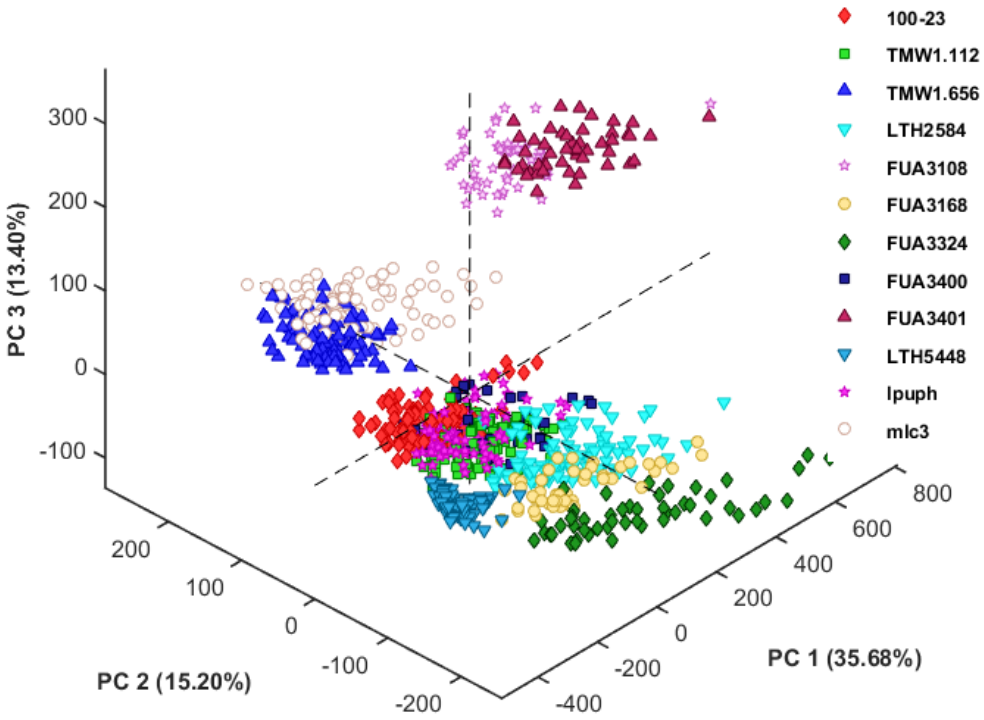


Fig. 5.3.1: PCA Plot of MALDI-TOF MS data from analysis of bacterial samples. The data is separated in three distinct groups, A - FUA3108 & FUA3401, B - TMW1.656 & mlc3, and C - 100-23, TMW1.112, LTH2584, FUA3168, FUA3324, FUA3400, LTH5448 & Lpuph.

predicted shows samples too close to the discrimination barrier.

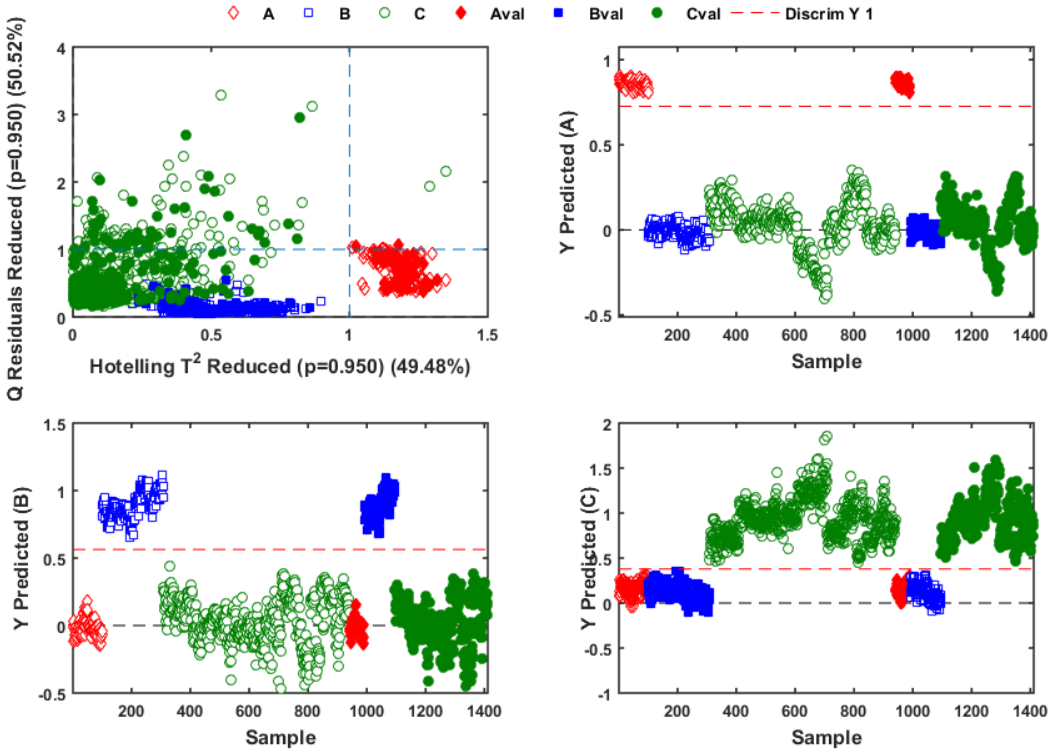


Fig. 5.3.2: PLS-DA y-predicted and Q residuals plot for bacteria samples belonging to Class A, B and C. These models were constructed using all the variables in the MALDI-TOF MS spectrum for all samples. Class A - red diamonds, Class B - blue squares and Class C - green circles. The hollow and filled markers represents samples in the training and validation sets, respectively.

To improve on the previous model, feature selection was performed with an F-ratio threshold as previously described. An *a* value of two was used which reduced the number of variables to 4278. A PLS-DA y-predicted plot for Class A, B and C using only the features retained is shown in Fig. 5.3.3. In Fig. 5.3.3, prediction sensitivity, specificity and accuracy for the validation sets were 100% each. Unlike Fig. 5.3.2, fewer variables were used (i.e., 4278). In addition the samples in Class A and Class B are further away from the discrimination

line. The hotelling T^2 of Class A is within the 95% confidence limit. Plots of the background corrected and smoothed average spectra of samples in Class A, B and C are shown in Fig 5.3.4. The 4278 variables retained are shown in red.

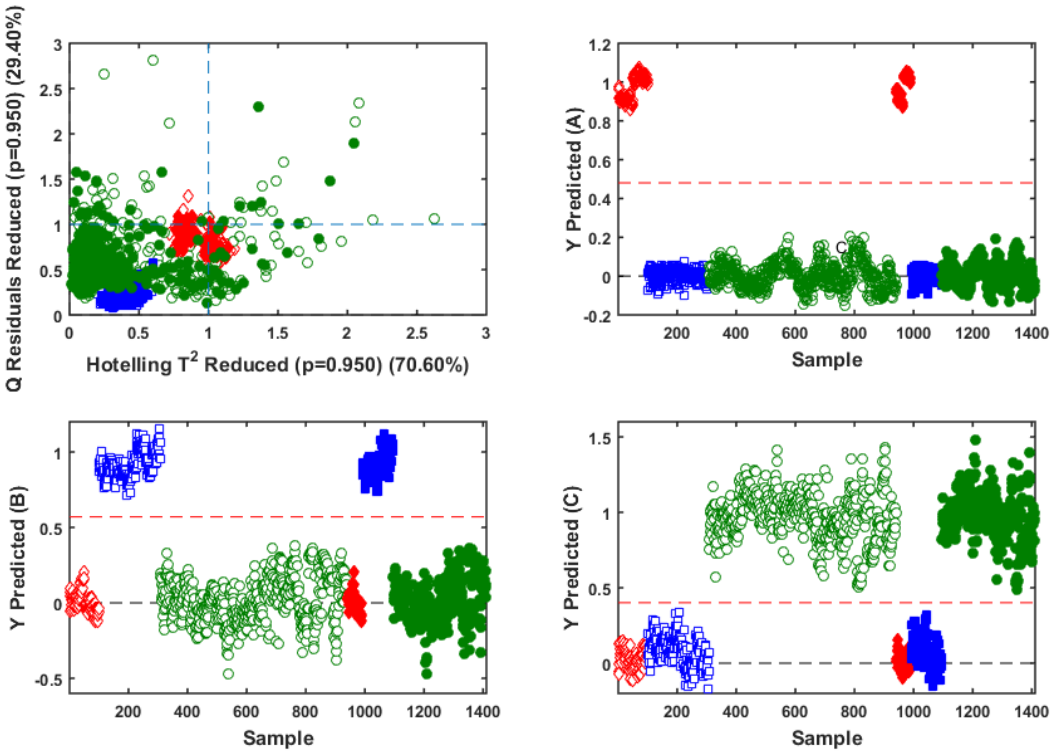


Fig. 5.3.3: PLS-DA y-predicted plot for bacteria samples in Class A, B and C. Model was constructed using only variables retained after the application of F-ratio threshold at $a = 2$. Total variables retained were 4278. Class A - red diamonds, Class B - blue squares and Class C - green circles. Hollow markers and filled markers are for training and validation sets, respectively.

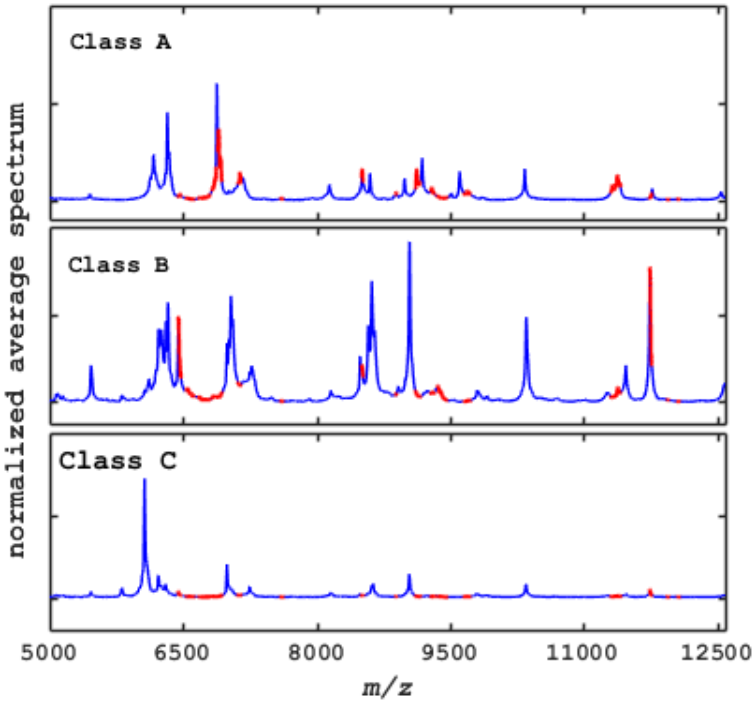


Fig. 5.3.4: Plots of average background-corrected spectra of classes A, B, and C showing m/z that survived the feature selection (4278 variables). Portions shown in red indicates variables that were retained after feature selection.

Class A data consisted of the strains FUA3108 and FUA3401. PCA and PLS-DA models constructed using all the variables are shown in Fig. 5.3.5. Model sensitivity, specificity and accuracy were all 100%. Since a more parsimonious model is desirable, an F-ratio threshold with $a = 5$ was used. All but 817 variables were eliminated. Fig. 5.3.6 shows the PCA and PLS-DA models constructed using the 817 variables. Again a similar prediction accuracies are obtained using approximately 10% of the total variables. Moreover, there is an increase in the explained variance captured in the first two PCs, from 63.5% to 99.2%. The y-predicted values from the PLS-DA models are close to ideal (i.e., 1). A plot of normalized average spectra for FUA3108 and FUA3401 are shown in Fig. 5.3.7. Variables retained are shown in red.

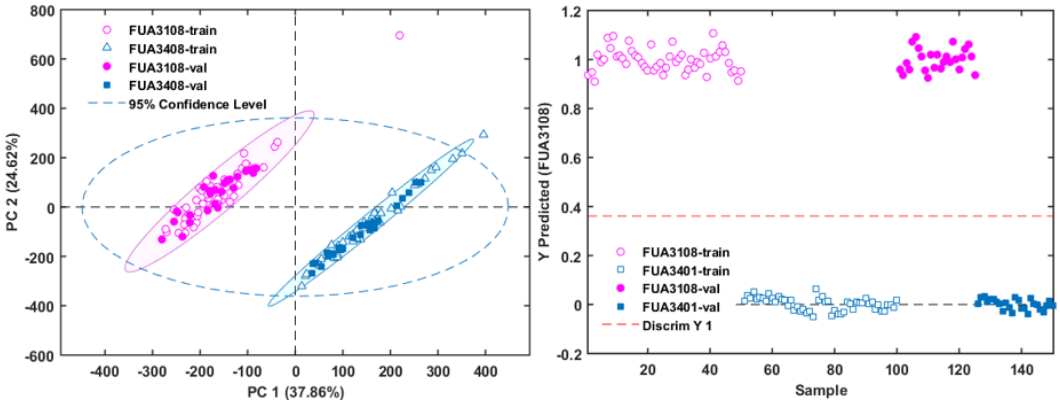


Fig. 5.3.5: PCA and PLS-DA models for Class A before feature selection. To distinguish between FUA3108 and FUA3408 using all the variables. Purple markers and blue markers represents FuA3108 and FUA3408, respectively. Hollow and filled markers represents training and validation sets, respectively.

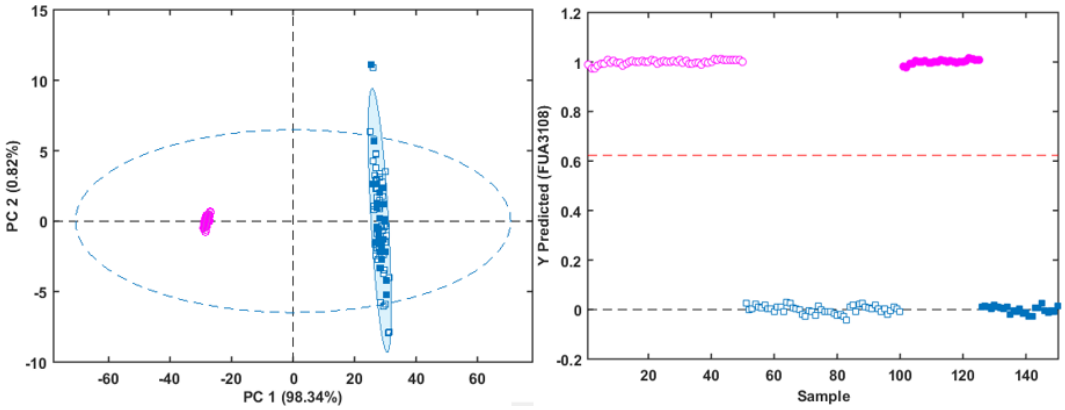


Fig. 5.3.6: PCA and PLS-DA models for Class A after feature selection. To distinguish between FUA3108 and FUA3408 using the variables retained after at an F-ratio threshold ($a = 5$).817 variables were retained. Purple and blue markers represents FuA3108 and FUA3408, respectively. Hollow and filled markers represents training and validation sets, respectively.

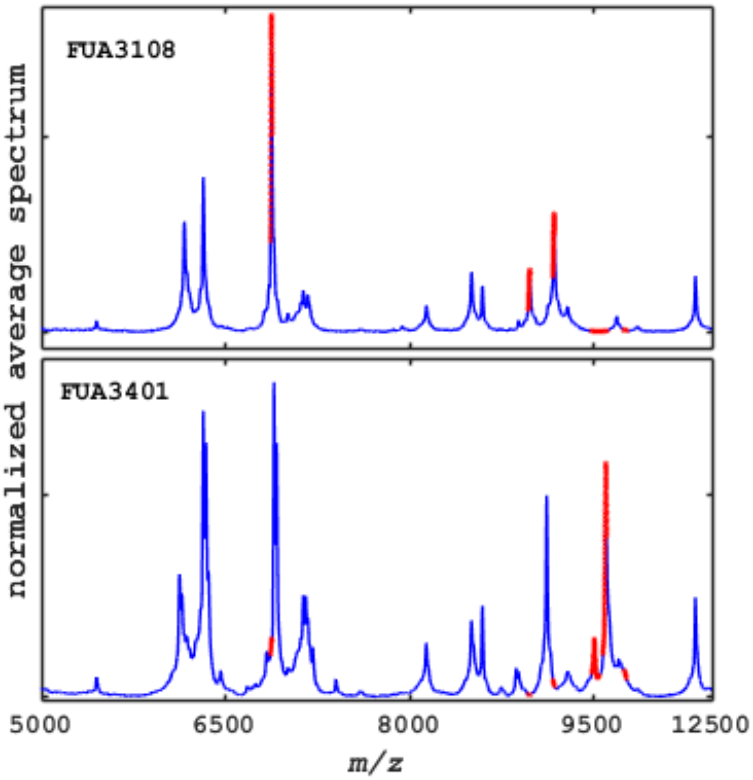


Fig. 5.3.7: Plots of average background-corrected spectra of Classes A (FUA3108 and FUA3408) showing m/z that survived the feature selection. Portions shown in red indicate variables that were retained (817 variables) after feature selection.

A similar approach to the classification of the strains of bacteria in Class A was used for Class B. This time however, the optimum value for a was 3. This led to 2757 variables being retained. PCA and PLS-DA models constructed using all the variables and the 2757 retained after feature selection are shown in Fig. 5.3.8 and Fig. 5.3.9, respectively. Once again, a much more parsimonious model with a prediction sensitivity, specificity and accuracy at par with using all the variables was obtained. A plot of the normalized average spectra for the two strains (TMW1.656 and mlc3) are shown in Fig. 5.1.10.

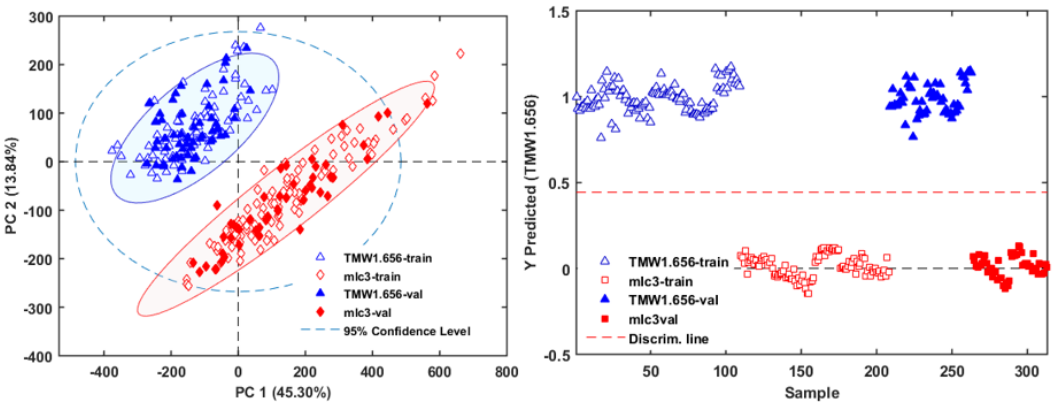


Fig. 5.3.8: PCA and PLS-DA models for Class A before feature selection. To distinguish between TMW1.656 and mlc3 using all the variables.Blue and red markers represents TMW1.656 and mlc3, respectively. The hollow and filled markers represents training and validation sets, respectively.

