AFRL-RH-WP-TR-2014-0131

A COMPREHENSIVE TOOL AND ANALYTICAL PATHWAY FOR DIFFERENTIAL MOLECULAR PROFILING AND BIOMARKER DISCOVERY

Claude C. Grigsby, Ph.D.
Human Signatures Branch
Human Centered ISR Division

OCTOBER 2014
Final Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2014-0131 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//	//signature//
Claude C. Grigsby, Ph.D. Work Unit Manager Human Signatures Branch	Louise A. Carter, Ph.D. Chief, Human-Centered-ISR Division Human Effectiveness Directorate 711th Human Performance Wing Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE ***(DD-MM-YY)*** 20 10 14	**2. REPORT TYPE** Final	**3. DATES COVERED** ***(From - To)*** January 2010 – September 2014

4. TITLE AND SUBTITLE
A Comprehensive Tool and Analytical Pathway for Differential Molecular Profiling and Biomarker Discovery

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Claude C. Grigsby, Ph.D.

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
C002

5f. WORK UNIT NUMBER
H04V (7184C002)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Human Centered ISR Division
Human Signatures Branch
Wright-Patterson Air Force Base, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHXB

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2014-0131

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
88ABW-2015-1753; Cleared 07 April 2015

14. ABSTRACT

Comprehensive and reproducible sample collection techniques were developed concomitantly with the informatics tool and used in multiple, independent studies for the validation and further development of generated software tools and approaches. Data from a dose-response study examining an organ specific environmental toxicant exposure was analyzed using the prototype software tool for discovery of LC/MS based metabolomic biomarkers.

15. SUBJECT TERMS
metabolomic biomarkers, gene expression profiles, post translational modifications

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	114	Claude Grigsby, Ph.D. **19b. TELEPHONE NUMBER** *(Include Area Code)*

THIS PAGE IS INTENTIONALLY LEFT BLANK.

TABLE OF CONTENTS

Section **Page**

LIST OF FIGURES

LIST OF TABLES

Table **Page**

1.0 SUMMARY

The key requirements to any empirically based study are to: (1) accurately measure and then compare the collected results in determining the result of the hypothesis being tested; and (2) collect a sample representative of the entities being studied. To address these requirements and their application to large scale analytical studies, we have developed and utilized an informatics tool for spectral registration, spectral and chromatographic alignment, visualization, and comparative analysis. Comprehensive and reproducible sample collection techniques were developed concomitantly with the informatics tool and used in multiple, independent studies for the validation and further development of generated software tools and approaches. Data from a dose-response study examining an organ specific environmental toxicant exposure was analyzed using the prototype software tool for discovery of Liquid Chromatography – Mass Spectometry (LC/MS)-based metabolomic biomarkers. This data set served as proof of concept in the development and illustration of the novel approach to spectral registration and visualization, and illustrates the rapid multi-sample analysis capability of the informatics tool. A variety of additional studies focused on volatile biomarker discovery, i.e., a murine model of infection to select agents, characterization of human and murine urine as it ages, human markers of age and ethnicity in axillary odors, and characterization of the binding between volatile ligands and murine major urinary proteins aided in algorithm and interface development for Gas Chromotography - Mass Spectometry (GC/MS) functionality implemented in the developed software. The final phase of this work focused on utilization of these analysis tools in combination with novel sampling techniques to create an end-to-end discovery pipeline for large-scale small molecule and volatile organic compound biomarker and differential profiling studies. This combination of biologically and environmentally-focused studies were successfully completed as final proof of concept for this work and demonstrate the universal utility of the approach. The results and data analysis of these five unique sets of experiments using the informatics tools are presented as chapters that answer the hypothesis that an informatics tool can be designed that provides spectral registration, spectral and chromatographic alignment, visualization, and comparative analysis for data generated from multiple analytical platforms, e.g., LC-MS and GC-MS.

2.0 INTRODUCTION

2.1 Overview

The key requirements to any empirically based study are to (1) accurately measure and then compare the collected results in determining the result of the hypothesis being tested, and (2) collect a sample representative of the entities being studied. To address this requirement and its application to large scale analytical studies, the research described here sought to develop and utilize a logically designed and successfully implemented informatics tool for spectral registration, spectral and chromatographic alignment, visualization, and comparative analysis in combination with reproducible and comprehensive sample collection techniques. This work incorporated multiple, independent studies for the development and validation of generated software tools and approaches. To both answer active research needs and aid in the development and illustration of the novel visualization and rapid multi-sample analysis capability of the tool for discovery of LC/MS-based metabolomic biomarkers, use of data from a dose-response study examining environmental toxicant exposures is also described. A variety of studies focused on volatile biomarker discovery were analyzed to aid in algorithm and interface development for GC/MS analysis. Experiments examined in support of this GC/MS analysis included: a murine model of infection to select agents, characterization of human and murine urine as it ages, human markers of age and ethnicity in axillary odors, and characterization of the binding between volatile ligands and murine Major Urinary Proteins (MUPs). The final phase of this work demonstrates utilization of the analysis tools and workflow developed in combination with both optimized and novel sampling techniques to create an end-to-end discovery pipeline for large-scale small molecule and volatile organic compound biomarker and differential profiling studies. A combination of biologically and environmentally-focused studies were assessed as final proof of concept for this analytical workflow and demonstrate the universal utility of the approach. Successful completion of the pipeline have facilitated and allowed the efficient and methodical analysis of multiple, large-scale small molecule and volatile organic compound based biomarker discovery and differential molecular profiling studies.

2.2 Omic Profiling, Sampling Consideration, and Data Analysis Tools

A brief overview of the various nuances in systems biology data space, analysis strategies, and sample collection considerations is required in order to adequately frame development and application of the analytical pipeline for large scale profiling studies which will be the focus of this proposal. Although the sciences of genomics and proteomics are peripheral to much of this research, detail on all three is described as the analysis approaches utilized for each are applicable to all, including metabolomics and volatile profiling. It is worth noting that a rich variety of data analysis and data mining software tools exist for genomics and proteomics in both open source format and from a myriad of vendors, in sharp contrast to the lack of comparable capability in the analysis of small molecule and volatile mass spectrometry-based differential profiling studies. It was this lack in availability for small molecule metabolomic and volatile informatics which served as the impetus and basis for much of the work proposed below.

2.3 Gene Expression Profiling

The unique gene expression profiles in different cells determine their overall functionality and phenotype. Phenotypic changes in response to perturbations resulting from external chemical, biological and physical insults are mediated by and the consequence of a concerted change in the expression of a large number of genes. Therefore, the global gene expression profile represents a unique signature of a specific cell state. By comparing gene expression profiles under different conditions, changes in a gene expression profile may be used for classification and predictive models for a variety of physiological and pathological states, as well as an indicator/predictor for the host response to various external and internal stimuli or insults. As has been previously demonstrated, gene expression changes elicited under different conditions are tightly regulated and highly distinctive [1-6]. Genomics studies have generated useful data highlighting the potential molecular mechanisms involved in many human diseases such as cancer, heart disease, chronic inflammatory disease, etc. [1-4], as well as providing better insights into the mode of action of different chemical toxicants [5, 6]. The Deoxyribonecleic Acid (DNA) microarray is a powerful and versatile tool for global gene expression profiling in biomedical research, which can be used without well-defined underlying hypotheses. It is therefore extremely useful in discovery-driven research in that it can provide mechanistic insights into various physiological or pathological processes. The expression profiles produced by microarray experiments represent the entire transcriptome (or a selected portion of it, dependent on the coverage of the array) of the cells or the tissue under a particular condition.