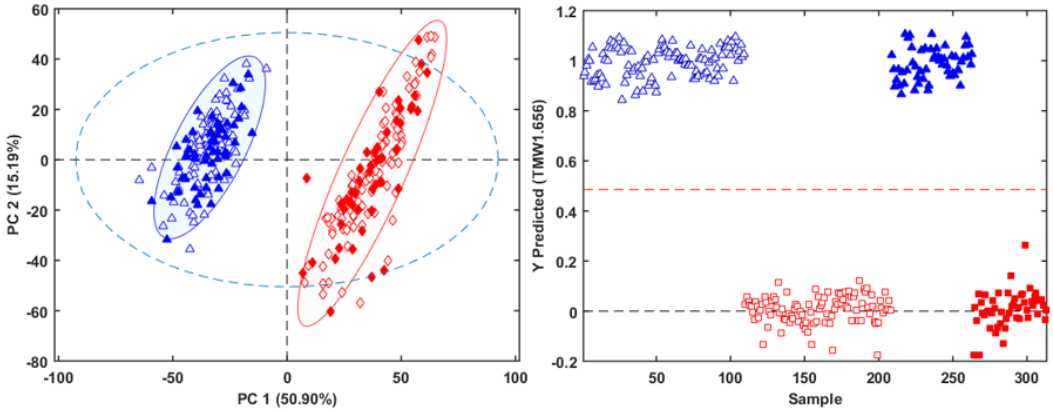


Fig. 5.3.9: PCA and PLS-DA models for Class B after feature selection. To distinguish between (TMW1.656 and mlc3) using the variables retained after an F-ratio threshold ($a = 3$). 2757 variables were retained. Blue markers and red markers represents TMW1.656 and mlc3, respectively. Hollow and filled markers represents training and validation sets, respectively.

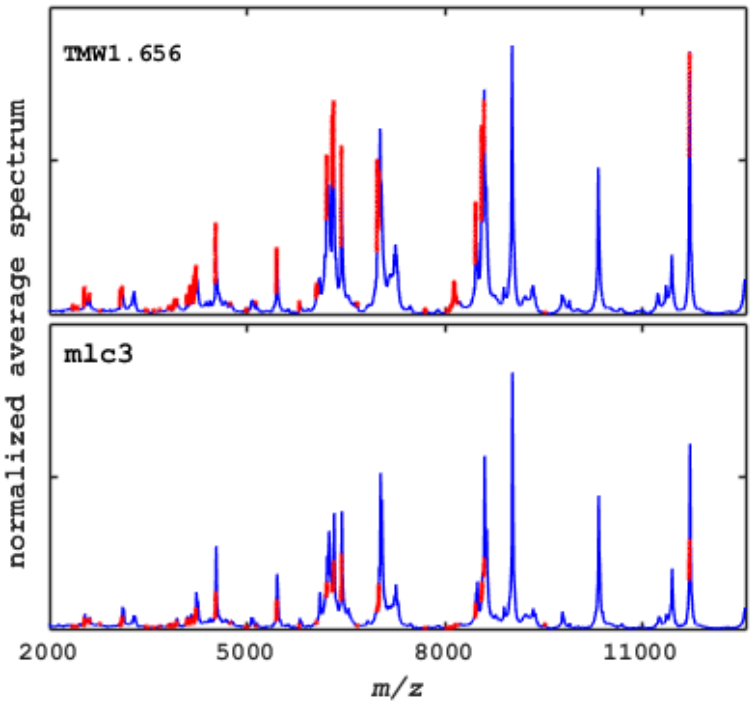


Fig. 5.3.10: Plots of average background-corrected spectra of Classes B (TMW1.656 and mlc3) showing m/z that survived the feature selection. Portions shown in red indicate variables that were retained (2757 variables).

Class C consisted of eight (8) strains with an initial PCA plot shown in Fig. 5.3.11. In PCA score space, three clusters could be identified and named Class C1 (5 strains: TMW1.112, LTH2584, FUA3168, FUA3324 and LTH5448), Class C2 (1 strain: lpuph) and Class C3 (2 strains: 100-23 and FUA3400).

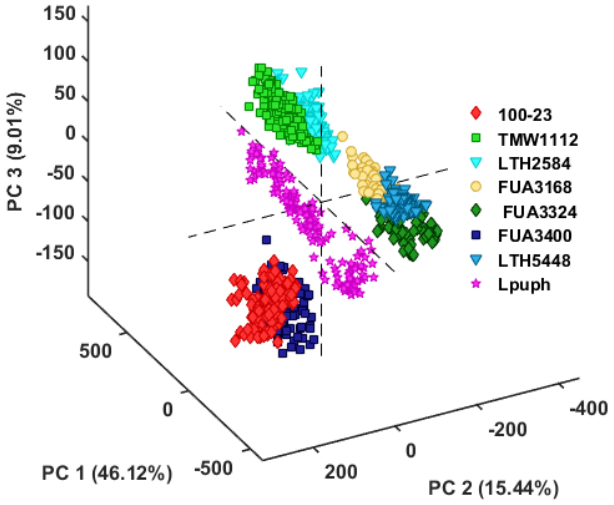


Fig. 5.3.11: PCA score plot of samples in Class C without feature selection. Class C1 (5 strains: TMW1.112, LTH2584, FUA3168, FUA3324 and LTH5448), Class C2 (1 strain: lpuph) and Class C3 (2 strains: 100-23 and FUA3400).

The PLS-DA model constructed using all the variables is shown in Fig. 5.3.12. Due to the high within class variances in the classes, CR-FS was used for feature selection which led to the retention of 2102 variables. PLS-DA model constructed with the 2102 features retained showed 100% classification accuracy (Fig.5.3.13) with fewer variables. A plot of the average spectra from C1, C2 and C3 showing the 2102 variables retained is also shown in Fig. 5.2.14.

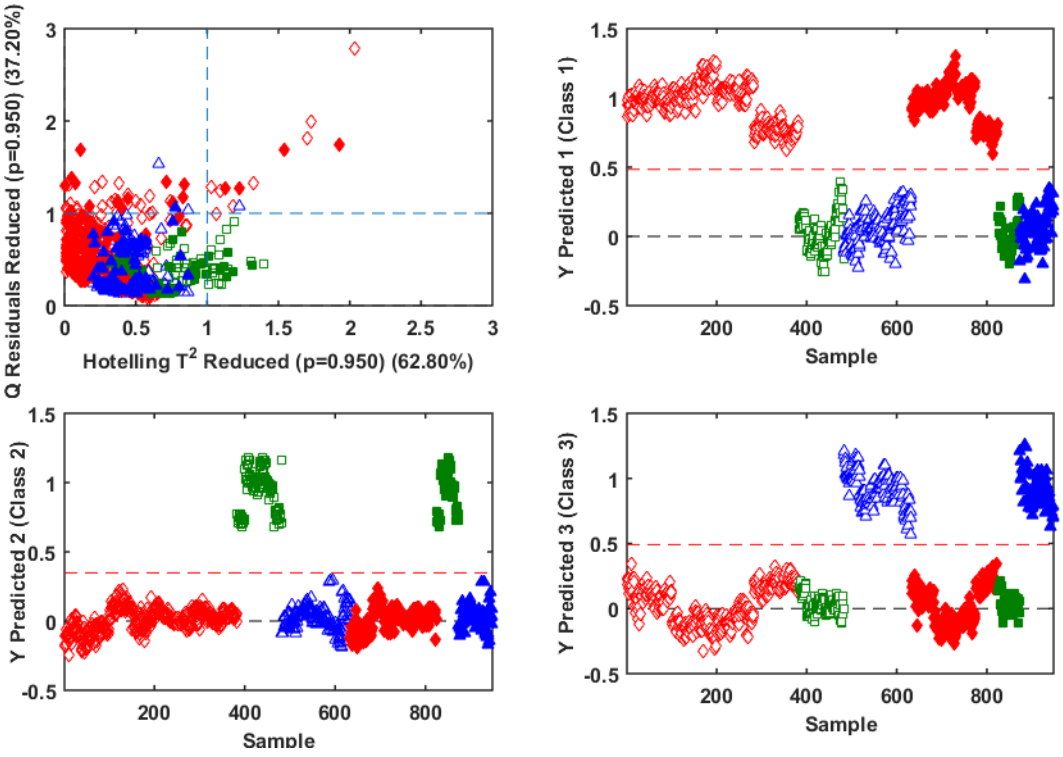


Fig. 5.3.12: PLS-DA models for Class C before feature selection. To distinguish between samples in Class C1, C2 and C3 using all the variables. Class C1 - red markers (5 strains: TMW1.112, LTH2584, FUA3168, FUA3324 and LTH5448), Class C2 - green markers (1 strain: lpuph) and Class C3 - blue markers (2 strains: 100-23 and FUA3400). Hollow and filled markers represents training and validation sets, respectively.

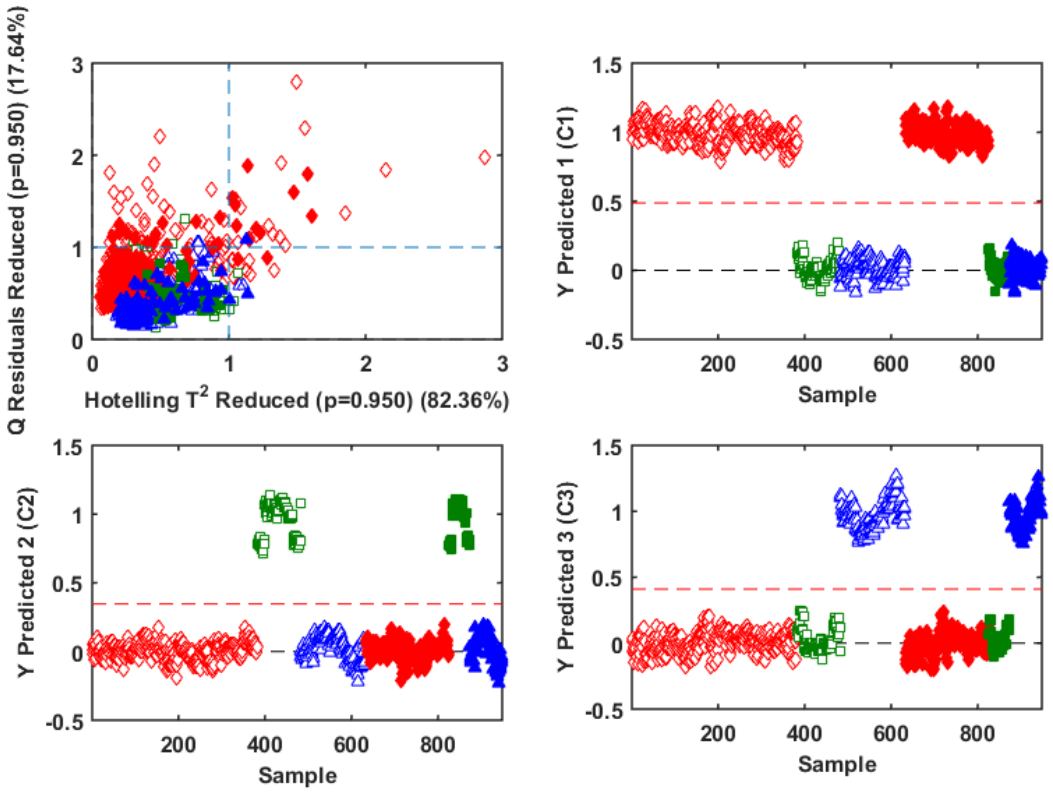


Fig. 5.3.13: PLS-DA models for Class C before feature selection. To distinguish between samples in Class C1, C2 and C3 using variables retained after feature selection with CR-FS (2102 variables). Class C1 - red markers (5 strains: TMW1.112, LTH2584, FUA3168, FUA3324 and LTH5448), Class C2 - green markers (1 strain: lpuph) and Class C3 - blue markers (2 strains: 100-23 and FUA3400). Hollow and filled markers represents training and validation sets, respectively.

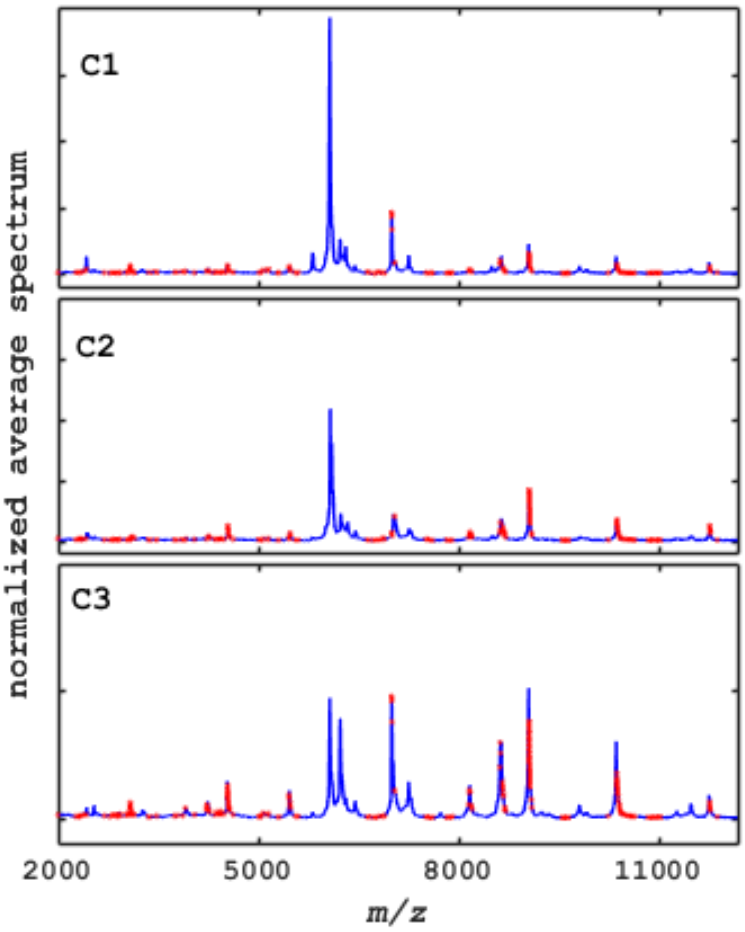


Fig. 5.3.14: Plots of average background-corrected spectra of Classes C1, C2 and C3 showing m/z that survived the feature selection. Portions shown in red indicate variables that were retained (2102 variables).

A PCA score plot of samples in Class C1 consisting of five strains showed two groups, namely, C1A (two strains: TMW1.112 and LTH2584) and C1B (three strains: FUA3168, FUA3324 and LTH5448). PCA and PLS-DA models of class C1 before feature selection is shown in Fig. 5.3.15. CR-FS was used and only 2008 variables were retained. PCA and PLS-DA models constructed using the variables retained after feature selection are shown in Fig. 5.3.16. Averaged spectrum of the samples in Class C1A and C1B showing variables retained after features selection with CR-FS is shown in Fig. 5.3.17.

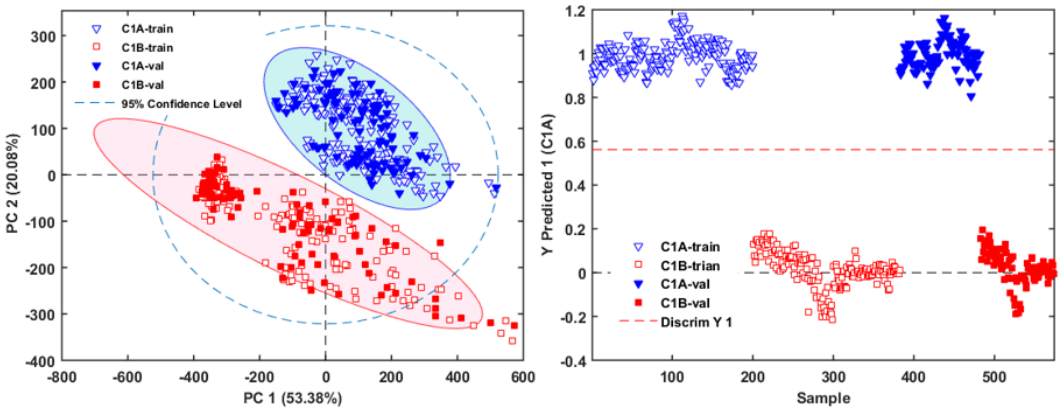


Fig. 5.3.15: PCA and PLS-DA models for Class C1 before feature selection to distinguish between Class C1A(TMW1.112 and LTH2584) and C1B (FUA3168, FUA3324 and LTH5448) using all the variables. Blue and red markers represents C1A and C1B, respectively. Hollow and filled markers represents training and validation sets, respectively.

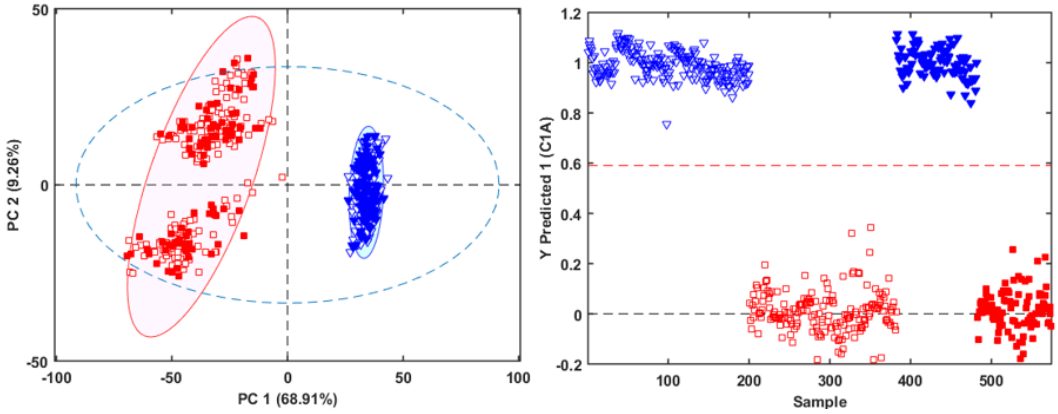


Fig. 5.3.16: PCA and PLS-DA models for Class C1 after feature selection to distinguish between Class C1A(TMW1.112 and LTH2584) and C1B (FUA3168, FUA3324 and LTH5448) using the variables retained after feature selection with CR-FS. 2008 variables were retained. Blue and red markers represents C1A and C1B, respectively. Hollow and filled markers represents training and validation sets, respectively.

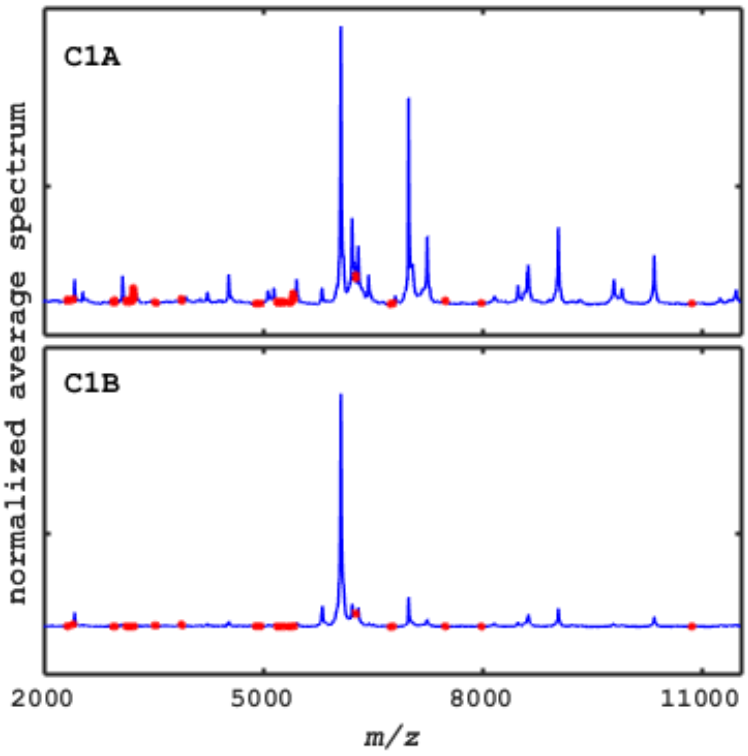


Fig. 5.3.17: Plots of average background-corrected spectra of Classes C1A and C1B showing m/z that survived the feature selection with CR-FS. Portions shown in red indicates variables that were retained (2008 variables).

Feature selection on C1A consisting of two strains (i.e., TMW1.112 and LTH2584), was performed with CR-FS algorithm. After feature selection 1115 variables were retained. Figure 5.3.18 and 5.3.19 show the PCA and PLS-DA y-predicted plot for before and after feature selection. A plot of average mass spectra showing the m/z locations of the retained features are shown in Fig. 5.3.20.

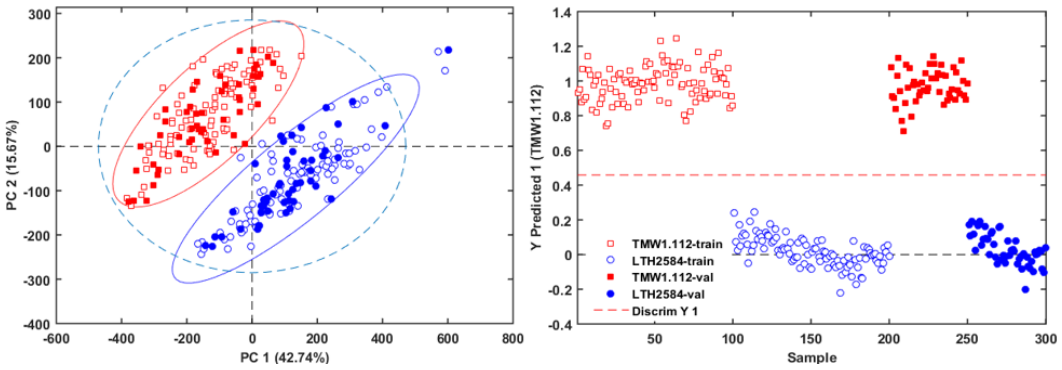


Fig. 5.3.18: PCA and PLS-DA models for Class C1A before feature selection to distinguish between TMW1.112 and LTH2584 using all the variables. Red and blue markers represents TMW1.112 and LTH2584, respectively. Hollow and filled markers represents training and validation sets, respectively.

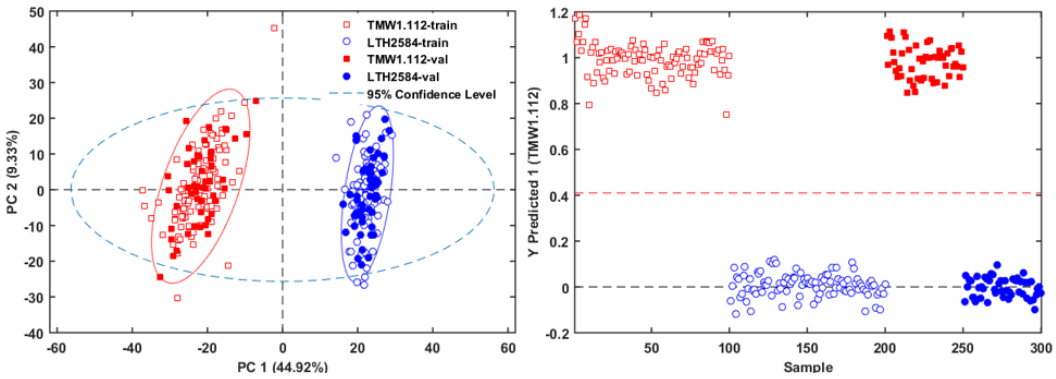


Fig. 5.3.19: PCA and PLS-DA models for Class C1A after feature selection to distinguish between TMW1.112 and LTH2584 using the variables retained after feature selection with CR-FS. 1115 variables were retained. Red and blue markers represents TMW1.112 and LTH2584, respectively. Hollow and filled markers represents training and validation sets, respectively.

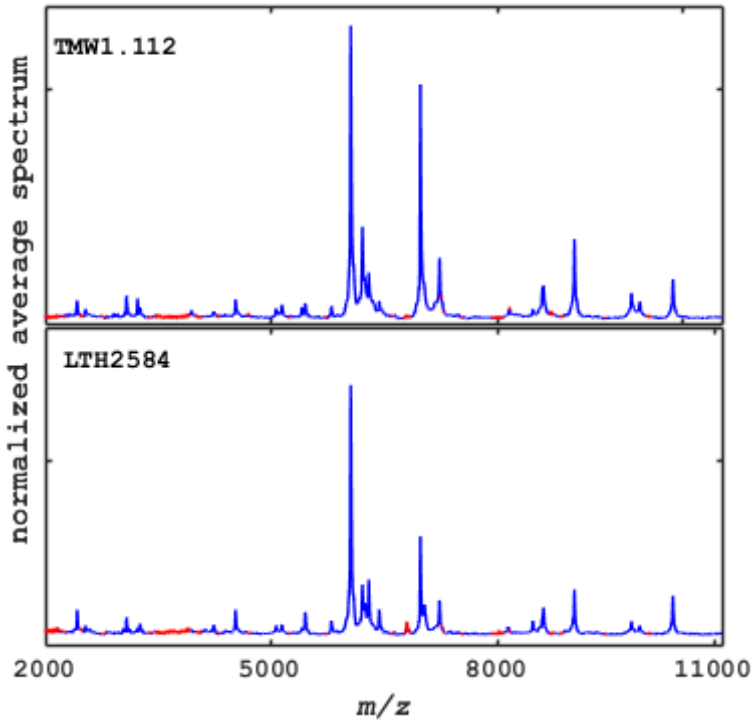


Fig. 5.3.20: Plots of average background-corrected spectra of TMW1.112 and LTH2584 showing m/z that survived the feature selection with CR-FS. Portions shown in red indicate variables that were retained (1115 variables).

Class C1B which consisted of strains FUA3168, FUA3324 and LTH5448, was also optimized using CR-FS. The number of variables retained was 2008. PLS-DA y-predicted for the three classes are shown in Fig. 5.3.21 and 5.3.22, respectively.

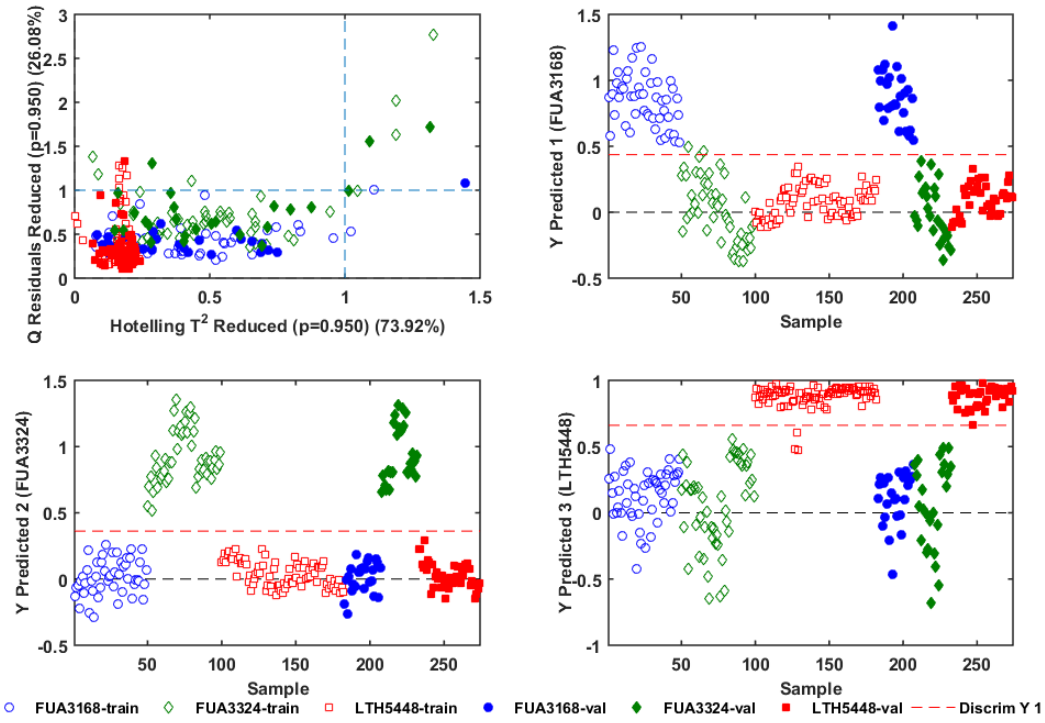


Fig. 5.3.21: PCA and PLS-DA models for Class C1A before feature selection to distinguish between FUA3168, FUA3324 and LTH5448 using all the variables. Blue, green and red markers represents, FUA3168, FUA3324 and LTH5448, respectively. Hollow and filled markers represents training and validation sets, respectively.

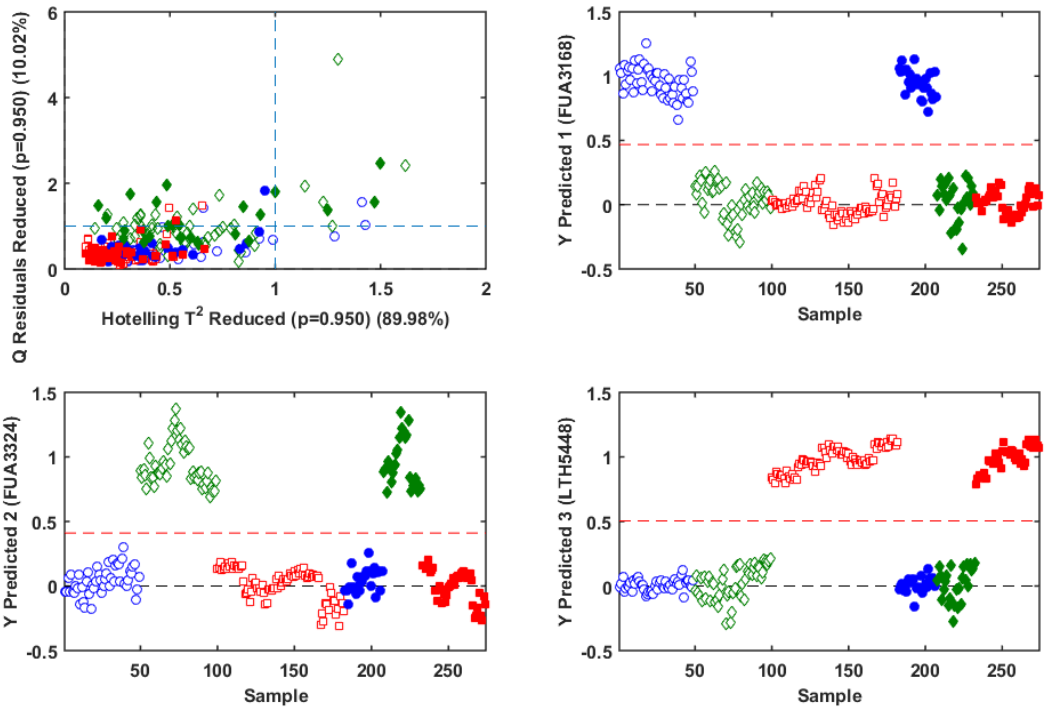


Fig. 5.3.22: PCA and PLS-DA models for Class C1A after feature selection to distinguish between FUA3168, FUA3324 and LTH5448 using the variables retained after feature selection with CR-FS. 1108 variables were retained. Blue, green and red markers represents, FUA3168, FUA3324 and LTH5448, respectively. Hollow and filled markers represents training and validation sets, respectively.

The cluster identified as Class C2 contained only one strain. Class C3 was composed of two strains (100-23 and FUA3400). Feature selection was performed with CR-FS. Results of PLS-DA models for Class C3 before and after feature selection are shown in Fig. 5.3.21 and 5.3.22, respectively. A plot of average spectra for 100-23 and FUA3400 showing the m/z locations for the 1119 variables selected are shown in Figure 5.3.23.

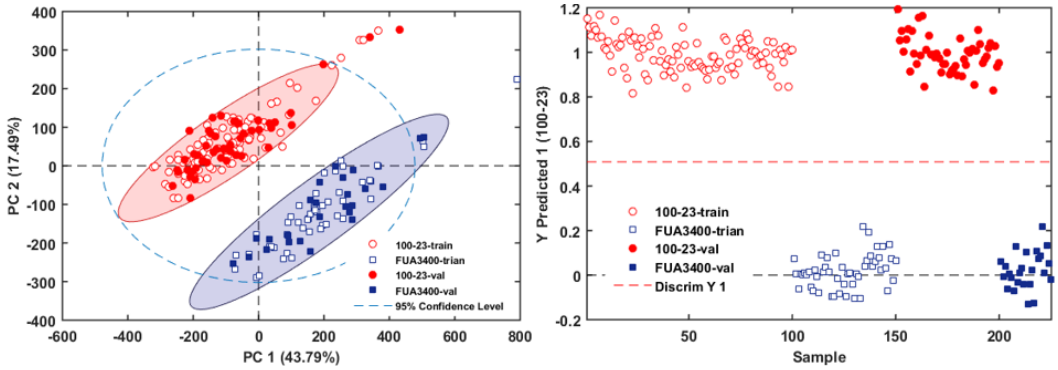


Fig. 5.3.23: PCA and PLS-DA models for Class C3 before feature selection to distinguish between 100-23 and FUA3400 using all the variables. Red and blue markers represents 100-23 and FUA3400, respectively. Hollow and filled markers represents training and validation sets, respectively.

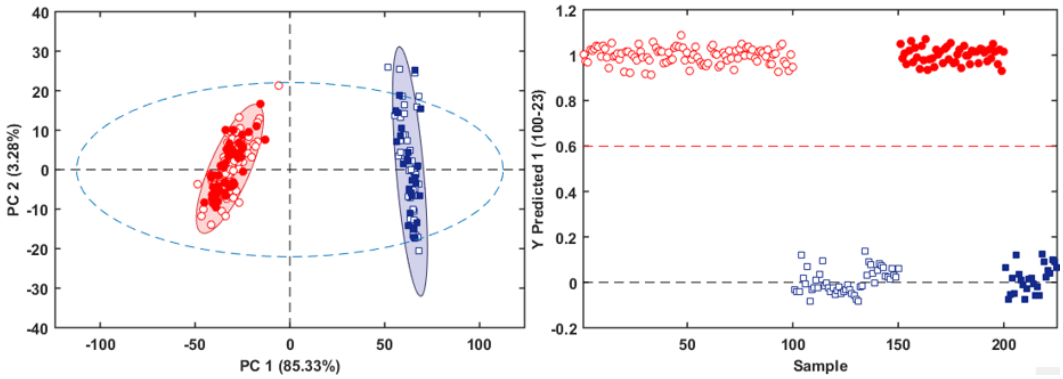


Fig. 5.3.24: PCA and PLS-DA models for Class C3 after feature selection to distinguish between 100-23 and FUA3400 using the variables retained after feature selection with CR-FS. 1119 variables were retained. Red and blue markers represents 100-23 and FUA3400, respectively. Hollow and filled markers represents training and validation sets, respectively.

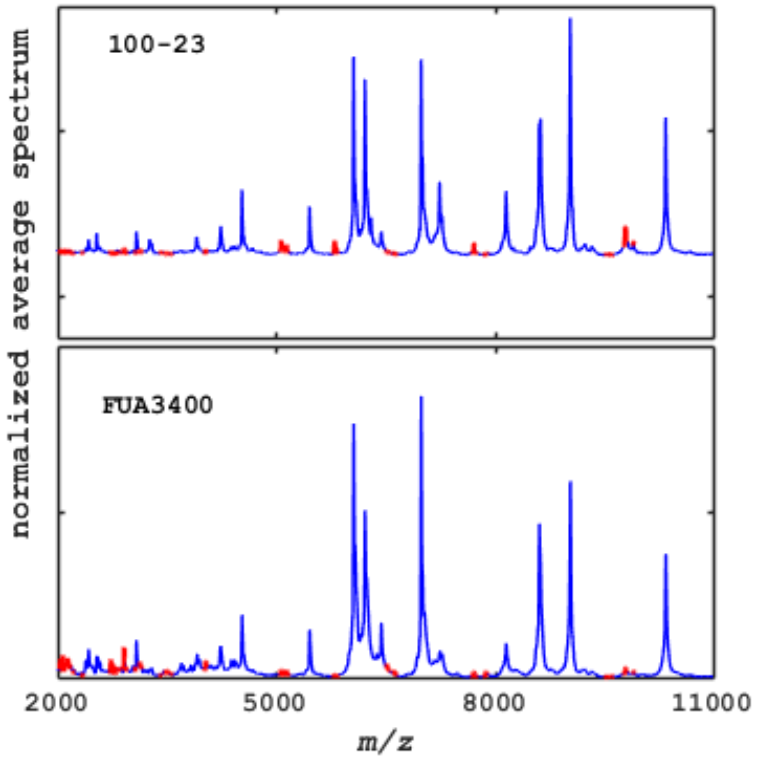


Fig. 5.3.25: Plots of average background-corrected spectra of 100-23 and FUA3400 showing m/z that survived the feature selection with CR-FS. Portions shown in red indicates variables that were retained (1119 variables).

A summary classification scheme for *Lactobacillus reuteri* strains, showing the feature selection used (i.e., F-ratio threshold or CR-FS), as well as the number of variables retained is shown in Fig. 5.3.26.