Analysis of DNA microarray data for biomarker discovery normally involves feature selection and class discovery / class prediction using multivariate statistical methods and/or pattern recognition techniques [7-9]. The objective is to identify informative genes (i.e. a subset of genes that have undergone significant expression changes) for a specific condition, such as the onset of a disease or exposure to a chemical toxicant. Before the data analysis can be completed, the data set has to be “preprocessed” [10, 11]. The first step of data preprocessing involves data normalization to compensate for inter-array variations due to technical reasons. The next step is to remove the genes with low-quality signals using statistical methods or platform-specific algorithms [12]. This will assure that noise of the assay system can be effectively filtered out. Finally, different data transformation techniques (e.g. log-transformation, mean-centering, etc) may be applied to facilitate data manipulation in subsequent data analysis steps [13].

After the data is preprocessed, an exploratory visualization of the data may be performed to examine the general grouping of the samples using an unsupervised clustering method (e.g. hierarchical clustering, K-means clustering, self-organizing maps, or principal component analysis) [14, 15]. Since these unsupervised methods do not depend on any a priori assumption on class repartition, they will provide good information regarding the effects of various variables on gene expression profiles, as well as the presence of potential confounding factors. If the variable(s) of interest appears to be the most dominant factor governing the gene expression profiles, informative genes can readily be identified using traditional statistical methods (i.e. student t-test, Analysis of Variance (ANOVA), etc.) [16]. However if confounding factors with strong effects on gene expression exist, algorithms developed for DNA microarray analysis (i.e. Significance Analysis of Microarrays, Predictive Analysis of Microarray, supervised pattern recognition techniques or machine learning algorithms) [17-21] will be needed with appropriate re-sampling techniques for cross-validation [22, 23]. Despite the inclusion of the cross-validation step, over-fitting of the model may still occur [24-26], which will result in the selection of gene expression changes due to experimental artifacts and noises. Therefore, validation using a fully independent data set should be performed to ensure that the features selected are authentic informative genes.

Identification of informative genes will allow one to focus on a smaller subset of genes with direct relevance to the problem of interest. For instance, genes that are differentially expressed in test groups compared with control groups identified in DNA microarray experiments can be further validated by analyzing in more detail and in larger numbers of subjects using other technology platforms (i.e. quantitative real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR)) [27]. After additional testing and validation using computational and experimental techniques, genes with the highest predictive power can be identified as potential biomarkers for a specific physiological or pathological state.

2.4 Proteomic Profiling

Proteomics is the large-scale study of proteins, particularly their structures and functions. The proteome in both humans and other higher life forms is orders of magnitude more complex than the genome. While the genome is a rather constant entity, the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment. One organism will have radically different protein expression in different parts of its body, in different stages of its life cycle and under different environmental conditions. Numerous post translational modifications combined with amino acid substitutions have been reported for the myriad of proteins coded by the genome. Discovering these differentially expressed protein profiles along with their corresponding Post Translational Modifications (PTMs) is a critical step in understanding the molecular mechanisms regulating function in both single cells and systemically. One of the core tasks assigned to modern proteomics is the identification of differential biological markers that are definitive for various states of organism health (i.e. disease vs. no disease, successful treatment vs. unsuccessful treatment). Because levels of proteins in blood and urine can reflect an individual's state of health or disease, proteomics is especially suitable for the identification of biomarkers [28]. The field of proteomics is still in development and various approaches for separating and identifying proteins are being validated. Additionally, separation and relative quantitation of complex protein mixtures remain two of the most challenging aspects of proteomics, as will be discussed briefly below. Sample preparation is a critical first step in any protein profiling experiment. Both the source of the sample as well as the subsequent processing steps are greatly dependent on the objectives of the study. Detecting and identifying all of the given proteins or peptides in a complex biological mixture such as serum or tissue media is an extremely difficult task due to both the wide dynamic range of protein expression levels (10^{12} variance in concentration) and the differing physical characteristics and resultant behavior of proteins during the separation process. In addition, in a mass spectral analysis, the ion signal of low concentration analytes can easily become saturated by highly abundant proteins, such as albumin, that are present in most raw biofluid samples. This signal suppression compromises the identification of low abundant proteins that are typically of highest interest in a biomarker discovery program. Multiple commercially available devices are available to assist in the depletion of up to the 20 most abundant proteins from serum/plasma samples, thereby facilitating detection of the less abundant proteins present. Other sample processing steps typically include a combination of the following: solubilization, fractionation, equalization, precipitation, desalting, and concentration.

Several techniques are currently in wide spread use for proteomic profiling. Early Mass Spectrometry (MS)-based biomarker profiling had previously been largely focused on the Surface Extraction Laser Desorption/Ionization (SELDI) mass spectrometry technology (see [29] for a comprehensive review). Early publications on the success of this approach to detect disease fingerprints in complex samples were received with some degree of skepticism and critique given subsequent failed attempts to reproduce the results in addition to some obvious flaws in the protocol design [30]. The high dimensionality of the data, small sample sizes, lack of rigorous analysis of technical and biological expressed. However, many reports employing SELDI Time of Flight (TOF) MS-based approaches demonstrated success in the early detection of many different cancers as well as other diseases [31-39]. Two dimensional (2D) gel electrophoresis, despite its 30-year history [40] and the many improvements that have been made to the technology over the years [41-44], is still faced with problems in reproducibility and comparability of individually generated gel images. Even with these challenges, many laboratories have successfully used this approach to sort out differential changes in complex protein mixtures and are routinely able to resolve over 1000 proteins on a single gel, representing an adequate sampling of the expressed phenotype of a specific sample. 2 dimensional (2D) gel separation of proteins also allows for the easy detection of critical posttranslational modifications and when performed on silver-stained gels, 2D gel electrophoresis is highly sensitive, detecting proteins down to the low ng levels. A new approach introduced in the past several years is 2D difference in-gel electrophoresis (2D-DIGE), a technique designed to minimize gel-to-gel variations. With 2D-DIGE, protein extracts from two different samples are covalently labeled with two different fluorescent dyes. The labeled samples are then mixed and subsequently separated on a single 2DE gel. The gel is subsequently scanned at different wavelengths revealing visually differentially expressed proteins as well as PTMs. Another recently emerging and powerful approach employs Beckman-Coulter's Proteome PF 2D system combined with proprietary analogs [45-49]. This two dimensional separation of complex mixtures is performed using Isoelectric Focusing (IEF) in the liquid phase (first dimension) followed by Reverse-Phase High Pressure Liquid Chromatography (RP-HPLC) in the second dimension. On-line Electro Spray Ionization (ESI) TOF mass spectrometry (see Figure 1) enables detection even in third dimension [48]. The accompanying vendor supplied software package generates color-coded protein maps that are easy to interpret and the differential display of protein profiles enables clear visual differentiation of two samples. Unfortunately the existing Beckman software suite does not allow for the comparison of multiple samples as part of a larger study, requiring integration with outside bioinformatics and statistical tools typically relegated to genomic array analysis.