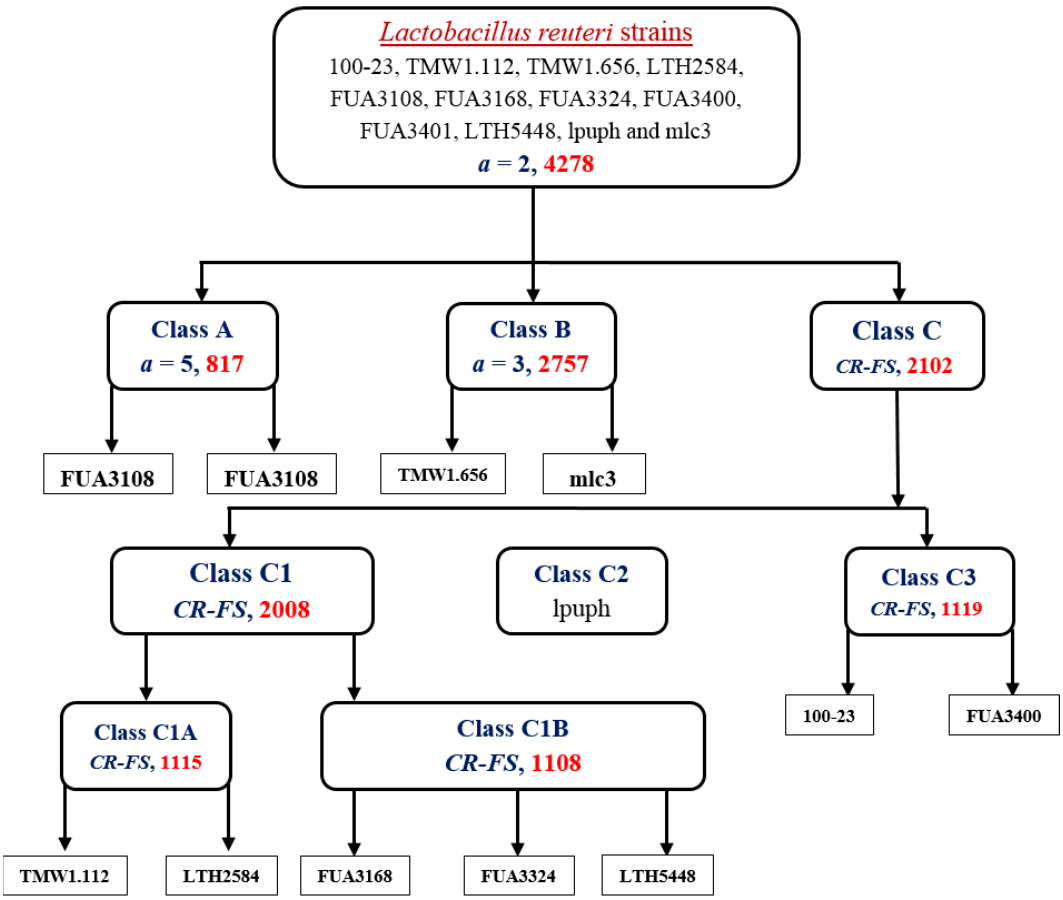


Fig. 5.3.26: Hierarchical flow chart for the classification of *Lactobacillus reutri* strains. The feature selection method used at each branching point on the flow chart is stated. Where the F-ratio is used for feature selection, the *a* value is specfied, otherwise, CR-FS is used. The number of features retained by whichever feature selection technique is shown in red.

5.3.5 Conclusion

Classification of twelve strains of *Lactobacillus reuteri* was achieved using MALDI-TOF-MS spectra. The number of variables retained for the various feature selection steps ranged from 817 to 4278. Compared to the total number of variables, this represents 0.1 - 5%. Hence a large portion of the MALDI-TOF MS spectrum contained variables irrelevant to the identification of the various strains of *Lactobacilus reutri*. Using the concept of exploratory data analysis, sub clustering allowed each strains to be classified correctly. This presents an opportunity for the creation of hierarchical model with each sub-model employing the most relevant variables. Even though this is a preliminary study, the potential of this work to contribute to an automated approach to bacterial identification using MALDI-TOF-MS is palpable.

" To every thing there is a season....".

Ecclesiastes 3

6

General Conclusions and Prospects for Future Work

6.0.1 General Conclusions

The need for feature selection prior to the application of chemometric techniques can not be over-emphasized. This means that for data sets with a large number of variables, it is beneficial to find smart ways of reducing the size of the data prior to the analysis. Smart data reduction implies reducing the data size while maintaining as much as possible the integrity of the original data.

The large data size challenge associated with the use raw GC-MS and GC×GC-MS data was addressed by the development of a data reduction strategy termed unique ion filter (UIF) in Chapter 2. UIF is a novel feature reduction approach for preprocessing of multivariate data. UIF1D and UIF2D were successfully applied to GC-MS and GC×GC-MS data, respectively. UIF eliminates redundant features and noise from the data. Consequently, feature

selection time is greatly reduced. The implementation of UIF does not alter the chemically relevant information in the data.

The challenge of the high number of variables associated with raw GC-MS data is re-visited in Chapter 3. This time a data reduction technique that builds on TIS was developed as an improved alignment-free preprocessing step for GC-MS data for fire debris analysis. STIS retains the advantages of TIS in addition to partial preservation of retention information. In general, STIS performs better than TIS. The partial preservation of retention information by STIS helps in identifying the potential source of selected variables which is not possible with TIS. Even though both TIS and STIS benefit from noise elimination, the effect is much more pronounced in TIS.

In Chapter 4, an empirical approach towards the estimation of start and stop number for the CR-FS algorithm was successfully developed. This was achieved though the analysis of true and null F-ratios obtained from a dataset for classification models. This resulted in the development of two equations to estimate the start and stop numbers for CR-FS. All but one of the parameters in this equation are obtained by comparing the PDFs of the true and null F-ratios. Through various experiments the final parameter, d, was estimated to be $0.48 \leq d \leq 0.57$. The validity of these empirical equations was experimentally confirmed.

In Chapter 5, the CR-FS algorithm was demonstrated as a feature selection algorithm which is useful in other fields of chemistry. Prediction of

crystal structure of binary (AB) and ternary (AIB) compounds were achieved by using their elemental properties. Strain level classification of *Lactobacillus reutri* was shown to be achievable with 100% prediction sensitivity, specificity and accuracy using MALDI-TOFMS data.

6.0.2 Prospects for Future Work

The completion of this dissertation does not in anyway suggest the completion of the research reported here. On the contrary, the progress made here has sprung up new research ideas to be explored.

In Chapter 3, STIS were generated using perdeuterated anchors which were added to the samples before analysis. However, it is possible to identify markers common to a particular dataset to use as a guide in the generation of STIS. This idea can be explored especially in the case of fire debris where some compounds produced as a result of pyrolysis of substrates are always present. In human metabolomic samples, there exist compounds that are always present in almost all subjects. This can be employed in an attempt to generate STIS for metabolomic data analysis. Exploration of STIS for use as a data reduction tool in GC$\times$GC-MS data is also an interesting project.

The null F-ratio analysis in the estimation of start an stop number was developed for a binary class classification problem. From the content of this dissertation, it is obvious that CR-FS is well suited for n-class problems, where $n > 2$. Hence, it is very relevant to develop an approach to the estimation of start an stop number when there are more than two classes in the dataset to be optimized.

The classification/identification of *Lactobacillus reutri* was performed on data obtained from a short period of study. The MALDI-TOF MS data were collected within the same period. Primary and secondary metabolites produced by micro-organism can be altered by the growth conditions.[252] Hence, testing the robustness of the selected features by perturbing the growth conditions could provide some useful information towards the validation of this technique.

References

[1] John Park and Steve Mackay. *Practical data acquisition for instrumentation and control systems*. Newnes, 2003.

[2] Data Acquisition Handbook. A reference for DAQ and analog & digital signal conditioning. *Measurement Computing Corporation (2004-2012)*, 2012.

[3] Bruce R. Kowalski. Chemometrics. *Analytical Chemistry*, 52(5):112R–122R, 1980.

[4] Ildiko E Frank and Bruce R Kowalski. Chemometrics. *Analytical Chemistry*, 54(5):232–243, 1982.

[5] Steven D. Brown, Stephen T. Sum, Frederic Despagne, and Barry K. Lavine. Chemometrics. *Analytical Chemistry*, 60(12):252–271, 1988.

[6] Steven D. Brown, Thomas B. Blank, Stephen T. Sum, and Lois G. Weyer. Chemometrics. *Analytical chemistry*, 66(12):315R–359R, 1994.

[7] Barry K. Lavine, Jerome Workman, Steven D. Brown, Stephen T. Sum, and Frederic Despagne. Chemometrics. *Analytical Chemistry*, 68(12):21–62, 1996.

[8] Barry K. Lavine. Chemometrics. *Analytical Chemistry*, 70(12):209–228, 1998.

[9] Barry K. Lavine. Chemometrics. *Analytical Chemistry*, 72(12):91R–97R, 2000.

[10] Barry K. Lavine and Jerome Workman. Chemometrics. *Analytical Chemistry*, 74(12):2763–2769, 2002.

[11] Barry Lavine and Jerome J. Workman. Chemometrics. *Analytical Chemistry*, 76(12):3365–3372, 2004.

[12] Barry Lavine and Jerome Workman. Chemometrics. *Analytical Chemistry*, 78(12):4137–4145, 2006.

[13] Barry Lavine and Jerome Workman. Chemometrics. *Analytical Chemistry*, 80(12):4519–4531, 2008.

[14] Barry K. Lavine and Jerome Workman. Chemometrics. *Analytical Chemistry*, 85(2):705–714, 2013.

[15] Isabelle Guyon. An Introduction to Variable and Feature Selection 1 Introduction. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[16] Isabelle Guyon and Andre Elisseeff. Feature Extraction, Foundations and Applications: An introduction to feature extraction. *Studies in Fuzziness and Soft Computing*, 207:1–25, 2006.

[17] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS one*, 6(12):e28210, 2011.

[18] Nikolai A. Sinkov and James J. Harynuk. Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta*, 83(4):1079–87, 2011.

[19] Nikolai A. Sinkov and James J. Harynuk. Three-dimensional cluster resolution for guiding automatic chemometric model optimization. *Talanta*, 103:252–259, 2013.

[20] Nikolai A. Sinkov, P. Mark L. Sandercock, and James J. Harynuk. Chemometric classification of casework arson samples based on gasoline content. *Forensic Science International*, 235:24–31, 2014.

[21] Nikolai A. Sinkov, Brandon M. Johnston, P. Mark L. Sandercock, and James J. Harynuk. Automated optimization and construction of chemometric models based on highly variable raw chromatographic data. *Analytica Chimica Acta*, 697(1-2):8–15, 2011.

[22] Svante Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995.

[23] Desiré Luc Massart, Bernard G. Vandeginste, L. M. C. Buydens, P. J. Lewi, S. de Jong and J Smeyers-Verbeke. *Handbook of chemometrics and qualimetrics: Part A*. Elsevier Science Inc., Amsterdam, 1st Edition, 1997.

[24] Desiré Luc Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman. *Chemometrics: a textbook*. Elsevier Amsterdam, Amsterdam, Volume 2, 5th Edition, 1988.

[25] Mathias Otto. *Chemometrics, Statistics and Computer Application in Analytical Chemistry*. Wiley VCH, Weinheim, 2nd Edition, 2007.

[26] Richard G. Brereton. *Applied chemometrics for scientists*. John Wiley & Sons, West Sussex, 1st Edition, 2007.

[27] Laura R. Snyder, Jamin C. Hoggard, Thomas J. Montine, and Robert E. Synovec. Development and application of a comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method for the analysis of l-ß-methylamino-alanine in human tissue. *Journal of Chromatography A*, 1217(27):4639–4647, 2010.

[28] Piotr S. Gromski, Howbeer Muhamadali, David I. Ellis, Yun Xu, Elon Correa, Michael L. Turner, and Royston Goodacre. A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879:10–23, 2015.

[29] Ian D. Wilson, Robert Plumb, Jennifer Granger, Hilary Major, Rebecca Williams, and Eva M. Lenz. HPLC-MS-based methods for the study of metabonomics. *Journal of Chromatography. B*, 817(1):67–76, 2005.

[30] E. Szymanska, M. J. Markuszewski, X. Capron, A. M. van Nederkassel, Y. Vander Heyden, M. Markuszewski, K. Krajka, and R. Kaliszan. Increasing conclusiveness of metabonomic studies by cheminformatic preprocessing of capillary electrophoretic data on urinary nucleoside profiles. *Journal of Pharmaceutical and Biomedical Analysis*, 43(2):413–420, 2007.

[31] Tobias Kind, Vladimir Tolstikov, Oliver Fiehn, and Robert H. Weiss. A comprehensive urinary metabolomic approach for identifying kidney cancerr. *Analytical Biochemistry*, 363(2):185–95, 2007.

[32] Xianlin Han, Steve Rozen, Stephen H. Boyle, Caroline Hellegers, Hua Cheng, James R. Burke, Kathleen A. Welsh-Bohmer, P. Murali Doraiswamy, and Rima Kaddurah-Daouk. Metabolomics in early Alzheimer's disease: Identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS ONE*, 6(7), 2011.

[33] Maud M. Koek, Frans M. van der Kloet, Robert Kleemann, Teake Kooistra, Elwin R. Verheij, and Thomas Hankemeier. Semi-automated non-target processing in GC × GC-MS metabolomics analysis: Applicability for biomedical studies. *Metabolomics*, 7:1–14, 2011.

[34] Kishore Kumar Pasikanti, Kesavan Esuvaranathan, Yanjun Hong, Paul C Ho, Ratha Mahendran, Lata Raman Nee Mani, Edmund Chiong, and Eric Chun Yong Chan. Urinary metabotyping of bladder cancer using two-dimensional gas chromatography time-of-flight mass spectrometry. *Journal of Proteome Research*, 12(9):3865–73, 2013.

[35] David G. Stork, Richard O. Duda, Peter E. Hart. *Pattern Classification*. Wiley VCH, New York, 2nd Edition, 2000.

[36] Ying Wu Wesam, Ashour Barbakh and Colin Fyfe. Review of Clustering Algorithms. In *Non-standard parameter adaptation for exploratory data analysis. Vol. 249*, 7–28. Springer-Verlag, Heidelberg, 2009.

[37] Barry K. Lavine. Source identification of underground fuel spills by pattern recognition analysis. *Analytical Chemistry*, 67(27):3846–3852, 1995.

[38] P. Mark L. Sandercock and Eric Du Pasquier. Chemical fingerprinting of unevaporated automotive gasoline samples. *Forensic Science International*, 134(1):1–10, 2003.

[39] Philip Doble, P. Mark L. Sandercock, Eric Du Pasquier, Peter Petocz, Claude Roux and Michael Dawson. Classification of premium and regular gasoline by gas chromatography/mass spectrometry, principal component analysis and artificial neural networks. *Forensic Science International*, 132(1):26–39, 2003.

[40] P. Mark. L. Sandercock and Eric Du Pasquier. Chemical fingerprinting of gasoline. 2. Comparison of unevaporated and evaporated automotive gasoline samples. *Forensic Science International*, 140(1):43–59, 2004.

[41] P. Mark. L. Sandercock and Eric Du Pasquier. Chemical fingerprinting of gasoline. Part 3. Comparison of unevaporated automotive gasoline samples from Australia and New Zealand. *Forensic Science International*, 140(1):71–77, 2004.

[42] Robert K. Nelson, Brian M. Kile, Desiree L. Plata, Sean P. Sylva, Li Xu, Christopher M. Reddy, Richard B. Gaines, Glenn S. Frysinger and Stephen E. Reichenbach. Tracking the weathering of an oil spill with comprehensive two-dimensional gas chromatography. *Environmental Forensics*, 7(1):33–44, 2006.

[43] Jan H. Christensen and Giorgio Tomasi. Practical aspects of chemometrics for oil spill fingerprinting. *Journal of Chromatography. A*, 1169(1-2):1–22, 2007.

[44] C. Pizarro, I. Esteban-Díez, C. Sáenz-González and J. M. González-Sáiz. Vinegar classification based on feature extraction and selection from headspace solid-phase microextraction/gas chromatography volatile analyses: a feasibility study. *Analytica Chimica Acta*, 608(1):38–47, 2008.

[45] Berhane T. Weldegergis and Andrew M. Crouch. Analysis of volatiles in Pinotage wines by stir bar sorptive extraction and chemometric profiling. *Journal of Agricultural and Food chemistry*, 56(21):10225–36, 2008.

[46] Davide Ballabio, Thomas Skov, Riccardo Leardi and Rasmus Bro. Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques. *Journal of Chemometrics*, 22(8):457–463, 2008.

[47] Xiang Li, Zhiliang Xu, Xin Lu, Xuehui Yang, Peiyuan Yin, Hongwei Kong, Ying Yu and Guowang. Xu. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: Biomarker discovery for diabetes mellitus. *Analytica Chimica Acta*, 633(2):257–62, 2009.

[48] Andrew C. Beckstrom, Elizabeth M. Humston, Laura R. Snyder, Robert E. Synovec and Sandra E. Juul. Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model. *Journal of Chromatography A*, 1218(14):1899–1906, 2011.

[49] Elisa Pietracci, Ana María Bermejo, Iván Álvarez, Pamela Cabarcos, Walter Balduini and María-Jesús Tabernero. Simultaneous determination of new-generation antidepressants in plasma by gas chromatography–mass spectrometry. *Forensic Toxicology*, 31(1):124–132, 2012.

[50] Haidy A Gad, Sherweit H El-Ahmady, Mohamed I Abou-Shoer and Mohamed M Al-Azizi. Application of chemometrics in authentication of herbal medicines: a review. *Phytochemical analysis : PCA*, 24(1):1–24, 2013.

[51] Bozena M Lukasiak, Rita Faria, Simeone Zomer, Richard G Brereton and John C Duncan. Pattern recognition for the analysis of polymeric materials. *The Analyst*, 131(1):73–80, 2006.

[52] J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *Computers, IEEE Transactions on*, C-23(9):881–890, 1974.

[53] James R. Beniger, Vic Barnett and Toby Lewis. Outliers in Statistical Data. *Contemporary Sociology*, 9(4):560, 1980.

[54] Daniel A. Keim and Hans-Peter Kriegel. Visualization Techniques for Mining Large Databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923–938, 1996.

[55] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[56] N. Jardine and R. Sibson. The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, 11(2):177–184, 1968.

[57] Erin E. Waddell, Jessica L Frisch-Daiello, Mary R. Williams and Michael E. Sigman. Hierarchical Cluster Analysis of Ignitable Liquids Based on the Total Ion Spectrum. *Journal of Forensic Sciences*, 59(5):1198–1204, 2014.

[58] Trevor Hastie and Werner Stuetzle. Principal Curves. *Journal of the American Statistical Association*, 84(406):502, 1989.

[59] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[60] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[61] Svant Wold. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37–52, 1987.

[62] Wei Guan, Manshui Zhou, Christina Y. Hampton, Benedict B. Benigno, L. Deette Walker, Alexander Gray, John F McDonald and Facundo M Fernández. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10:259, 2009.

[63] Sarah J. Dixon and Richard G. Brereton. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on. *Chemometrics and Intelligent Laboratory Systems*, 95(1):1–17, 2009.

[64] Ildiko E. Frank and Jerome H. Friedman. Classification: Oldtimers and newcomers. *Journal of Chemometrics*, 3(3):463–475, 1989.

[65] Chih Jen Lin and Ruby C. Weng. Simple probabilistic predictions for support vector regression. Technical report, National Taiwan University, 2004.

[66] Tze-wey Loong. Clinical review Understanding sensitivity and specificity with the right. *Bmj*, 327(September):716–19, 2003.

[67] Jasper Engel, Jan Gerretzen, Ewa Szymańska, Jeroen J. Jansen, Gerard Downey, Lionel Blanchet and Lutgarde M C Buydens. Breaking with trends in pre-processing? *TrAC - Trends in Analytical Chemistry*, 50:96–106, 2013.

[68] Rasmus Bro and Age K. Smilde. Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1):16–33, 2003.

[69] F. Cuesta Sánchez, P. J. Lewi and D. L. Massart. Effect of different preprocessing methods for principal component analysis applied to the composition of mixtures: Detection of impurities in HPLC-DAD. *Chemometrics and Intelligent Laboratory Systems*, 25(2):157–177, 1994.

[70] David E. Axelson. *Data Preprocessing for Chemometrics and Metabonomic Analysis*. 2010.

[71] Margaret M. W. B. Hendriks, Leyre Cruz-Juarez, Dries De Bont and Robert D. Hall. Preprocessing and exploratory analysis of chromatographic profiles of plant extracts. *Analytica Chimica Acta*, 545(1):53–64, 2005.

[72] Andrew Craig, Olivier Cloarec, Elaine Holmes, Jeremy K. Nicholson and John C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267, 2006.

[73] Wiktoria Struck, Paweł Wiczling, Małgorzata Waszczuk-Jankowska, Roman Kaliszan and Michał Jan Markuszewski. New supervised alignment method as a preprocessing tool for chromatographic data in metabolomic studies. *Journal of Chromatography A*, 1256:150–159, 2012.

[74] Mikko Katajamaa and Matej Orešič. Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, 1158(1-2):318–328, 2007.

[75] David C Stone. Application of median filtering to noisy data. *Canadian Journal of chemistry*, 73(10):1573–1581, 1995.

[76] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[77] Manfred U. a. Bromba and Horst. Ziegler. Application hints for Savitzky-Golay digital smoothing filters. *Analytical Chemistry*, 53(11):1583–1586, 1981.

[78] David A. McNulty and Halliday J. H. MacFie. The effect of different baseline estimators on the limit of quantification in chromatography. *Journal of Chemometrics*, 11:1–11, 1997.

[79] Zhi-Min Zhang, Shan Chen and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *The Analyst*, 135(5):1138–46, 2010.

[80] Ł Komsta. Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia*, 73:721–731, 2011.

[81] Zhi Min Zhang and Yi Zeng Liang. Comments on the baseline removal method based on quantile regression and comparison of several methods. *Chromatographia*, 75(5-6):313–314, 2012.

[82] Georg Schulze, Andrew Jirasek, Marcia M. L. Yu, Arnel Lim, Robin F. B. Turner and Michael W. Blades. Investigation of selected baseline removal techniques as candidates for automated implementation. *Applied Spectroscopy*, 59(5):545–574, 2005.

[83] Feng Gan, Guihua Ruan and Jinyuan Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):59–65, 2006.

[84] Mauro Mecozzi. A Polynomial Curve Fitting Method for Baseline Drift Correction in the Chromatographic Analysis of Hydrocarbons in Environmental Samples. *APCBEE Procedia*, 10:2–6, 2014.

[85] Shao Xueguang, Wensheng Cai and Pan Zhongxiap. Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis. *Chemometrics and Intelligent Laboratory Systems*, 45:249–256, 1999.

[86] Xue-Guang Shao, Alexander Kai-Man Leung and Foo-Tim Chau. Wavelet: a new trend in chemistry. *Accounts of Chemical Research*, 36(4):276–283, 2003.

[87] Xiao-Guo Ma and Zhan-Xia Zhang. Application of wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry. *Analytica Chimica Acta*, 485:233–239, 2003.

[88] Hans F. M. Boelens, Reyer J. Dijkstra, Paul H. C. Eilers, Fiona Fitzpatrick and Johan A. Westerhuis. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *Journal of Chromatography A*, 1057(1-2):21–30, 2004.

[89] Martin Lopatka, Andrei Barcaru, Marjan J. Sjerps and Gabriel Vivo-Truyols. Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples. *Journal of Chromatography A*, 1431:122–130, 2016.

[90] A. M. van Nederkassel, C. J. Xu, P. Lancelin, M. Sarraf, D. A. MacKenzie, N. J. Walton, F. Bensaid, M. Lees, G. J. Martin, J. R. Desmurs, D. L. Massart, J. Smeyers-Verbeke and Y. Vander Heyden. Chemometric treatment of vanillin fingerprint chromatograms. Effect of different signal alignments on principal component analysis plots. *Journal of Chromatography A*, 1120(1-2):291–298, 2006.

[91] Trond Brekke, Olav M. Kvalheim and Einar Sletten. Prediction of physical properties of hydrocarbon mixtures by partial-least-squares calibration of carbon-13 nuclear magnetic resonance data. *Analytica Chimica Acta*, 223(C):123–134, 1989.

[92] Jenny Forshed, Ina Schuppe-Koistinen and Sven P. Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487(2):189–199, 2003.

[93] Jenny Forshed, Ralf J O Torgrip, K. Magnus Åberg, Bo Karlberg, Johan Lindberg and Sven P. Jacobsson. A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 38(5):824–832, 2005.

[94] Bradley Worley and Robert Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.

[95] Nils Hoffmann, Matthias Keck, Heiko Neuweger, Mathias Wilhelm, Petra Högy, Karsten Niehaus and Jens Stoye. Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC Bioinformatics*, 13(1):214, 2012.

[96] Melissa D. Krebs, Robert D. Tingley, Julie E. Zeskind, Maria E. Holmboe, Joung-Mo Kang and Cristina E. Davis. Alignment of gas chromatography–mass spectrometry data by landmark selection from complex chemical mixtures. *Chemometrics and Intelligent Laboratory Systems*, 81(1):74–81, 2006.

[97] Mark P. Styczynski, Joel F. Moxley, Lily V. Tong, Jason L. Walther, Kyle L. Jensen and Gregory N. Stephanopoulos. Systematic Identification of Conserved Metabolites in GC/MS Data for Metabolomics and Biomarker Discovery. *Analytical Chemistry*, 79(3):966–973, 2007.

[98] N. P. V. Nielsen, Jens Michael Carstensen and Jørn Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805:17–35, 1998.

[99] V. Pravdova, B. Walczak and D. L. Massart. A comparison of two algorithms for warping of analytical signals. *Analytica Chimica Acta*, 456(1):77–92, 2002.

[100] Giorgio Tomasi, Frans Van Den Berg and Claus Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18(5):231–241, 2004.

[101] F. Savorani, G. Tomasi and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[102] Jason W H Wong, Caterina Durante and Hugh M. Cartwright. Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Analytical Chemistry*, 77(17):5655–5661, 2005.

[103] Beata Walczak and Wen Wu. Fuzzy warping of chromatograms. *Chemometrics and Intelligent Laboratory Systems*, 77(1–2):173–180, 2005.

[104] Xiang Zhang, John M Asara, Jiri Adamec, Mourad Ouzzani and Ahmed K Elmagarmid. Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics*, 21(21):4054–4059, 2005.

[105] Karisa M. Pierce, Bob W. Wright and Robert E. Synovec. Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatographic data using the piecewise alignment algorithm. *Journal of Chromatography A*, 1141(1):106–116, 2007.

[106] David Clifford, Glenn Stone, Ivan Montoliu, Serge Rezzi, Philippe Guy, Stephen Bruce and Sunil Kochhar. Alignment Using Variable Penalty Dynamic Time Warping. *Analytical chemistry*, 81(3):1000–1007, 2009.

[107] A. M. van Nederkassel, M. Daszykowski, P. H C Eilers and Y. Vander Heyden. A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118(2):199–210, 2006.

[108] Behrooz Zekavat and Touradj Solouki. Chemometric data analysis for deconvolution of overlapped ion mobility profiles. *Journal of the American Society for Mass Spectrometry*, 23(11):1873–1884, 2012.

[109] Song Yang, Jeremy S Nadeau, Elizabeth M Humston-Fulmer, Jamin C Hoggard, Mary E Lidstrom and Robert E Synovec. Gas chromatography-mass spectrometry with chemometric analysis for determining ^{12}C and ^{13}C labeled contributions in metabolomics and ^{13}C flux analysis. *Journal of chromatography. A*, 1240:156–64, 2012.

[110] Chao Yang, Zengyou He and Weichuan Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC bioinformatics*, 10(1):4, 2009.

[111] Armin Schwartzman, Yulia Gavrilov and Robert J. Adler. Multiple testing of local maxima for detection of peaks in 1D. *Annals of Statistics*, 39(6):3290–3319, 2011.

[112] S. V. Chekanov and M. Erickson. A nonparametric peak finder algorithm and its application in searches for new physics. *Advances in High Energy Physics*, 2013, 2013.

[113] Felix Scholkmann, Jens Boss and Martin Wolf. An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. *Algorithms*, 5(4):588–603, 2012.

[114] M. Lopatka, G. Vivó-Truyols and M. J. Sjerps. Probabilistic peak detection for first-order chromatographic data. *Analytica chimica acta*, 817:9–16, 2014.

[115] Robert A van den Berg, Huub C J Hoefsloot, Johan a Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7:142, 2006.

[116] Svante Wold, Michael Sjöström and Lennart Eriksson. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[117] Joanet Maree, Guy Kamatou, Simon Gibbons, Alvaro Viljoen and Sandy Van Vuuren. The application of GC–MS combined with chemometrics for the identification of antimicrobial compounds from selected commercial essential oils. *Chemometrics and Intelligent Laboratory Systems*, 130:172–181, 2014.

[118] Geore H John and Ron Kohavi. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[119] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[120] Mehmed Kantardzic. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[121] N Sánchez-Maroño, A Alonso-Betanzos and M Tombilla-Snaromán. Filter methods for feature selection–a comparative study. *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pages 178–187, 2007.

[122] G. H. John, Ron Kohavi and K. Pfleger. Irrelevant features and the subset selection problem. In *11th International Conference on Machine Learning*, 121–129, New Brunswick, NJ, 1994.

[123] R. A. Caruana and D. Freitag. How useful is relevance? In *AAAI Fall Symposium on Relevance*, pages 25–29, New Orleans, LA, 1994.

[124] José M. Cadenas, M. Carmen Garrido and Raquel Martínez. Feature subset selection filter–wrapper based on low quality data. *Expert Systems with Applications*, 40(16):6241–6252, 2013.

[125] Kenji Kira and Larry A. Rendell. The feature selection problem: Traditional methods and a new algorithm. *Aaai -* , 129 – 134, 1992.

[126] Hussein Almuallim and Thomas G Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1-2):279–305, 1994.

[127] Mark Hall. Correlation-based Feature Selection for Machine Learning. *Methodology*, 21i195-i20:1–5, 1999.

[128] Brian P. Flannery, Saul A. Teukolsky, William H. Press and William T. Vetterling. *Numerical Recipes in C - The Art of Scientific Computing*. Cambridge University Press, 2nd Edition, 1992.

[129] Zheng Zhao and Huan Liu. Searching for Interacting Features. In *IJCAI International Joint Conference on Artificial Intelligence*, 1156–1161, 2007.

[130] Jinjie Huang, Yunze Cai, and Xiaoming Xu. A Hybrid Genetic Algorithm for Feature Selection Wrapper based on Mutual Information. *Pattern Recognition Letters*, 28(13):1825–1844, 2007.

[131] L Zhuo, J Zheng, F Wang, X Li, B Ai, and J Qian. A Genetic Algorithm Based Wrapper Feature Selection Method for Classification of Hyperspectral images using support vector machine. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37:397–402, 2008.

[132] Othman Soufan, Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B. Bajic. DWFS: A wrapper feature selection tool based on a parallel Genetic Algorithm. *PLoS ONE*, 10(2), 2015.

[133] Nabil M. Hewahi and Eyad A. Alashqar. Wrapper Feature Selection based on Genetic Algorithm for Recognizing Objects from Satellite Imagery. *Journal of Information Technology*, 8(3), 1–20 2015.

[134] Béatrice Duval and Jin-Kao Hao. Advances in Metaheuristics for Gene Selection and Classification of Microarray data. *Briefings in Bioinformatics*, 11(1):127–41, 2010.

[135] Susana M. Vieira, João M. C. Sousa and Uzay Kaymak. Fuzzy Criteria for Feature Selection. *Fuzzy Sets and Systems*, 189(1):1–18, 2012.

[136] Caitlin N. Rinke, Mary R. Williams, Christopher Brown, Matthieu Baudelet, Martin Richardson and Michael E Sigman. Discriminant analysis in the presence of interferences: Combined application of target factor analysis and a Bayesian soft-classifier. *Analytica Chimica Acta*, 753:19–26, 2012.

[137] Mireia Farrés, Stefan Platikanov, Stefan Tsakovski and Romà Tauler. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, 29(10):528–536, 2015.

[138] Tarja Rajalahti, Reidar Arneberg, Ann C. Kroksveen, Magnus Berle, Kjell-Morten Myhr and Olav M Kvalheim. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical chemistry*, 81(7):2581–2590, 2009.

[139] André S Fialho, Federico Cismondi, Susana M Vieira, João MC Sousa, Shane R Reti, Michael D Howell and Stan N Finkelstein. Predicting outcomes of septic shock patients using feature selection based on soft computing techniques. *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, 65–74, 2010.

[140] Lawrence A. Adutwum and James J. Harynuk. Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis. *Analytical Chemistry (Washington, DC, United States)*, 86(15):7726–7733, 2014.

[141] Anton O. Oliynyk, Lawrence A. Adutwum, James J. Harynuk and Arthur Mar. Classifying crystal structures of binary compounds AB through cluster resolution feature selection and support vector machine analysis. *Chemistry of Materials*, 28(18):6672–6681, 2016.

[142] Michael E. Sigman, Mary R. Williams, Joseph A. Castelbuono, Joseph G. Colca and C. Douglas. Clark. Ignitable Liquid Classification and Identification Using the Summed-Ion Mass Spectrum. *Instrumentation Science & Technology*, 36(4):375–393, 2008.

[143] J. Calvin Giddings. Two-dimensional separations: concept and promise. *Analytical Chemistry*, 56(12):1258A–1270A, 1984.

[144] John B. Phillips and Jingzhen Xu. Comprehensive multi-dimensional gas chromatography. *Journal of Chromatography A*, 703(1–2)::327–334, 1995.

[145] Russell M. Kinghorn and Philip J. Marriott. Comprehensive two-dimensional gas chromatography using a modulating cryogenic trap. *Journal of Separation Science*, 21(11):620–622, 1998.

[146] John V. Seeley, Frederick Kramp and Christine J. Hicks. Comprehensive two-dimensional gas chromatography via differential flow modulation. *Analytical Chemistry*, 72(18):4346–4352, 2000.

[147] Eric Stauffer and John J. Lentini. ASTM standards for fire debris analysis: A review. *Forensic Science International*, 132(1):63–67, 2003.

[148] Marta Ferreiro-González, Jesús Ayuso, José A. Álvarez, Miguel Palma and Carmelo G. Barroso. Application of an HS–MS for the detection of ignitable liquids from fire debris. *Talanta*, 142:150–156, 2015.

[149] Hadi Parastar, Jagoš R. Radović, Mehdi Jalali-Heravi, Sergi Diez, Josep Maria Bayona and Roma Tauler. Resolution and quantification of complex mixtures of

polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC×GC-TOFMS combined to multivariate curve resolution. *Analytical Chemistry*, 83(24):9289–9297, 2011.

[150] Thomas Dutriez, Marion Courtiade, Jeremie Ponthus, Didier Thiébaut, Hugues Dulot, and Marie Claire Hennion. Complementarity of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry and high temperature comprehensive two-dimensional gas chromatography for the characterization of resin fractions from vacuum gas oils. *Fuel*, 96:108–119, 2012.

[151] Bárbara M. F. Ávila, Ricardo Pereira, Alexandre O. Gomes and Débora a. Azevedo. Chemical characterization of aromatic compounds in extra heavy gas oil by comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry. *Journal of Chromatography A*, 1218:3208–3216, 2011.

[152] James J. Harynuk, Aleisha D. Rossé, G. Bryce McGarvey. Study of alkyl phosphates in industrial petroleum mixtures by comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *Analytical and Bioanalytical Chemistry*, 401(8):2415–2422, 2011.

[153] Sara Mostafapour and Hadi Parastar. N-way partial least squares with variable importance in projection combined to GC × GC-TOFMS as a reliable tool for toxicity identification of fresh and weathered crude oils. *Analytical and Bioanalytical Chemistry*, 407:285–295, 2014.

[154] Katie D. Nizio and James J. Harynuk. Profiling alkyl phosphates in industrial petroleum samples by comprehensive two-dimensional gas chromatography with nitrogen phosphorus detection (GC×GC-NPD), post-column deans switching, and concurrent backflushing. *Energy and Fuels*, 28(3):1709–1716, 2014.

[155] Nobuo Ochiai, Teruyo Ieda, Kikuo Sasamoto, Yoshikatsu Takazawa, Shunji Hashimoto, Akihiro Fushimi and Kiyoshi Tanabe. Stir bar sorptive extraction and comprehensive two-dimensional gas chromatography coupled to high-resolution time-of-flight mass spectrometry for ultra-trace analysis of organochlorine pesticides in river water. *Journal of Chromatography A*, 1218(39):6851–6860, 2011.

[156] Miren Pena-Abaurrea, Adrian Covaci and Lourdes Ramos. Comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry for the identification of organobrominated compounds in bluefin tuna. *Journal of Chromatography A*, 1218(39):6995–7002, 2011.

[157] Caixiang Zhang, Robert P. Eganhouse, James Pontolillo, Isabelle M. Cozzarelli and Yanxin Wang. Determination of nonylphenol isomers in landfill leachate and municipal wastewater using steam distillation extraction coupled with comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. *Journal of Chromatography A*, 1230:110–116, 2012.

[158] Amy Dindal, Elizabeth Thompson, Erich Strozier, and Stephen Billets. Application of GC–HRMS and GC×GC–TOFMS to aid in the understanding of a dioxin assay's performance for soil and sediment samples. *Environmental science & technology*, 45(24):10501–10508, 2011.

[159] Garth T. Llewellyn, Frank Dorman, J. L. Westland, D. Yoxtheimer, Paul Grieve, Todd Sowers, E. Humston-Fulmer and Susan L. Brantley. Evaluating a groundwater supply contamination incident attributed to Marcellus Shale gas development. *Proceedings of the National Academy of Sciences of the United States of America*, 112(20):6325–6330, 2015.