To identify the chromatographically or electrophoretically separated proteins once they are digested with a suitable enzyme (i.e. trypsin), the resultant peptides are subjected to different MS techniques that may or may not include additional on-line liquid chromatography such as a strong cation exchange column to separate peptides before they enter the mass spectrometer.

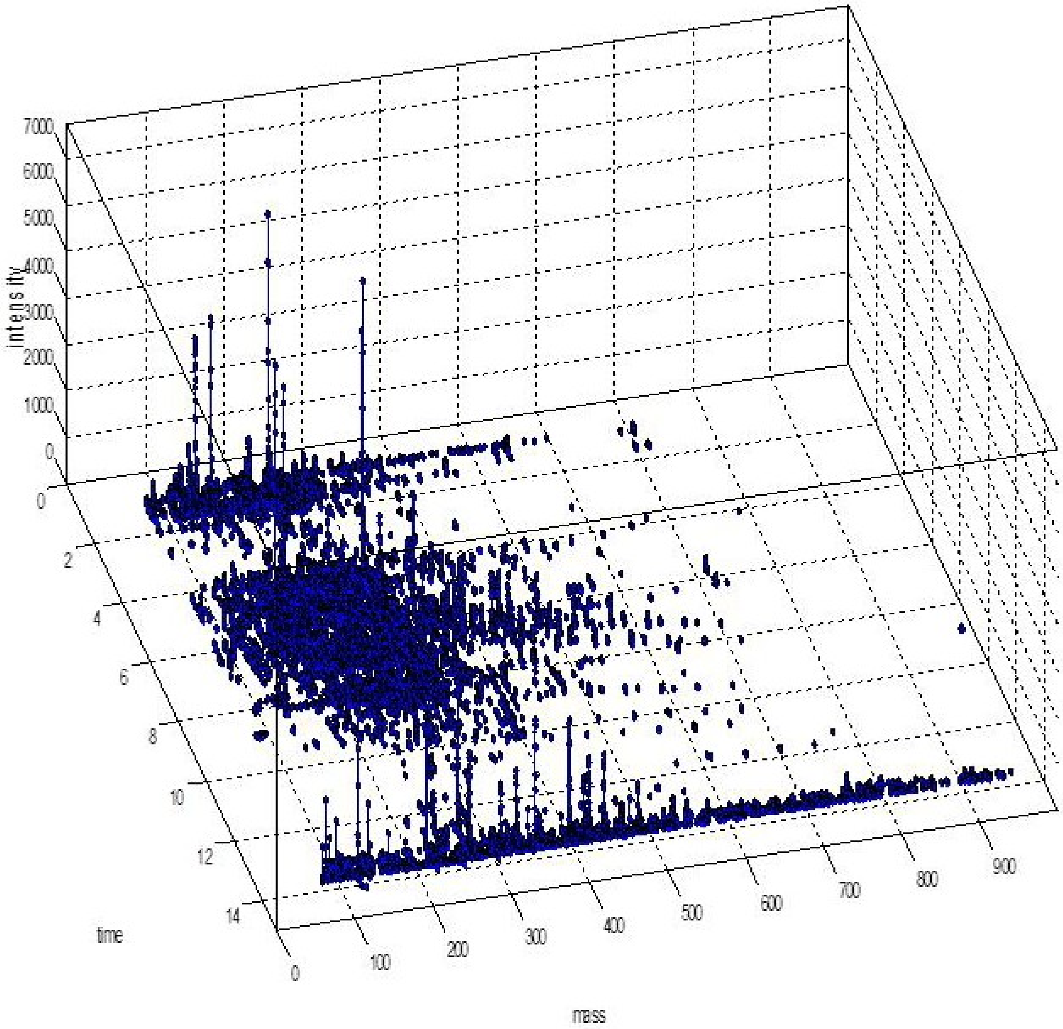


Figure 1: ESI LC/MS Acquired

Example of three dimensionality seen in chromatographic / mass spectral data

The MS techniques used for protein identification can be classified by the type of mass spectrometer being utilized (quadrupole, TOF, ion trap, ion cyclotron resonance etc. as well as hybrid instruments) as well as by the ionization technique used as a source and interface with the HPLC. ESI and Matrix Assisted Laser Desorption Ionization (MALDI) are the two major techniques commonly used by researches for protein profiling, with nano-spray ionization being basically a nano- or reduced volume process of ESI. Nano-spray offers a higher sensitivity versus ESI due to the increased efficiency of ionization [51, 52] and has become the defacto standard for LC-MS based proteomics.

Other widely used approaches for protein profiling include techniques such as 2 dimensional liquid chromatography (2-D LC) and capillary electrophoresis [53]. In these experiments the complex mixtures of proteins are digested with a suitable enzyme (trypsin) followed by two physical separation techniques prior to ionization and MS detection. 2D LC systems typically consist of two columns, the first being a Strong Cation Exchange (SCX) column and the second a Reversed Phase (RP) or C18 column [54-58]. The peptides are initially loaded onto the SCX column and are gradually eluted onto the RP column using increasing salt steps with an aqueous/organic gradient between increasing salt concentrations. This approach has facilitated unambiguous identification of up to 1,500 proteins from one sample [59].

2.5 Metabolomic Profiling

Metabolomics is a rapidly growing field used to characterize the metabolic profile of a specific tissue or biofluid. Metabolic profiling, originally pioneered by Jeremy Nicholson, Elaine Holmes, and John Lindon at the Imperial College in London [60] utilizing Nuclear Magnetic Resonance (NMR)-based analysis, has evolved to become one of the most common applications of LC-MS [61-63]. Metabolomics is an attractive approach to the study of time-related quantitative multivariate metabolic responses to pathophysiological processes by which biological and chemical agents, e.g., drugs, can cause perturbations in the concentrations and flux of endogenous metabolites involved in critical cellular pathways [64]. Thus, cells respond to toxic insult or other stressors by altering their intra-and/or extra-cellular environment in an attempt to maintain a homeostatic intracellular environment.

This metabolic alteration is expressed as a "fingerprint" of biochemical perturbations characteristic of the type and target of a toxic insult or disease process [65]. These metabolic alterations are often seen in body fluids as changes in metabolic profiles in response to toxicity or disease, as the body attempts to maintain homeostasis by eliminating substances from the body. Therefore, because many biofluids can be easily obtained either non-invasively (urine) or minimally invasively (blood), they are typically used in metabolomic studies [66]. Additionally, if a significant number of trace molecules can be identified and monitored, the overall pattern produced may be more consistent and predictive than any single biomarker [67], which would prove of great value in the development of deployable devices for testing toxic or infectious exposures.