[160] Elizabeth M. Humston, Kenneth M. Dombek, Benjamin P. Tu, Elton T. Young, and Robert E. Synovec. Toward a global analysis of metabolites in regulatory mutants of yeast. *Analytical and Bioanalytical Chemistry*, 401:2387–2402, 2011.

[161] Jason H. Winnike, Xiaoli Wei, Kevin J Knagge, Steven D. Colman, Simon G. Gregory and Xiang Zhang. Comparison of GC-MS and GC×GC-MS in the analysis of human serum samples for biomarker discovery. *Journal of proteome research*, 14(4):1810–1817, 2015.

[162] Glenn S. Frysinger and Richard B. Gaines. Forensic analysis of ignitable liquids in fire debris by comprehensive two-dimensional gas chromatography. *Journal of Forensic Science*, 47(3):471–482, 2002.

[163] John V. Seeley and Stacy K. Seeley. Multidimensional gas chromatography: Fundamental advances and new applications. *Analytical Chemistry*, 85(2):557–578, 2013.

[164] Karisa M. Pierce, Benjamin Kehimkar, Luke C. Marney, Jamin C. Hoggard and Robert E. Synovec. Review of chemometric analysis techniques for comprehensive two dimensional separations data. *Journal of Chromatography A*, 1255:3–11, 2012.

[165] Chiara Cordero, Johannes Kiefl, Peter Schieberle, Stephen E. Reichenbach and Carlo Bicchi. Comprehensive two-dimensional gas chromatography and food sensory properties: potential and challenges. *Analytical and Bioanalytical Chemistry*, 407(1):169–191, 2014.

[166] Sung-Tong Chin and Philip J. Marriott. Multidimensional gas chromatography beyond simple volatiles separation. *Chemical Communications*, 50:8819–8833, 2014.

[167] Kevin J. Johnson and Robert E. Synovec. Pattern recognition of jet fuels : comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 60:225–237, 2002.

[168] José Figueira, Hugo Câmara, Jorge Pereira and José S Câmara. Evaluation of volatile metabolites as markers in *Lycopersicon esculentum L. cultivars* discrimination by multivariate analysis of headspace solid phase microextraction and mass spectrometry data. *Food Chemistry*, 145:653–63, 2014.

[169] Yun Xu, Stephen J. Fowler, Ardeshir Bayat and Royston Goodacre. Chemometrics models for overcoming high between subject variability: applications in clinical metabolic profiling studies. *Metabolomics*, 10(3):375–385, 2013.

[170] Mrinal Kumar Das, Subasa Chandra Bishwal, Aleena Das, Deepti Dabral, Ankur Varshney, Vinod Kumar Badireddy, and Ranjan Nanda. Investigation of Gender-Specific Exhaled Breath Volatome in Humans by GC×GC-TOF-MS. *Analytical Chemistry*, 86(2):1229–1237, 2013.

[171] Yin-Hua Xiong, Ying Xu, Li Yang and Zheng-Tao Wang. Gas chromatography-mass spectrometry-based profiling of serum fatty acids in acetaminophen-induced liver injured rats. *Journal of Applied Toxicology : JAT*, 34(2):149–157, 2014.

[172] H. K. Lim, S. Stellingweif, S. Sisenwine and K. W. Chan. Rapid drug metabolite profiling using fast liquid chromatography, automated multiple-stage mass spectrometry and receptor-binding. *Journal of Chromatography. A*, 831(2):227–241, 1999.

[173] Thomas Gröger and Ralf Zimmermann. Application of parallel computing to speed up chemometrics for GC×GC-TOFMS based metabolic fingerprinting. *Talanta*, 83:1289–1294, 2011.

[174] Karin Kjeldahl and Rasmus Bro. Some common misunderstandings in chemometrics. *Journal of Chemometrics*, 24(7-8):558–564, 2010.

[175] Rachel E. Mohler, Kenneth M. Dombek, Jamin C. Hoggard, Karisa M. Pierce, Elton T. Young, and Robert E. Synovec. Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: combining discovery-mode and deconvolution chemometric software. *The Analyst*, 132(8):756–767, 2007.

[176] Luke C. Marney, W. Christopher Siegler, Brendon A. Parsons, Jamin C. Hoggard, Bob W. Wright, and Robert E. Synovec. Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data. *Talanta*, 115:887–895, 2013.

[177] Brendon A. Parsons, Luke C. Marney, W. Christopher Siegler, Jamin C. Hoggard, Bob W. Wright and Robert E. Synovec. Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC-TOFMS) Data Using a Null Distribution Approach. *Analytical Chemistry*, 87(7): 3812–3819, 2015.

[178] Martin Lopatka, Andjoe A. Sampat, Steffan Jonkers, Lawrence A. Adutwum, Hans G. J. Mol, Guido van der Weg, James J. Harynuk, Peter J. Schoenmakers, Arian van Asten, Marjan J. Sjerps, and Gabriel Vivó-Truyols. Local Ion Signatures (LIS) for comparison of comprehensive two-dimensional gas chromatography applied to fire debris analysis. *Forensic Chemistry*, 3:1–13, 2016.

[179] Mohammadreza Khanmohammadi, Amir Bagheri Garmarudi, and Miguel de la Guardia. Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis. *Talanta*, 104:128–34, 2013.

[180] Kanet Wongravee, Nina Heinrich, Maria Holmboe, Michele L. Schaefer, Randall R. Reed, Jose Trevejo, and Richard G. Brereton. Variable selection using iterative reformulation of training set models for discrimination of samples: application to gas chromatography/mass spectrometry of mouse urinary metabolites. *Analytical Chemistry*, 81(13):5204–5217, 2009.

[181] Artur J. Ferreira and Mário A. T. Figueiredo. An unsupervised approach to feature discretization and selection. *Pattern Recognition*, 45(9):3048–3060, 2012.

[182] Sonja Peters, Gabriel Vivó-Truyols, Philip J. Marriott, and Peter J. Schoenmakers. Development of an algorithm for peak detection in comprehensive two-dimensional chromatography. *Journal of Chromatography A*, 1156:14–24, 2007.

[183] Imhoi Koo, Seongho Kim, and Xiang Zhang. Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry. *Journal of Chromatography A*, 1298:132–138, 2013.

[184] Karisa M. Pierce, Janiece L. Hope, Kevin J. Johnson, Bob W. Wright, and Robert E. Synovec. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *Journal of Chromatography. A*, 1096(1-2):101–10, 2005.

[185] Eric Stauffer, Julia A. Dolan and Reta Newman. *Fire debris analysis*. Academic Press, 2007.

[186] José R. Almirall and Kenneth G. Furton. *Analysis and interpretation of fire scene evidence*. CRC Press, Boca Raton, 2004.

[187] ASTM Standard. E1618-06.?Standard Test Method for Ignitable Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography-Mass Spectrometry?; ASTM International: West Conshahocken, PA, 2006.

[188] Jamie M. Baerncopf, Victoria L. McGuffin and Ruth W. Smith. Association of Ignitable Liquid Residues to Neat Ignitable Liquids in the Presence of Matrix Interferences Using Chemometric Procedures. *Journal of Forensic Sciences*, 56(1):70–81, 2011.

[189] J. I. Cacho, N. Campillo, M. Aliste, P. Viñas and M Hernández-Córdoba. Headspace sorptive extraction for the detection of combustion accelerants in fire debris. *Forensic Science International*, 238:26–32, 2014.

[190] Minho Chae, Robert J. Shmookler Reis, and John J. Thaden. An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas chromatography-mass spectrometry peaks. *BMC Bioinformatics*, 9(9), S15, 2008.

[191] Jonas Gros, Deedar Nabi, Petros Dimitriou-Christidis, Rebecca Rutler and J. Samuel Arey. Robust algorithm for aligning two-dimensional chromatograms. *Analytical Chemistry*, 84(21):9033–9040, 2012.

[192] Mark Nowlan, Allan W. Stuart, Gene J. Basara and P. Mark L Sandercock. Use of a solid absorbent and an accelerant detection canine for the detection of ignitable liquids burned in a structure fire. *Journal of Forensic Sciences*, 52(3):643–648, 2007.

[193] Martin Lopatka, Michael E. Sigman, Marjan J. Sjerps, Mary R. Williams and Gabriel. Vivo-Truyols. Class-conditional feature modeling for ignitable liquid classification with substantial substrate contribution in fire debris analysis. *Forensic Science International*, 252:177–186, 2015.

[194] Nathanial E. Watson, Matthew M. VanWingerden, Karisa M. Pierce, Bob W. Wright and Robert E Synovec. Classification of high-speed gas chromatography–mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection. *Journal of Chromatography A*, 1129(1):111–118, 2006.

[195] Yawei Wang, Jiemin Liu, Ning Li, Guoqing Shi, Guibin Jiang and Weiping Ma. Preliminary study of the retention behavior for different compounds using cryogenic chromatography at different initial temperatures. *Microchemical Journal*, 81(2):184–190, 2005.

[196] M. J. Zhang, S. D. Li and B. J. Chen. Compositional studies of high-temperature coal tar by GC/FTIR analysis of light oil fractions. *Chromatographia*, 33(3):138–146, 1992.

[197] A Paulina de la Mata, Rachel H McQueen, Seo Lin Nam, and James J Harynuk. Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics. *Analytical and Bioanalytical Chemistry*, 409(7):1905–1913, 2017.

[198] Murray S. Weitzman. *Measures of overlap of income distributions of white and Negro families in the United States*, volume 22. US Bureau of the Census, 1970.

[199] Henry F. Inman and Edwin L. Bradley. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap

of two normal densities. *Communications in Statistics - Theory and Methods*, 18(10):3851–3874, 1989.

[200] Kameo Matusita. Decision rule, based on the distance, for the classification problem. *Annals of the Institute of Statistical Mathematics*, 8(1):67, 1956.

[201] Madhuri S. Mulekar and Satya N. Mishra. Confidence interval estimation of overlap: Equal means case. *Computational Statistics and Data Analysis*, 34(2):121–137, 2000.

[202] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. *Proceeding of the Second International Symposium on Information Theory*,267–281, 1973.

[203] Shuhua Hu. Akaike information criterion statistics. *Mathematics and Computers in Simulation*, 29(5):452, 1987.

[204] Brucker Daltonics Inc. Metabolic profiling of different coffee types on the bruker compact™ qtof system, 2013.

[205] Samuel Kotz and Norman L. Johnson Breakthroughs in Statistics, Foundations and Basic Theory, Springer, 1992.

[206] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, volume 172. Springer Science & Business Media, 2002.

[207] Michael Snipes and D. Christopher Taylor. Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3(1):3–9, 2014.

[208] Linus Pauling. Historical perspective of crystallography. In *Struct. Bonding Cryst.*, Volume 1, 1–12. Academic Press,v1981.

[209] L Pauling. *The Nature of the Chemical Bond*. Cornel University Press, Ithaca, NY, 1960.

[210] J C. Phillips. Structure and properties: Mooser-Pearson plots. *Helvetica Physica Acta*, 58(2-3):209–215, 1985.

[211] James Charles Phillips and J A. Van Vechten. Dielectric classification of crystal structures, ionization potentials, and band structures. *Physical Review Letters*, 22(14):705–708, 1969.

[212] D. G. Pettifor. A chemical scale for crystal-structure maps. *Solid State Communications*, 51(1):31–34, 1984.

[213] Alex Zunger. Systematization of the stable crystal structure of all AB-type binary compounds: a pseudopotential orbital-radii approach. *Physical Review B: Condensed Matter and Materials Physics*, 22(12):5839–5872, 1980.

[214] P. Villars. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *Journal of the Less-Common Metals*, 92(2):215–238, 1983.

[215] P. Villars, K. Cenzual, J. Daams, Y. Chen and S. Iwata. Data-driven atomic environment prediction for binaries using the Mendeleev number. Part 1. Composition AB. *Journal of Alloys and Compounds*, 367(1-2):167–175, 2004.

[216] Stepan Sergeevich Batsanov. Dielectric Methods of Studying the Chemical Bond and the Concept of Electronegativity. *Russian Chemical Reviews*, 51(7):684, 1982.

[217] P. Villars, Phil Karlton, and Scott McGregor. Pearson's crystal data – crystal structure database for inorganic compounds (on dvd), 2015.

[218] P. Villars, H. Okamoto and K. Cenzual. ASM Alloy Phase Diagrams Database, 2016.

[219] J. Nowacki. Phosphorus in iron alloys surface engineering. *Journal of Achievements in Materials and Manufacturing Engineering*, 24(1):57–67, 2007.

[220] Horie Toshio. Conductive film, corrosion-resistant conduction film, corrosion-resistant conduction material and process for producing the same, 2010.

[221] H. Barz, H. C. Ku, G. P. Meisner, Z. Fisk and B. T. Matthias. Ternary transition metal phosphides: High temperature superconductors. *Proceedings of the National Academy of Science of the United States of America*, 77(6):3132, 1980.

[222] Stephanie L. Brock, Susanthri C. Perera and Kimber L. Stamm. Chemical routes for production of transition-metal phosphides on the nanoscale: Implications for advanced magnetic and catalytic materials. *Chemistry - A European Journal*, 10(14):3364–3371, 2004.

[223] Rusen Yang, Yu Lun Chueh, Jenny Ruth Morber, Robert Snyder, Li Jen Chou, and Zhong Lin Wang. Single-crystalline branched zinc phosphide nanostructures: Synthesis, properties, and optoelectronic devices. *Nano Letters*, 7(2):269–275, 2007.

[224] Diana C. Phillips, Stephanie J. Sawhill, Randy Self, and Mark E. Bussell. Synthesis, Characterization, and Hydrodesulfurization Properties of Silica-Supported Molybdenum Phosphide Catalysts. *Journal of Catalysis*, 207(2):266–273, 2002.

[225] Mme Roy-Montreuil, A. Michel B. Deyris, A. Rouault, P. L'Héritier, A. Nylund, J.P. Sénateur, and R. Fruchart. Nouveaux composes ternaires MM'P et MM'As interactions metalliques et structures. *Material Research Bulletin*, 7:813–826, 1972.

[226] Y. F. Lomnytska and Y. B. Kuz'ma. New phosphides of IVa and Va group metals with TiNiSi-type. *Journal of Alloys and Compounds*, 269(1-2):133–137, 1998.

[227] V. Raghavan. Fe-P-Ti (Iron-Phosphorus-Titanium). *Journal of Phase Equilibria and Diffusion*, 29(6):529–531, 2008.

[228] Oksana Toma, Mariya Dzevenko, Anton Oliynyk, and Yaroslava Lomnytska. The Ti-Fe-P system: phase equilibria and crystal structure of phases. *Open Chemistry*, 11(9), 2013.

[229] L. D Gulay, Y. M. Kalychak, M Wołcyrz and K Łukaszewicz. Crystal structure of RNiPb (R=Y, Nd, Sm, Gd, Tb, Dy, Ho, Er, Tm, Lu) compounds. *Journal of Alloys and Compounds*, 313(1-2):42–46, 2000.

[230] M. L. Fornasini, F. Merlo, A. Palenzona and M. Pani. Valency changes of ytterbium in YbMnGe and in the YbMnSi1-xGex pseudo-ternary system. *Journal of Alloys and Compounds*, 335(1-2):120–125, 2002.

[231] Jan F Riecken, Ute Ch, Gunter Heymann, Sudhindra Rayaprol, Hubert Huppertz, Rolf-dieter Hoffmann, and P Rainer. Synthesis , Structure and Properties of the High-pressure Modifications of the Ternary Compounds RE PtSn (RE = La , Pr , Sm). *Anorg. Allg. Chem*, 2006.

[232] Jan Freerks Riecken, Gunter Heymann, Hubert Huppertz, and Rainer Pöttgen. The high-temperature phases HT-YPtSn, HT-GdPtSn, and HT-TbPtSn. *Zeitschrift fur Anorganische und Allgemeine Chemie*, 633(5-6):869–872, 2007.

[233] Jan F Riecken, Gunter Heymann, Wilfried Hermes, and Ute Ch. High-pressure / High-temperature Studies on the Stannides RE NiSn (RE = Ce , Pr , Nd , Sm) and RE PdSn (RE = La , Pr , Nd). *Anorg. Allg. Chem*, 61(1):66, 2008.

[234] Villars P., Phil Karlton, and Scott McGregor. Pearson's Crystal Data – Crystal Structure Database for Inorganic Compounds (on DVD), 2016.

[235] P. Villars, H. Okamoto, K. Cenzual, Villars P., Okamoto H. and Cenzual K. ASM Alloy Phase Diagrams Database, 2015.

[236] Ronald Kshikhundo and Shayalethu Itumhelo. Bacterial species identification. *World News of Natural Sciences*, 3:26–38, 2016.

[237] H.C. Gram. Gram staining. *Fortschritte der Medicin*, 2:185–189, 1884.

[238] D. F. Welch. Applications of cellular fatty acid analysis. *Clinical Microbiology Reviews*, 4(4):422–438, 1991.

[239] Giorgia Purcaro, Peter Quinto Tranchida, Paola Dugo, Erminia La Camera, Giuseppe Bisignano, Lanfranco Conte, and Luigi Mondello. Characterization of bacterial lipid profiles by using rapid sample preparation and fast comprehensive two-dimensional gas chromatography in combination with mass spectrometry. *Journal of Separation Science*, 33(15):2334–2340, 2010.

[240] María Teresa Núñez-cardona. Fatty Acids Analysis of Photosynthetic Sulfur Bacteria by Gas Chromatography. *Gas Chromatography - Biochemicals, Narcotics and Essential Oils*, pages 117–138, 2012.

[241] P. Pattanapipitpaisal, N. L. Brown, and L. E. Macaskie. Chromate reduction and 16s rRNA identification of bacteria isolated from a Cr(VI)-contaminated site. *Applied Microbiology and Biotechnology*, 57(1-2):257–261, 2001.

[242] Ok Sun Kim, Yong Joon Cho, Kihyun Lee, Seok Hwan Yoon, Mincheol Kim, Hyunsoo Na, Sang Cheol Park, Yoon Seong Jeon, Jae Hak Lee, Hana Yi, Sungho Won, and Jongsik Chun. Introducing EzTaxon-e: A prokaryotic 16s rRNA gene sequence database with phylotypes that represent uncultured species. *International Journal of Systematic and Evolutionary Microbiology*, 62(PART 3):716–721, 2012.

[243] Claudio Foschi, Luca Laghi, Carola Parolin, Barbara Giordani, Monica Compri, Roberto Cevenini, Antonella Marangoni, and Beatrice Vitali. Novel approaches for the taxonomic and metabolic characterization of lactobacilli: Integration of 16S rRNA gene sequencing with MALDI-TOF MS and 1H-NMR. *PLoS ONE*, 12(2):1–18, 2017.

[244] C. D. Calvano, R. A. Picca, E. Bonerba, G. Tantillo, N. Cioffi and F. Palmisano. MALDI-TOF mass spectrometry analysis of proteins and lipids in Escherichia coli exposed to copper ions and nanoparticles. *Journal of Mass Spectrometry*, 51(9):828–840, 2016.

[245] J. O. Lay. MALDI-TOF mass spectrometry and bacterial taxonomy. *TrAC - Trends in Analytical Chemistry*, 19(8):507–516, 2000.

[246] Sascha Sauer and Magdalena Kliem. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology*, 8(1):74–82, 2010.

[247] J. A. Falkner, D. M. Veine, M. Kachman, A. Walker, J. R. Strahler and P. C. Andrews. Validated MALDI-TOF/TOF mass spectra for protein standards. *Journal of the American Society of Mass Spectrometry*, 18, 2007.

[248] Silpak Biswas and Jean Marc Rolain. Use of MALDI-TOF mass spectrometry for identification of bacteria that are difficult to culture. *Journal of Microbiological Methods*, 92(1):14–24, 2013.

[249] Phaik Lyn Oh, Andrew K. Benson, Daniel A. Peterson, Prabhu B. Patil, Etsuko N. Moriyama, Stefan Roos, and Jens Walter. Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *The ISME Journal*, 4(3):377–387, 2010.

[250] Marcia Shu Wei Su, Phaik Lyn Oh, Jens Walter, and Michael G. Gänzle. Intestinal origin of sourdough Lactobacillus reuteri isolates as revealed by phylogenetic, genetic, and physiological analysis. *Applied and Environmental Microbiology*, 78(18):6777–6780, 2012.

[251] Jinshui Zheng, Xin Zhao, Xiaoxi B Lin, and Michael Gänzle. Comparative genomics lactobacillus reuteri from sourdough reveals adaptation of an intestinal symbiont to food fermentations. *Scientific reports*, 5, 2015.

[252] Jessica M. A. Blair, Grace E. Richmond, Andrew M. Bailey, Al Ivens and Laura J. V. Piddock. Choice of bacterial growth medium alters the transcriptome and phenotype of salmonella enterica serovar typhimurium. *PLoS One*, 8(5):e63912, 2013.

Appendix A

Extraction and Preparation of Fire Debris Samples

Samples were, stored, extracted, and analyzed using established protocols at the Royal Canadian Mounted Police (RCMP)1 trace analysis laboratory. Protocols. These protocols were in line with ASTM methods E1618 and E1412. The headspace of samples were extracted with activated carbon strips onto (Albrayco Technologies, Cromwell, CT) for 16 h at 60 °C. An elution solution of CS_2 containing perdeuterated alkane ladder, namely, of *n*-heptane (d16), *n*-nonane (d20), *n*-undecane (d24), *n*-tridecane (d28), *n*-pentadecane (d32), *n*-heptadecane (d36), *n*-nonadecane (d-40) and *n*-heneicosane (d-44) (CDN Isotopes, Pointe-Claire, QC) at concentrations of 16 µg L−1 was prepared. This elution solution was used to elute the content of the activated carbon strips. Eluted with CS2 and the eluent injected into GC–MS for anlaysis.2

GC-MS Separation Conditions for Fire Debrist Samples

The eluates were analyzed with Agilent Technologies 7890A gas chromatographs (GC) with 5975 quadrupole mass spectrometers (MS) and 7683 auto samplers (Agilent Technologies, Mississauga, ON). Automation and data acquisition was done with MSD ChemStation (Agilent). Separation was performed on a 30 m × 250 µm × 0.25 µm HP-1MS columns (Agilent). The GC oven temperature program was an initial 40 °C (held for 3.0 min) and a ramp to 250 °C at a rate of 8 °C min−1, and held at 250 °C 0.75 min. Split mode sample injection was used with an injection volume of 1 µL, a split ratio of 20:1 and injector temperature of 250 °C. The carrier gas was High Purity Hydrogen with a flow rate of 1.1 mL/min. The MS source and transfer line temperatures were set at 230 and 300 °C °C, respectively. ***Chemometric classification of casework arson samples based on gasoline content Sinkov, Nikolai A. et al. Forensic Science International , Volume 235 , 24 - 31***.

Solid Phase Micro Extraction (SPME) of Green Tea Samples

A 200 mg portion of tea sample was weighed into a 6 mL clear headspace (Chromatgraphic Specialties Inc, Brockville, ON, Canada) and caped. The vial was inserted about 1.5 cm deep into an oil bath set on a hotplate stirrer (VWR, Edmonton, AB, Canada). It was allowed to equilibrate for 10 min at 60 °C. Volatiles in the headspace were extracted using a DVB/CARB/PDMS SPME fibre (Supelco, Bellefonte, PA,USA) for 20 min at 60 °C.

Chromatographic Separation and TOF MS Conditions for Green Tea Samples

A Pegasus®4D GC×GC-TOFMS equipped with a liquid nitrogen cryogenic modulator (Leco, St Joseph, MI) was used. Sample was desorbed for 2 min via splitless injection with the injector temperature kept at 230 °C. Separation was performed with a 30 m × 250 µm internal diameter ×1 µm Rtx 5 film (Restek, Bellefonte, PA, USA) and a 1.44 m × 250 µm internal diameter × 0.18 µm DB Wax (Agilent / J & W Technologies, Santa Clara, CA, USA) as primary and secondary columns, respectively. Ultra high purity helium (Praxair, Edmonton, AB, Canada) was used as the carrier gas at a constant flow of 1.5 mL/min. A temperature program was used with an initial oven temperature set at 40 °C for 3 min and ramped at8 °C/min until it reached 230 °C. It was held at this temperature for 5 min giving a total run time of 31.75 min. The secondary oven and modulator were set to be 10 °C above the primary oven with a cap at 240 °C. A modulation period of 5 s was used with 0.8 s and 1.7 s for hot and cold jets, respectively. The ion source was set at 200 °C with an optimized detector acquisition voltage of 1495 V and 200 V offset. Signals within the range of 35 - 300 amu were collected at 100 spectra/s.

Identification of Anchors for GC×GC-MS alignment

A set of thirteen (13) compounds present in all chromatograms were used as anchors to model the shift in each modulation. One chromatogram was selected and the apexes of the anchor compounds were manually identified. The compounds were grouped into lower (<1.5 - 7 anchors) and higher (>1.5 - 6 anchors) retentions (Fig. A1). The peak apexes of the lower and higher retention compounds were fitted separately to a first degree polynomial (Fig. A2). The anchors are located in the chromatogram to be aligned by comparing the mass spectra using the weighted cosine correlation score described by Kim et. al. (S. Kim, A. Fang, B. Wang, J. Jeong, and X. Zhang, An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure, Bioinformatics, vol. 27, no. 12, pp. 1660 - 1666, 2011). Lower and upper anchors in the chromatogram to be aligned are also fitted separately to a first degree polynomial.

Modulation to Modulation Alignment for GC×GC-MS Chromatogram

Since these fitted lines runs across the length of the entire chromatogram, each modulation is segmented into three (parts), providing six (6) anchor points for each modulation. Each modulation in the sample chromatogram is aligned to that of the target chromatogram using the six (6) anchor points.