2.6 Volatile Organic Compound (VOC)-Based Metabolomic Profiling

A specific area of study within metabolomics is that focused on the volatile metabolites. These VOC- based metabolites generally have a boiling point less than 300°C and contain fewer than 12 carbon atoms[68]. Clinicians frequently associate peculiar body odors with a disease state and for several disorders the odors are distinctive enough to be diagnostic [69, 70]. These odors are often composed of “organic volatiles” such as nutrients, metabolic intermediates, waste products, environmental contaminants, and other compounds of low molecular mass involved in metabolism [68]. The concept of metabolic profiling was summarized by Jellum et al. as follows, "it seems reasonable to assume that if one were able to identify and determine the concentration of all compounds inside the human body, including high molecular weight as well as low molecular weight substances, one would probably find that almost every known disease would result in characteristic changes of the biochemical composition of the cells and the body fluids" [71]. Metabolic disorders are often characterized by the accumulation of a small number of metabolites in body fluids, generally because a deficiency in enzymatic activity blocks the normal biochemical pathway [72].

In the 1970's, studies of volatile metabolites in human urine by Zlatkis and Liebich [73, 74], composed of nearly 200 samples from adults were analyzed by GC-MS and resulted in the identification of the 40 constituents listed in Table 1. The identities of these volatiles were subsequently confirmed by comparing both the retention times and mass spectra of the compounds with known standards. Key components found in the profiles of normal urines are ketones (i.e. 2-butanone, 2-pentanone, and 4-heptanone), dimethyl disulfide, several alkyl furans, pyrrol, and carvone. Pyrazines are present in trace quantities. During a subsequent study performed in 1973 and 1980, mouse urine was shown to contain well over 100 volatile components [75] similar to those in humans, and an additional study in 1989 [76] identified the compounds listed in Table 2 in an examination of dominant and subordinate male mouse urine.

Table 1: Volatile Constituents in Human Urine

(Adapted from [77])

Diethyl ether (solvent)	2-Heptanone
3-Methyl-2-butanone	3-Methylcyclopentanone (tent)
2,3(?)-Dimethylfuran	Limonene
2,4-Dimethylfuran	2-n-Pentylfuran
2 2-Pentanone	4-Ethoxy-2-pentanone
2-Methyl-3-pentanone	Cyclohexanone
3-Methyl-2-pentanone	3-Octanone
4-Methyl-2-pentanone	5 Allyl isothiocyanate
1- Propanol	2-Octanone
2-Methyl-5-ethylfuran (tent)	Acetic acid
Dimethyl disulfide	6 Pyrrole
3-Hexanone	Benzaldehyde
3 2,3,5-Trimethylfuran	2,3-Butanediol
2-Hexanone	γ - Valerolactone
2-Methyl-1-propanol	α -Terpineol
5-Methyl-3-hexanone (tent)	γ - Hexalactone
3- Penten-2-one	Carvone
4-Methylpent-3-en-2-one (tent-)	δ - Hexalactone
4 4-Heptanone	9 Dimethyl sulfone
Cyclopentanone	4-Methyl-5-hydroxyhexanoic acid
2-Methyltetra hyd rofu ran-3-one (tent)	Lactone(tent)
3-Heptanone	ρ - Cresol B

Table 2: Volatile Compounds in Urine of ICR/ALB Male Mice
(Adapted From [76]).

Class of compounds	Structure
Dihydrofurans	mol wt 126
	mol wt 126
	mol wt 126
Ketones	2-Heptanone
	5-Heptene-2-one
	4-Heptene-2-one
	3-Heptene-2-one
	6-Methyl-6-hepten-3-one
	6-Methyl-5-hepten-3-one
	Acetophenone
Acetates	*n*-Pentyl acetate
	2-Penten-l-yl acetate
Dehydro-*exo*-brevicomin	
2-(*sec*-butyl)-4 ,5-dihydrothiazole	
Sesquiterpenes	β-Famesene
	α-Farnesene

2.7 Approaches to Volatile Sampling

Volatile sampling may be performed using a variety of approaches, each of which come with their own advantages and disadvantages in such areas as analyte sensitivity and selectivity of volatiles sampled, form factor (sampler/sensor footprint), and time/cost of analysis. Additional consideration must be given to any volatile study to determine the best approach with the principal concerns being: experimental scale/number of samples, are the compound(s)/subspecies of interest known, and is absolute quantitation required. To understand the implications of the above in the real-world applications described and proposed below, a brief overview of volatile sampling techniques is required.

The majority of air sampling for volatiles performed in both the laboratory setting and environmental field monitoring is a variant of a "grab" sample. Grab samples consist of a single, finite time slice of the atmosphere captured during the sampling time period and provide the concentration of a volatile mixture for only the actual time slice comprising the collection (i.e. a still picture versus video). The time slice may be extended through the manipulation of flow rates across a sensor or modifying diffusion to a media, but inevitably are limited in that the reported concentrations are a time weighted average of the time of collection, which does not allow for discrimination of short duration increases or decreases in analyte concentration. Time-series collections through the sequential acquisition of grab samples is possible as is the continuous, "real-time" monitoring of a volatile sample space through solid state sensors. Both of these approaches have limitations, with space, power, and increased analysis requirements being principal concerns for the time-series grab samples and limited analytical resolution and discrimination being foremost for real-time sensors (i.e. only good for a given analyte or analyte sub-set). Below is a brief description of some of the most commonly used volatile sampling techniques along with their associated advantages and disadvantages.

The gold standard for volatile "grab" sampling is the summa canister, a stainless steel electro-polished (or "summa" polished) passivated vessel, generally in a spherical or cylindrical format. Summa canister are available in a variety of sizes (100ml to 10L), contain a valve and gas tight fitting for sample inlet, and are "re-conditioned" (cleaned) and evacuated prior to collection. Additional accessories which can be used in conjunction with summa canisters to allow for time-weighted, longer duration collections are flow controllers and critical apertures/orifice assemblies, and analog gauges. For typical air grab sampling using summa canisters, the summa canister valve is opened and the canister is left in a designated area for a period of time to allow the surrounding air to fill the canister and achieve a representative sample. The valve is then closed and the canister is sent to a laboratory for analysis. Summa canisters are typically analyzed via GC/MS according to the Environmental Protection Agency's (EPA) guidelines for air, such as methods TO-14 and TO-15 [78] and are the only way to trap the lightest of the freons and inorganic (permanent) gases such as O_2, O_3, CO_2, SO_2, NO_2, etc. However, compared to other grab sampling techniques, summa canisters are more expensive, have a large physical and logistical footprint, require manual sampling and additional high-cost equipment to be compatible with most GC systems, and are limited in their volatile capture range to compounds with 10 carbons or less [79]. Another common approach to volatile grab sampling are gas sampling bags such as the Tedlar® bag. Tedlar® bags require either manual (i.e. hand pump) or active pumping of air using a portable sampling pump into these relatively less expensive bags which can then be returned to the lab for analysis. Analysis is generally performed either by injecting the air in the bag directly into a sorptive device to concentrate low levels of VOCs, followed by desorption into the GC/MS, or by aspiration and injection via a gas tight syringe into a sealed headspace vial (less sensitive) and analysis by GC/MS. Principal limitations for gas bag sampling are limited sample stability and associated difficulties in sample transport/shipment, large sampling footprint, and volatile condensation on the bag surface requiring heating immediately prior to analysis.

Another approach to air sampling is the passive collection of volatiles onto a sorbent media. Passive sampling is widely used in both the industrial and environmental setting and is particularly useful in many applications that are space limited and/or require long term monitoring, such as in workplace personnel assessments. The devices used for passive sampling are usually based on diffusion through a well-defined diffusion barrier or permeation through a membrane. The analyte concentrations obtained using the passive samplers are a time weighted average, and are limited to the capture range of the media utilized. One of the most commonly utilized passive samplers is the badge type device, which is commercially available from numerous manufacturers such as 3M and SKC [80]. These devices do not require any active sampling pumps but generally need exposure to the sample air for considerably longer times (4 hrs or more) for low level detection. Quantitation requires calculation based upon published uptake rates for each individual volatile on a given passive sampler and both open source literature [81] and recent consultation with National Institute for Occupational Safety and Health (NIOSH) have shown that this sampling approach has been proven less reliable and reproducible for quantitative studies than actively, pumped samples.

One of the most commonly used passive media formats is Solid Phase Microextraction (SPME). SPME, which has other commercial analogues such as SPE-td (or SBSE aka "Twister") is a licensed, passive sampler available from Supelco (Bellefonte, PA). Most GC sample preparation procedures using based on solvent extractions have been time-consuming, labor intensive, multi-stage operations. All of these steps, especially concentration via solvent evaporation, introduce errors and losses, especially when analyzing volatile compounds. Additionally, waste solvent has to be disposed of, adding to the expense of the procedure. Many of these limitations have been reduced through the use of Solid Phase Extraction (SPE), but this technique is still time consuming and generally requires a concentration step. SPME was invented by Pawlisyzn at the University of Waterloo (Ontario, Canada) in 1989 [82, 83], and addresses these limitations by integrating sampling, extraction, concentration, and sample introduction into a single step. SPME utilizes a short, thin, solid rod of fused silica (typically 2 cm long and 0.11 mm OD) coated with an adsorbent polymer. The fiber is the same type of chemically inert fused silica used to make capillary GC columns and is stable at high temperatures [84]. The SPME extraction consists of two processes: analytes partition between the sample and the fiber coating; and the concentrated analytes desorb from the coated fiber to an analytical instrument. SPME extraction is a complex multiphase equilibrium process. An extraction can be considered complete when the concentration of analytes has reached distribution equilibrium between the sample and coating (amount extracted is independent of further increases in time). The higher the distribution constant of a compound, the higher the affinity of that compound for the SPME fiber coating [85]. In general, the distribution constant for an analyte increases with increasing molecular weight and boiling point. The equilibrium conditions can be described as: $n = \frac{K_{fs} V_f V_s C_0}{K_{fs} V_f + V_s}$ where n is the amount extracted by the coating, K_{fs} is a fiber coating/sample matrix distribution constant, V_f is the fiber coating volume, V_s is the sample volume, and C_0 is the initial concentration of a given analyte in the sample [85].

There also exists a linear relationship between the amount of the analyte on the fiber extracted and its initial concentration $C_{s,0}$ in the sample if the extraction time is long enough for equilibration, the dynamic range of the method is not exceeded and all other experimental conditions (sample composition, temperature, volumes of sample, and headspace) are held constant [85]:

$$n = \text{const x } C_{s,0} \text{ where } n = \frac{K_{fh} K_{hs} V_f V_s}{K_{fh} K_{hs} V_s + K_{hs} V_h + V_s}$$

The above equation still holds true if the extraction is interrupted after a constant time before the equilibrium is attained [86]. In SPME, detection limits depend on the distribution constants and polarity of the analytes but are often in the ppt. For example, with a 100 um polydimethylsiloxane (PDMS) fiber, nonpolar compounds with high distribution constants have lower minimum detection limits than more polar analytes with lower distribution constants. Branched aromatic compounds and chlorinated alkenes exhibit the lowest detection limits [87].

As the extremely volatile compounds can begin to leave the fiber before the sample is injected, it is best to minimize the time between extraction and desorption and to maintain consistent timing for each step. Some of the advantages of SPME over other approaches such as summa canisters are: it is a passive sampling technique and takes only a few minutes for equilibration; the "clean" sampling matrix provides an accurate volatile baseline; lower sample cost and smaller footprint versus a summa canister; and will detect material with higher boiling points than benzene, like engine or mineral oil hydrocarbons [88]. Significant downsides to SPME are that sufficient analyte may not be absorbed by the SPME fiber for quantitation if the total concentration is very low, SPME volatile mixture capture is limited low sample capacity and competitive displacement, there is no easy way to automate analysis for high volume testing, and sample breakthrough/loss becomes a large concern for extended exposures or in a high flow environment and limits stability on fiber. Additionally, only one analysis per sampling event is possible as the fiber is completely reconditioned after each injection. With other techniques listed above (i.e. summa canister and gas bag), the analyst may perform multiple trials / sample injections into GC/MS with just one sampling event. For the field sampling pipeline proposed in this study, the fibers would require transport at -4°C. Even under these conditions, prior experience in our laboratory [89] has demonstrated that it is likely not possible to store the SPME fiber for three days or more post sampling and recover highly volatile organic compounds such as chloromethane, but the technique does remain viable for less volatile compounds (i.e. naphthalene and dichlorobenzenes). It must also be noted that after discussions with scientists directly involved with development of SPME in the laboratory of Dr. Janusz Pawliszyn [90], the use of sorption tubes, described below, is certain to be more quantitatively accurate than SPME due to the competitive binding nature and limited surface area of the SPME fibers.

Arguably the most versatile and accurate approach to unknown atmospheric mixtures characterization are the active sampling techniques. Active sampling requires forced airflow through the collection media using either a pump or in-line series/parallel flow. The most commonly utlized active media format in environmental monitoring for this purpose is the Thermal Desorption (TD) tube.

Although TD tubes may be used in the passive role (see below), they are primarily used in conjunction with an actively pumped system. These tubes are analogous to thermal desorption traps used in GC/MS to concentrate VOCs in a headspace analysis or trap columns used in liquid chromatography. TD tubes allow for collection of volatiles by channeled air flow though the capture packing material using miniaturized sampling pumps at the testing site. The TD tubes are then either thermally desorbed (i.e. via commercial autosamplers) or solvent desorbed for analysis. When utilized for passive sampling, as gases will diffuse in (and out of) both ends of the tube, all the passive sampling work is generally done with a single sorbent bed in the tube. Doing this limits the boiling point range covered, making this method somewhat constrained in that the analyte being measured should be previously determined. As diffusion drives this flow, breakthrough is not a concern, however, low boiling materials may diffuse back off of the tube (i.e., diffusion is a two way process). This is particularly important if the nature of the gas being monitored changes and has a lower concentration of the target analyte than the original gas. Depending on the duration for which the sample air is pumped, sorption tubes can either be considered as grab sampling or time weighted average sampling (i.e. ~5mins can be considered as grab sampling vs. 4hrs at low flow rates considered as time weighted average sampling). One feature which contributes greatly to the use of TD tubes for unknown or mixtures detection is the availability of multi- phase packing materials capable of capturing a wide range of volatiles, such as in the tribed tubes recommended for use in EPA method TO-17 ("Air Toxics"), example shown in Figure 2. It is due to this versatility and spectrum of commercially available sorbents that TD tube use is widely reported for breath sampling for biomarker discovery [91, 92].