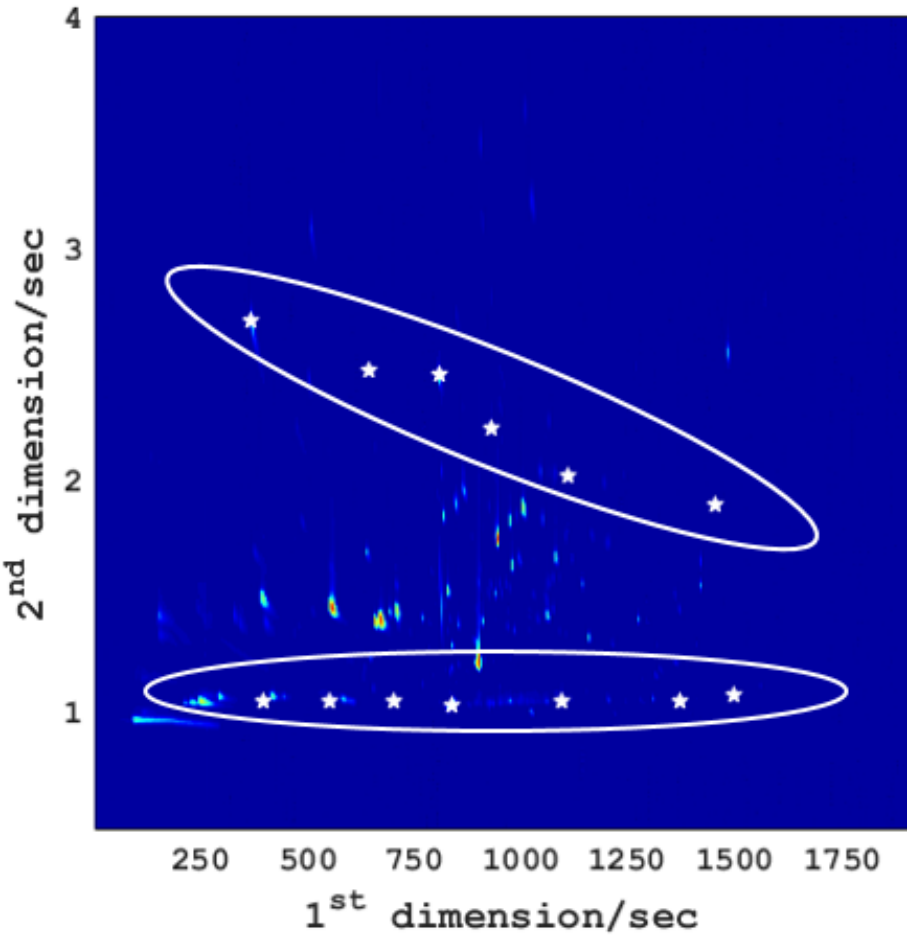


Fig. A1 - Anchors for alignment of GC×GC chromatogram

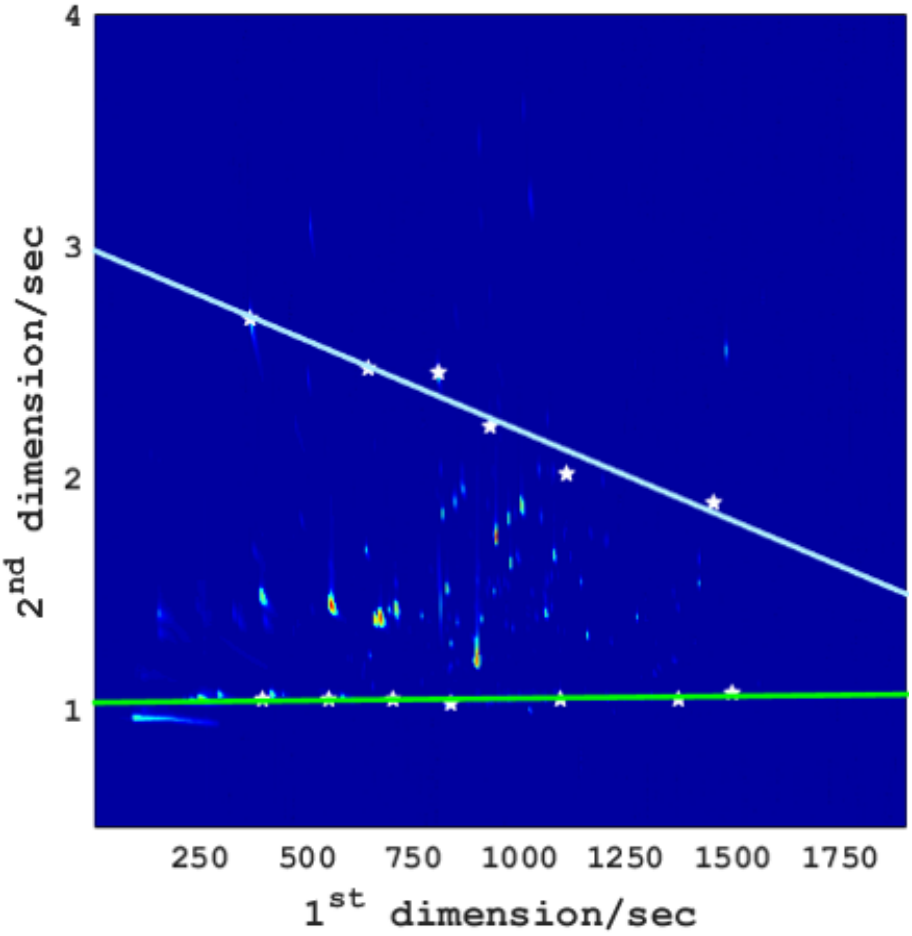


Fig. A2 - Anchors fitted to a first degree polynomial

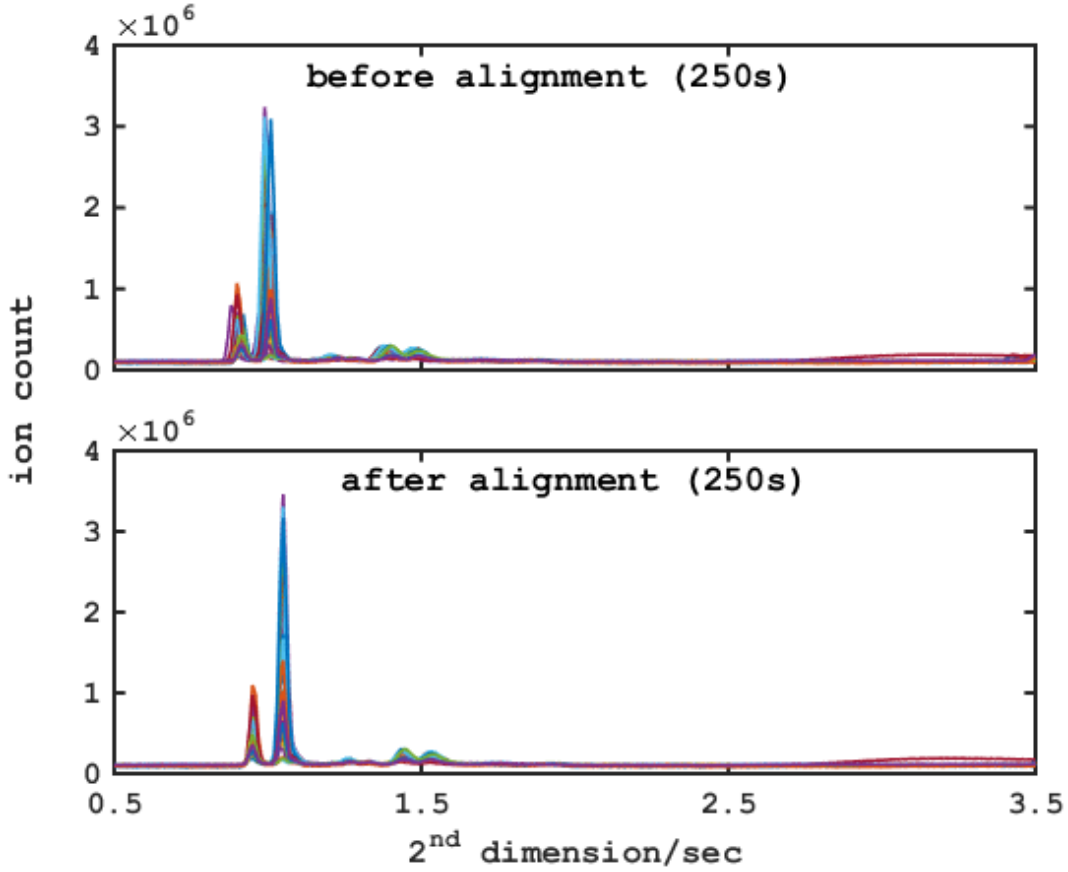


Fig. A3 - Comparison of the same modulation before and after alignment

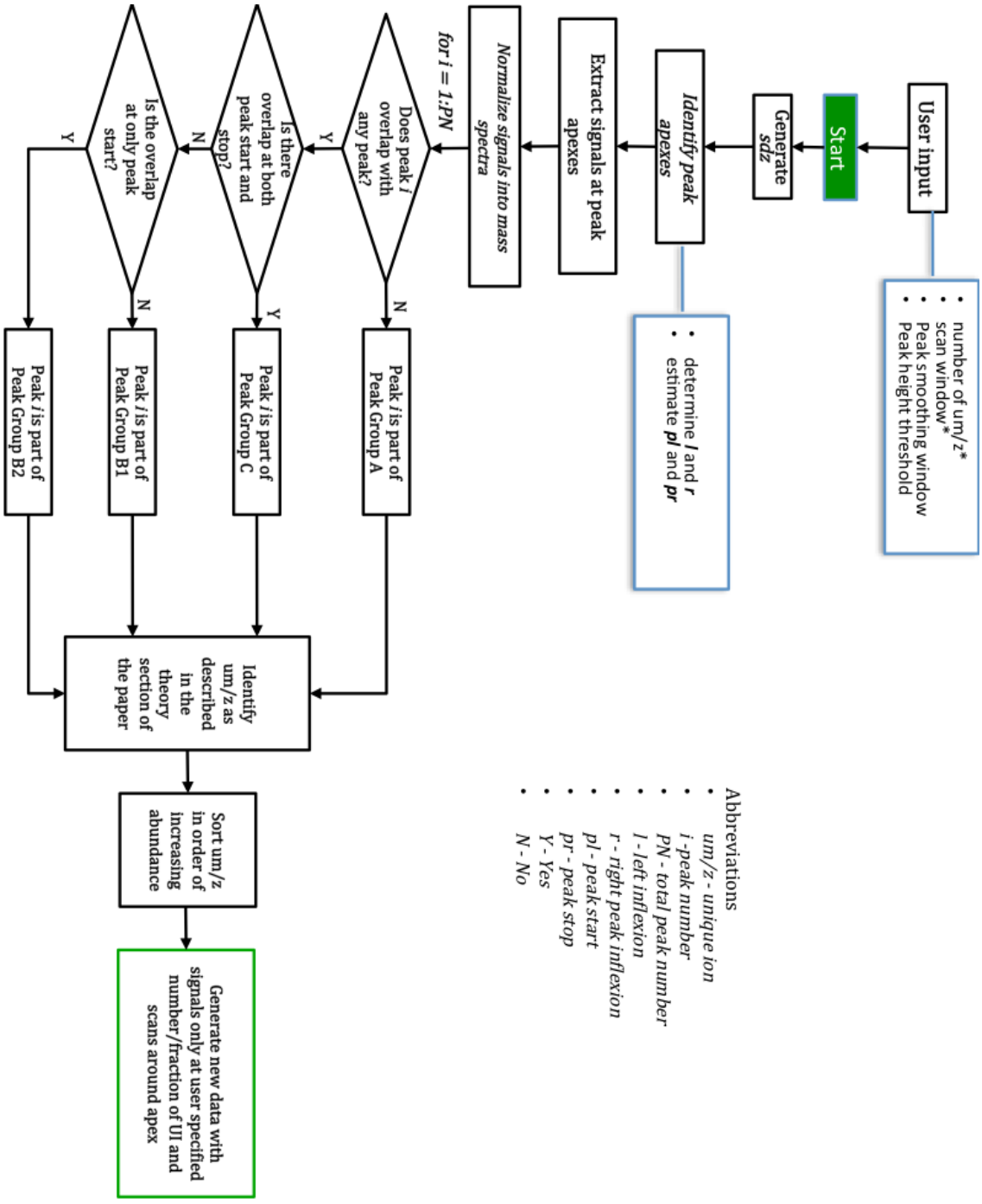


Fig. A4 - Algorithm for the Identification of UIF1D

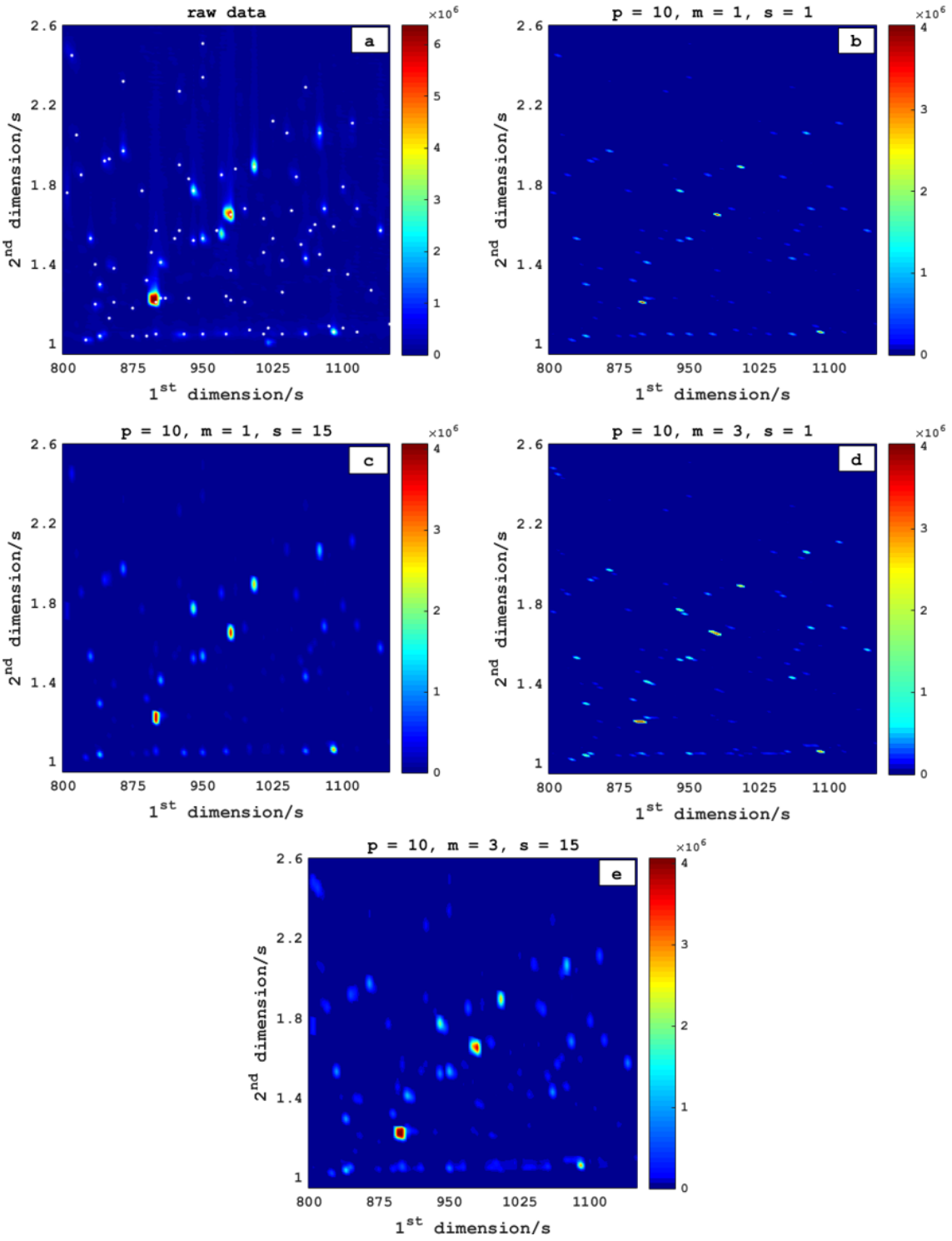


Fig. A4 - Comparison of Chromatograms Generated using various settings of UIF2D

Appendix B

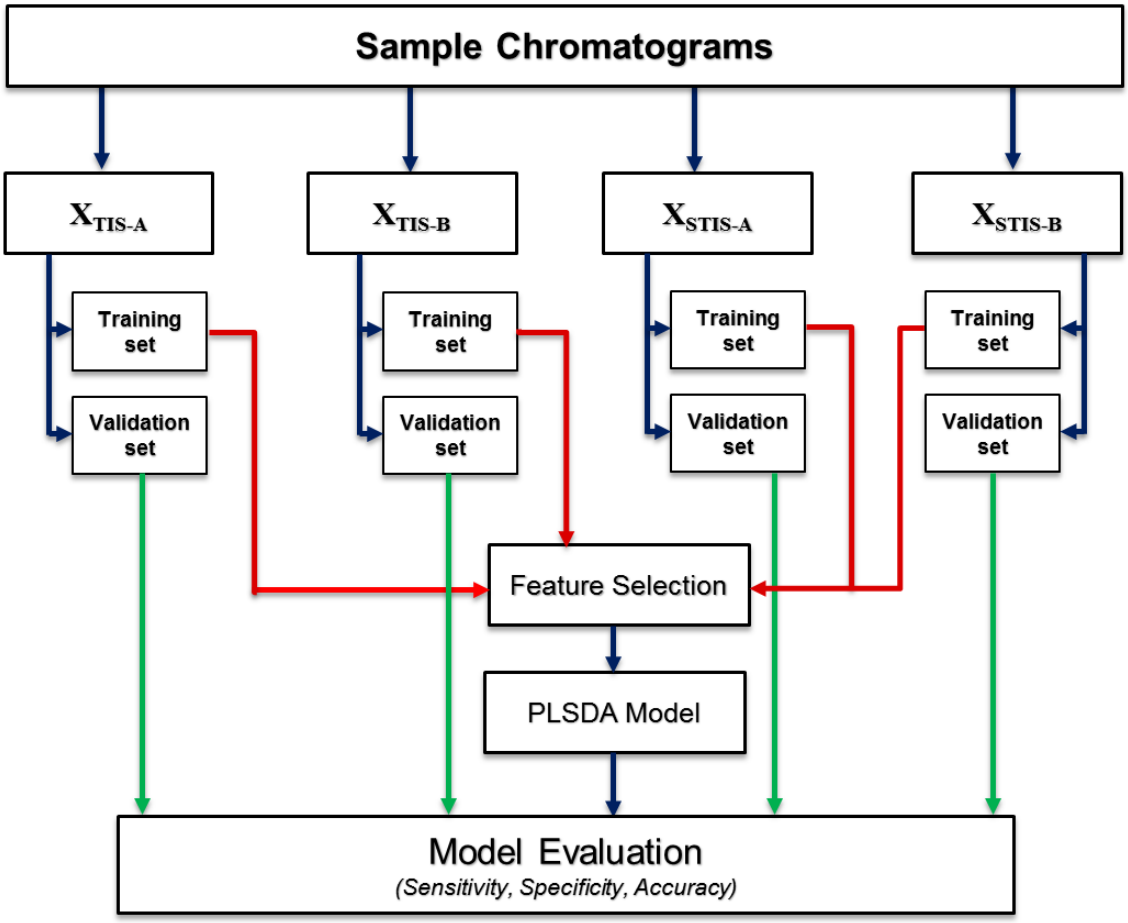


Fig. B1 - Chemometric analysis work flow for TIS and STIS datasets

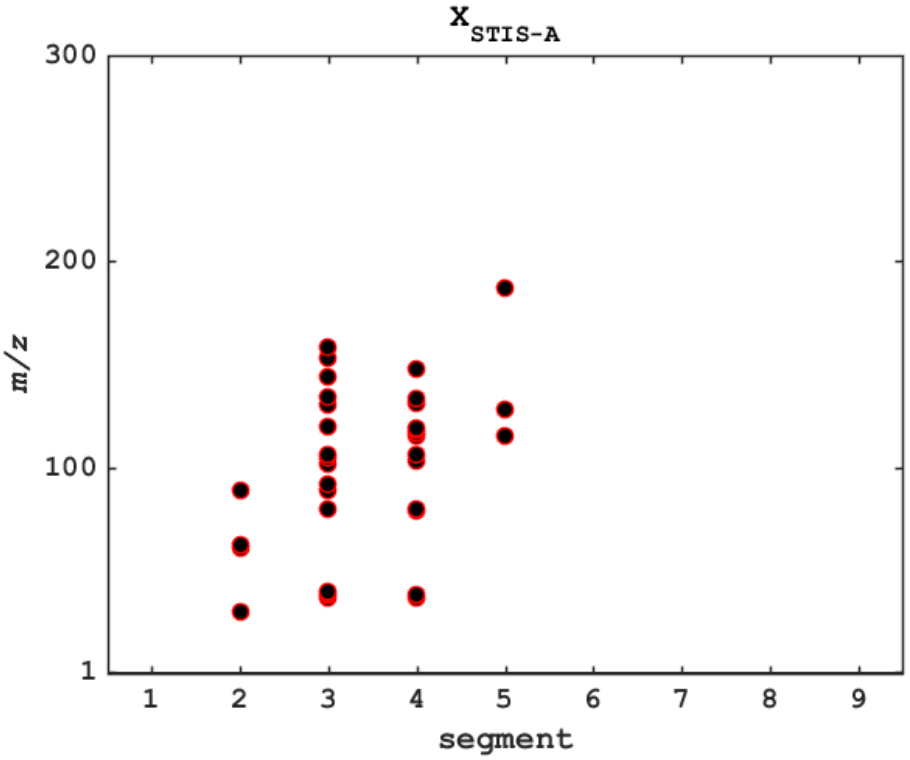


Fig. B2 - Ions selected in each segment by CR-FS using $\mathbf{X}_{\mathbf{STIS\text{-}A}}$ dataset (34 variables)

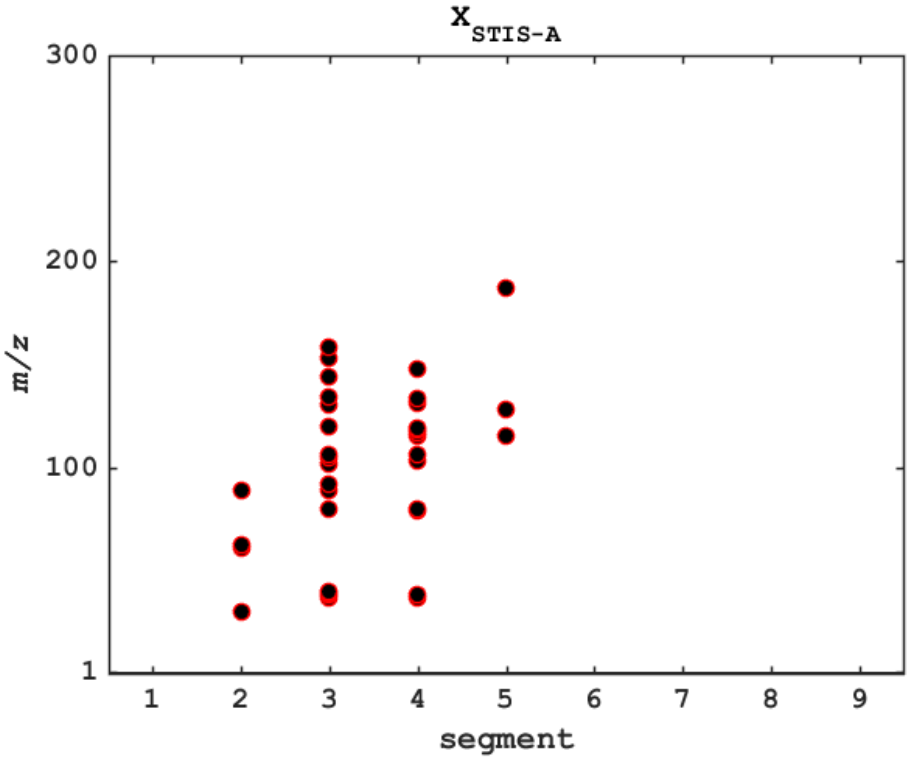


Fig. B3 - Ions selected in each segment by CR-FS using $\mathbf{X_{STIS\text{-}B}}$ dataset (36 variables)

Appendix C

1. ●Electronegativity difference (Pauling scale)
2. ★Electronegativity difference (Martynov-Batsanov scale)
3. ●Electronegativity difference (Gordy scale)
4. ●Electronegativity difference (Mulliken scale)
5. ★Electronegativity difference (Allred-Rochow scale)
6. ●Mean electronegativity (Pauling scale)
7. ★Mean electronegativity (Martynov-Batsanov scale)
8. ●Mean electronegativity (Gordy scale)
9. ★Mean electronegativity (Mulliken scale)
10. ★Mean electronegativity (Allred-Rochow scale)
11. Ionic character (Pauling scale)
12. ★Ionic character (Martynov-Batsanov scale)
13. Ionic character (Gordy scale)
14. Ionic character (Mulliken scale)
15. Ionic character (Allred-Rochow scale)
16. ●Sum of valence electrons
17. Mean number of electrons
18. ●Atomic number sum
19. Atomic number difference
20. Mean atomic number
21. Atomic weight difference
22. Mean atomic weight
23. Atomic weight sum
24. ●Atomic radius sum ($d_{A\text{-}B}$)
25. Mean atomic radius
26. ★Atomic radius ratio
27. 2×atomic radius difference ($d_{A\text{-}A}$ - $d_{A\text{-}B}$)
28. ●Covalent radius sum ($d_{A\text{-}B}$)
29. Mean covalent radius
30. ●Covalent radius ratio
31. ●2×covalent radius difference ($d_{A\text{-}A}$ - $d_{A\text{-}B}$)
32. Zunger radius sum (r_S+r_P) sum
33. Mean Zunger radius sum (r_S+r_P)
34. ★Zunger radius sum (r_S+r_P) ratio
35. ★2×Zunger radius sum (r_S+r_P) difference
36. Zunger radius sum (r_S+r_P) difference
37. Ionic radius sum ($d_{A\text{-}B}$)
38. Mean ionic radius
39. ★Ionic radius ratio
40. ★2×ionic radius difference ($d_{A\text{-}A}$ - $d_{A\text{-}B}$)
41. ●Crystal radius sum ($d_{A\text{-}B}$)
42. Mean crystal radius
43. ●Crystal radius ratio
44. ●2×crystal radius difference ($d_{A\text{-}A}$ - $d_{A\text{-}B}$)
45. Period number sum
46. Mean period number
47. ●Period number difference
48. ★Group number sum
49. Mean group number
50. ●Group number difference
51. ★Family number sum
52. Mean Family number
53. Family number difference
54. ●Quantum number (l) sum
55. Mean quantum number (l) mean
56. ●Quantum number (l) difference

Fig. C1 - A complete list of all the 56 variables. Blue cirles indicates features that were retained during the SBE stage whiles red starts indicates those added during the SFS.

Table C1 - Refined Crystallographic Data for RhCd

Formula	***RhCd***
fw (amu)	215.31
Space group	Pm m (No. 221)
a (Å)	3.2191(7)
V (Å^3)	33.358(13)
Z	1
$''\rho_{calcd}(cm^{-3})$	10.718
T (K)	296(2)
crystal dimensions (mm)	0.05 × 0.03 × 0.03
radiation	graphite monochromated Mo Ka, $\lambda = 0.71073$
$''\mu$(Mo Ka)(mm^{-1})	27.489
transmission factors	0.285–0.666
2$''\theta$limits	17.96–65.48
data collected	$4 \leq \boldsymbol{h} \leq 4, 4 \leq \boldsymbol{k} \leq 4, 4 \leq \boldsymbol{l} \leq 4$
no. of data collected	234
no. of unique data, including F_o^2 <0	13 (R_{int} = 0.0152)
no. of unique data, with $F_o^2 > 2\sigma(F_o^2)$	13
no. of variables	4
R(F) for $F_o^2 > 2\sigma(F_o^2)^a$	0.0086
Rw$(F_o^2)^b$	0.0185
goodness of fit	1.297
$(\Delta\rho)_{max}$ $(\Delta\rho)_{min}$ (eÅ^{-3})	0.617 - 0.339
Positional and displacement parameters [c]	
Rh at 1a (0, 0, 0)	
U−*iso* (Å^2)	0.02(2)
Cd at 1b (1/2, 1/2, 1/2)	
U−*iso* (Å^2)	0.016(14)
Interatomic distances (Å)	
Rh–Cd (×8)	2.7878(6)
Cd–Cd (×6)	3.2191(7)
Rh–Rh (×6)	3.2191(7)

[a] $R(F) = \sum \|\|F_o\| - |F_c\| \,/\, |\sum F_c| \,[\sum [w(F_o^2 - F_c^2)^2)] \,/\, \sum wF_o^4]^{\frac{1}{2}}$
where $w^{-1} = [\sigma^2(F_o^2) + (Ap)^2 + Bp]$, *and* $p = [max(F_o^2, 0) + 2F_o^2]/3$

33 properties of elements, used to calculate descriptors	
1. Atomic number	17. Outer shell electrons
2. Atomic weight	18. Period number
3. Atomic radius	19. Group number
4. Covalent radius	20. Family number
5. Metallic radius	21. L quantum number
6. Single bond radius	22. Melting point
7. Zunger radii sum	23. Boiling point
8. Ionic radius	24. Density
9. Crystal radius	25. First ionization energy
10. Pauling electronegativity	26. Electrical conductivity
11. Martynov-Batsanov electronegativity	27. Specific heat
12. Gordy electronegativity	28. Heat of fusion
13. Mülliken electronegativity	29. Heat of vapourization
14. Allred-Rochow electronegativity	30. Thermal conductivity
15. Metallic valence	31. Heat atomization
16. Number of valence electrons	32. Polarizability
	33. Mendeleev number
30 formulae, used to calculate descriptors	
1. Average number (of 3)	16. Number average B and C
2. Number A	17. Sum of two largest numbers
3. Number B	18. Sum of two smallest numbers
4. Number C	19. Difference of two largest numbers
5. Number sum of A and B	20. Difference of two smallest numbers
6. Number sum of A and C	21. Ratio of two largest numbers
7. Number sum of B and C	22. Ratio of two smallest numbers
8. Number difference of A and B	23. Average of two largest numbers
9. Number difference of A and C	24. Average of two smallest numbers
10. Number difference of B and C	25. Sum of two extremes
11. Number ratio A/B	26. Difference of two extremes
12. Number ratio A/C	27. Ratio of two extremes
13. Number ratio B/C	28. Average of two extremes
14. Number average A and B	29. Smallest number
15. Number average A and C	30. Largest number

Fig. C2 - A list of descriptors/variables for the ABC structure prediction study. *The first part of the list are the elemental properties considered. The second part show the mathematical transformations of the properties to generate new variables. For example for the first property, i.e. atomic number, the first calculated descriptor will be the average atomic number for A, B and C. The second, third and fourth descriptors are the atomic numbers of A, B and C, respectively and so on.*