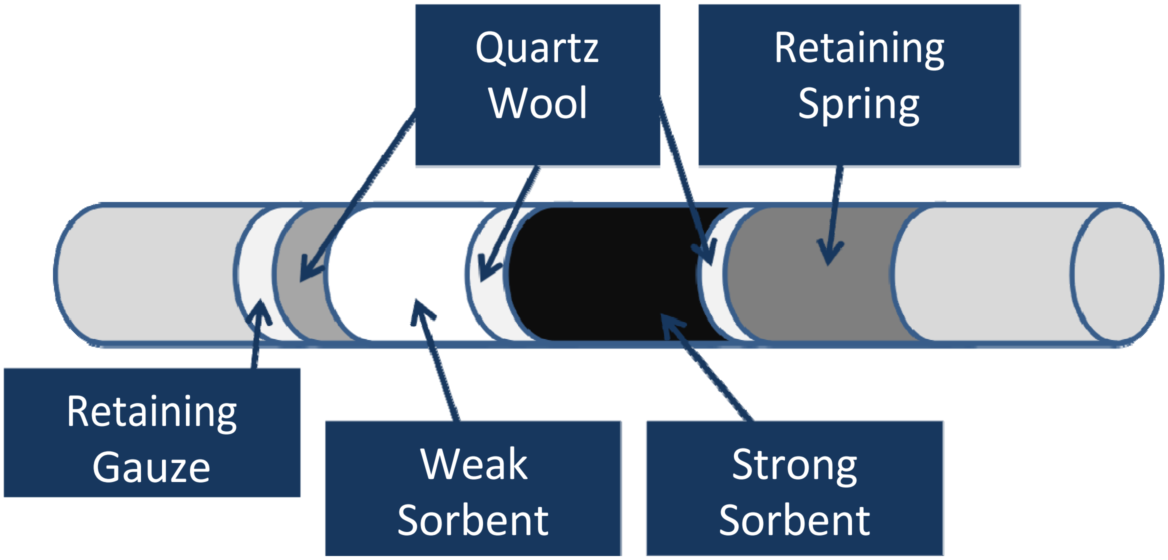
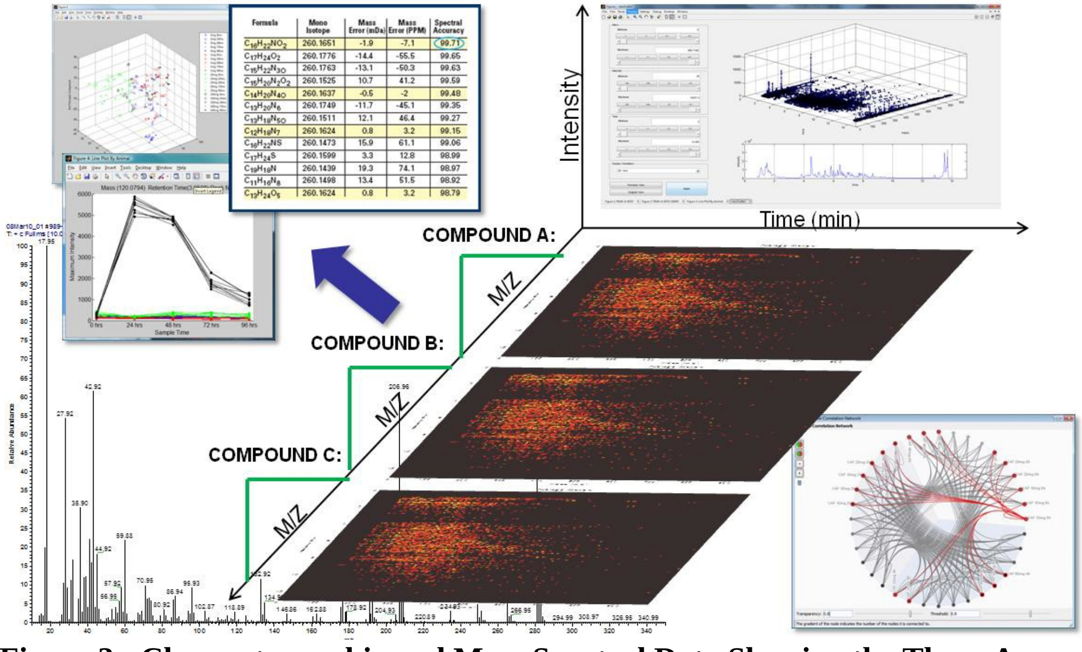


Figure 2: Multiple Sorbent Beds Present in Tri-Bed Thermal Desorption Tube Utilized for EPA Method TO-17

2.8 Data Analysis

To fully appreciate the parallels between GC-LC/MS and the other -omics data sets described above and the associated requirements for post-acquisition analysis, a basic understanding of the raw chromatographic and spectral data obtained is needed. Using the example of an acquisition performed by GC/MS, a typical sample consisting of an unknown mixture of interest is introduced into the heated sample injection port (i.e. via SPME) and flushed by an inert carrier gas, generally helium, onto the head of the capillary column. The columns used by our laboratory are Fused Silica Open Tubular (FSOT) format columns and consist of a fused-silica capillary tube whose walls are coated with liquid stationary phase. Without delving too deeply into issues surrounding stationary phase selection, thickness, and their relevance to the principals of optimized theoretical plate separation, column types are determined based on the physical property selected as the best candidate for separation of the individual components of the mixture of interest (i.e. polarity or hydrophobicity) as well as the sample capacity required. As the complex mixture is pushed through the column and interacts with the bound stationary phase, individual components are pulled apart based on the physical property of the column type, and under optimal conditions, elute as discrete, Gaussian peaks into the detector (mass spectrometer) over a chromatographic time scale, generally measure in minutes, giving the first dimension of each sample run. For the two, additional sample dimensions, we must look at the detector utilized, the mass spectrometer. According to Silverstein [93], the concept of mass spectrometry is relatively simple: a compound is ionized (i.e. via electron impact, or EI, in the case of GC/MS), the ions are separated on the basis of their mass/charge (m/z) ratio, our second dimension, and the number of ions representing each m/z unit is recorded as a spectrum. The intensity of each individual ion can be correlated to relative intensity, and it is this intensity which gives us our third dimension. When all three dimensions are viewed either as a heat map or surface (Figure 3), the data feature similarities between GC-LC/MS, genomic (i.e. gene-chip), and proteomic (i.e. 2D-SDS PAGE gels) become readily apparent. Analysis issues such as image alignment, feature registration, data visualization, and differential profiling are all required, and lessons learned in the more mature fields, such as genomics, are directly applicable to the analysis of metabolomic data.



Formula	Mono Isotope	Mass Error (mDa)	Mass Error (PPM)	Spectral Accuracy
$C_{16}H_{22}NO_2$	260.1651	-1.9	-7.1	99.71
$C_{17}H_{24}O_2$	260.1776	-14.4	-55.5	99.65
$C_{15}H_{22}N_3O$	260.1763	-13.1	-50.3	99.63
$C_{15}H_{20}N_2O_2$	260.1525	10.7	41.2	99.59
$C_{14}H_{20}N_4O$	260.1637	-0.5	-2	99.48
$C_{13}H_{20}N_6$	260.1749	-11.7	-45.1	99.35
$C_{13}H_{18}N_5O$	260.1511	12.1	46.4	99.27
$C_{12}H_{18}N_7$	260.1624	0.8	3.2	99.15
$C_{16}H_{22}NS$	260.1473	15.9	61.1	99.06
$C_{17}H_{24}S$	260.1599	3.3	12.8	98.99
$C_{19}H_{18}N$	260.1439	19.3	74.1	98.97
$C_{11}H_{16}N_8$	260.1498	13.4	51.5	98.92
$C_{13}H_{24}O_5$	260.1624	0.8	3.2	98.79

Figure 3: Chromatographic and Mass Spectral Data Showing the Three Axes *of time, m/z, and intensity and example analysis techniques utilized in differential molecular profiling studies*

3.0 RESEARCH QUESTION AND DESIGN

3.1 Problem Statement

To illustrate the data gap in both sampling and analyzing real-world needs in molecular profiling, following is a widely reported issue identified with the United States Air Force (USAF) aircraft where numerous knowledge and capability gaps exist in the molecular characterization of the operational environments faced by today's Airmen. In 2011, the entire F-22 fleet was grounded for nearly five months due to a series of incidents where pilots had experienced unexplained symptoms such as shortness of breath, disorientation, confusion, and headache, among other symptoms. This led to a myriad of formal investigations that have identified aircraft systems limitations as well as a need to improve our analytical toolset and understanding of the environments encountered in the F-22 and other high performance aircraft and their effects on the humans operating them. In the midst of these high profile investigations, ground maintenance personnel have also had a series of incidents where they have experienced symptoms such as dizziness, headache, and nausea while performing engine ground runs. These incidents have had a similar response effort undertaken for them where possible, to include environmental samples out of the cockpit air supply, but have found no specific cause from that methodology. Unfortunately, the response methodology was tailored to in-flight emergencies and given that testing capabilities are more numerous on the ground, this issue has not been explored as much as possible. A comprehensive exploration of the associated chemical environments will require both the application of existing sampling tools (described above), development of novel analysis techniques, and improvements in large scale environmental differential profiling. Additionally, extensive evidence exists on the uniqueness and dynamics of molecular, and particularly mass spectral, signatures of liquids and gases for use in identification of materials, phenotypic states, and environmental conditions. These operational concerns must be addressed through the development and application of both novel sampling tools and novel applications of available media for the detection and identification of potential contaminants and other performance degrading conditions experienced both on the ground and in-flight. The scope and scale of these studies will also be limited due to the availability of informatics tools capable of the 1000's of generated sample sets. Current LC/MS and GC/MS systems typically consist of a system of specialized instrumentation with customized support software. This software is generally proprietary, being supplied by the instrument manufacturer, and is primarily designed to facilitate user interaction with the analytical hardware. Most manufacturers also market add-on commercial software packages for the analysis of the results of LC/MS and GC/MS experiments, which provide a limited tool set for a specific type of data analysis (i.e. proteomic or metabolomic), generally with a focus on pharmaceutical development needs, and which cannot be readily modified or added to by the end-user as various study needs arise. For larger molecular profiling and biomarker discovery studies, such as the LC-GC/MS efforts undertaken by our laboratory, none of the software solutions reviewed [94-99] offer the ability to compare multiple time point and exposure groups, or handle data sets in significant sample numbers. Due to the need to implement a wide spectrum of differential algorithms required for individual studies as well as the undocumented "black box" nature prevalent in much of the vendor and recently available open source tools/algorithms (possibly inducing unknown bias in obtained results), creation of a novel, modular informatics tool set allowing processing of the raw, chromatographic and spectral data was required.

3.2 Hypothesis

The core hypothesis tested during the course of this work is that an informatics tool can be designed that provides spectral registration, spectral and chromatographic alignment, visualization, and comparative analysis for data generated from multiple analytical platforms, i.e. LC-MS and GC-MS. Comprehensive and reproducible sample collection techniques were developed concomitantly with the informatics tool and used in multiple, independent studies for the validation and further development of generated software tools and approaches.

3.3 Technical Objectives

As stated earlier, the key requirements to any empirically based study are to (1) accurately measure and then compare the collected results in determining the result of the hypothesis being tested, and (2) collect a sample representative of the entities being studied. For measurement in metabolomic studies focused on volatile and small molecule profiling (i.e. 41-1000 m/z), the standard method of approach is GC or LC/MS with associated support software. In addition to the proprietary instrument driver software supplied by the instrument manufacturer, most GC-LC/MS manufacturers also market add-on commercial software packages for the analysis of the results of MS experiments, which are meant to fill the gap in data processing and provide a very specific type of analysis (i.e. proteomic or metabolomic for drug discovery/metabolism) and cannot be modified or added to by the end-user. For larger metabolomic profiling studies, such as the efforts completed in support of this work, none of the software solutions available at the time this effort was initiated offered the ability to compare multiple sample groups and handle data sets in significant sample numbers (i.e. >500 samples). This bottleneck in data handling was the impetus for the development of the informatics tools generated. The second requirement for the creation of this analytical pipeline was the design and execution of novel capture techniques and/or the down-selection of the best commercially available (if experimental criteria are met), with a focus on collecting comprehensive samples to meet the need for the complete characterization of the samples/environments of interest. In summary, the overall objectives of this work were to develop and implement an implement an analytical pipeline to: (1) identify differential molecular signatures and biomarkers of materials, phenotypic states, environmental conditions, and individual/group differences; (2) enhance existing mass spectral differential profiling capability, allowing for optimally targeted unknown compound identification for potential subsequent incorporation into sensor platforms; and (3) conduct associated field sampling in support of operational needs. We achieved the objectives of this work through pursuit of the following three specific aims.

3.4 Specific Aim One

To test the hypothesis that the design and utilization of a novel, prototype software tool for feature registration, and spectral and chromatographic alignment will facilitate analysis of large scale studies for LC/MS based biomarker discovery and small molecule profiling, and allow the ability to visualize the data for a global view of an entire experiment, while still maintaining the ability to focus on individual metabolites and spectra for subsequent identification. In support of this aim, a prototype software tool (described below) was designed in Matlab 2010a (The MathWorks Inc., Natick, MA) for LC/MS based spectral registration and alignment. A preliminary data set from a biomarker discovery experiment to identify low lever markers of organ specific damage was utilized for this proof of concept study. LC/MS-based metabolomic analysis was performed using the Waters Acquity® ultra-performance liquid chromatograph coupled to a Waters QToF® hybrid tandem quadrupole/time of flight mass spectrometer equipped with a Lockspray electrospray source operating in positive ion mode. The resulting spectra were analyzed using this prototype software tool.

3.5 Specific Aim Two

To test the hypothesis that the combination of a suite of basic comparative algorithms, logical operators, and statistical filters in conjunction with an integrated set of machine learning tools for feature down-selection will allow for the efficient analysis of GC/MS based spectral data to support volatile biomarker and gas phased-differential molecular profiling studies. In support of this aim, modifications and enhancements were completed on the prototype software tool created in Aim 1 to allow for the visualization and analysis of GC/MS acquired experimental data in support of multiple VOC-based biomarker and molecular profiling research. Initial GC/MS based metabolomic analysis were performed using a Thermo Scientific (Waltham, MA) Trace GC Ultra® gas chromatograph interfaced to a Thermo Triplus® autosampler configured for automated SPME headspace sampling and in-line with a Thermo DSQII® single quadrupole mass spectrometer. Further validation of this tool and approach to analytical studies were conducted on a variety of volatile profiling studies, to include phenotypic characterizations of a urine murine model, as well as other data sets, including environmental studies, to demonstrate versatility of the approach.

3.6 Specific Aim Three

To test the hypothesis that the creation and execution of a logically designed work flow for large-scale analytical studies in differential molecular profiling, essential in both environmental and biological screening studies, requires not only an efficient data analysis pipeline, but also the design and execution of novel capture techniques focused on collecting comprehensive samples, allowing for the complete characterization necessary. Studies which require complete characterization of an unknown, whether it is biologically derived, such as human breath, or environmental, as in the case of fighter cockpit air, necessitate varied approaches to collection for accurate and broad-spectrum capture of all potential molecular species present. In support of this aim, a thermal desorption tube based approach to sample collection for in-flight, pilot air supply was performed with follow-on analysis being completed utilizing a modified version of EPA Method TO-17 for monitoring VOCs via automated, cryogen-free TD using a Markes Intl TD100® in line with a Thermo Scientific Trace GC Ultra® and ISQ® single quadrupole mass spectrometer. The thermal desorption GC/MS system was calibrated with a standard of target 65 VOC analytes, including selected aliphatic and aromatic hydrocarbons, ketones, alcohols, esters and halogenated organics such as Freons and chlorinated solvents. In addition to the target compound analysis, we identified detected compounds and estimated quantities for all major non-target chromatographic peaks using mass spectral library search procedures.

3.7 Experimental Design

As the goal of this work was to design and implement an end-to-end sample collection to results pipeline, to include a prototype software tool, incorporation of multiple, independent studies were required for development and validation of generated tools and approaches. The principal bottleneck to any study of this type is the data processing and analysis, thus the first aim was to address this through the development of a prototype spectral analysis platform. To both answer an active research need and aid in the development and illustration of the novel visualization and rapid multi-sample analysis capability of the tool for discovery of metabolomic biomarkers, use of data from a dose-response study examining environmental toxicant exposures was performed.

Table 3: Foci, Data Types, and Major Experiments Completed in Support of Aims

	Focus	Data Type(s)	Supporting Experiment(s)
Aim 1	Foundational software framework to include spectral registration and alignment	LC/MS	Small molecule biomarker discovery in F344 model for low level organ selective toxin exposure
Aim 2	GC/MS data functionality - chromatographic time binning and fold change	GC/MS	Characterization of the binding between volatile ligands and murine major urinary proteins Volatile characterization of murine urine as it ages Volatile characterization of human urine as it ages
	Data filtering and normalization techniques	GC/MS	Human volatile markers of age and ethnicity in axillary odor profiles
	Novel hybrid evolutionary classifier for biomarker discovery	GC/MS	Murine model of urine based volatile makers of infection to select agents
Aim 3	Novel sampling technique for comprehensive assessment of volatile contaminants	GC/MS	In-flight air quality assessment of the combat air fleet using a novel sampling approach

The next goal of this study was to expand both the capabilities of the informatics tools and develop a suite of optimized techniques for analysis of volatile based metabolites. Using the same approach as for the LC/MS data analysis, a variety of studies focused on volatile biomarker discovery were processed with the enhanced informatics tool to: aid in algorithm and interface development; provide validation of generated results; and to prove its utility. To demonstrate the various processing, algorithmic, and data filtering requirements addressed, a wide range of study types were processed. Experiments examined in support of this aim included: a murine model of infection to select agents, characterization of human and murine urine as it ages, human markers of age and ethnicity in axillary odors, and characterization of the binding between volatile ligands and murine MUPs.

The final goal of this work was to utilize the analysis tools and workflow developed in the first two aims in combination with both optimized and novel sampling techniques to create an end-to-end discovery pipeline for large-scale small molecule and volatile organic compound biomarker and differential profiling studies. An environmental health and safety focused study of the cockpit atmosphere was assessed as final proof of concept and demonstrate the universal utility of the approach.

4.0 FOUNDATIONAL SOFTWARE FRAMEWORK

4.1 Overview

The goal of this work was to design and implement a prototype software tool for the visualization and analysis of small molecule metabolite GC-MS and LC-MS data for biomarker discovery. The key features of the Metabolite Differentiation and Discovery Lab (MeDDL) software platform [100] include support for the manipulation of large data sets, tools to provide a multifaceted view of the individual experimental results, and a software architecture amenable to modification and addition of new algorithms and software components. The MeDDL tool, through its emphasis on visualization, provides unique opportunities by combining the following: easy use of both GC-MS and LC-MS data; use of both manufacturer specific data files as well as network Common Data Form (netCDF); preprocessing (peak registration and alignment in both time and mass); powerful visualization tools; and built in data analysis functionality.

To illustrate the novel visualization and rapid multi-sample analysis capability of MeDDL for discovery of metabolomic biomarkers, data from a portion of a study examining environmental toxicant exposure has been selected. Environmental exposures to toxins as well as therapeutic interventions often cause nephrotoxicity[101]. An expanded list of metabolites indicating kidney damage would be immensely helpful in the monitoring of renal conditions after exposure to external toxicants, not only in pharmaceutical drug safety evaluations and clinical studies, but also in the occupational and military operational setting.

As reported previously by our group [102], D-serine is ubiquitous in human plasma and composes up to 3% of total plasma serine level in humans, with plasma D-serine elevations observed in chronic renal failure, suggesting elimination by the kidney is responsible for control of D-serine concentrations. D-serine is reabsorbed in the pars recta region of the rat proximal tubule and subsequently metabolized by D-aminoacid oxidase (D-AAO), to produce α–keto acid, ammonia, and hydrogen peroxide [103, 104]. Other research has indicated that metabolism of D-serine by D-AAO is causative for initiation of toxicity in the kidney, with elevated levels generating selective necrosis to the pars recta region of the renal proximal tubules in the rat[105]. The choice to use the D-Serine model was made in order to reveal both early and sensitive biomarkers for epithelial cell injury in the kidney.

4.2 Urine Samples and Materials

Animal use in this study was conducted in accordance with the principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 1996, and the Animal Welfare Act of 1966, as amended. Male Fischer 344 rats weighing 222–258 g were obtained from Charles River Laboratories. Groups of five animals received a single intraperitoneal (IP) dose of d-serine at a dose of 5, 20, or 500 mg/kg (or vehicle only - 0.9% saline solution). Food (Purina Certified Rat Chow # 5002) and water was available for all animals ad libitum. The housing environment was maintained on a 12-hour light-darkness cycle at 25°C, and all animals were examined by Vivarium personnel twice daily. Urine samples were collected cold using plastic 50 mL conical tubes containing 1.0 mL of 1% sodium azide maintained at 6-10°C using I-Cups (Bioanalytical Systems, Inc.; stored at -80°C prior to use) 24 hours prior to dosing and daily thereafter, generating five 24 hour intervals (0, 24, 48, 72, and 96 hours post-dosing). The urine was then frozen at -20°C and thawed on ice prior to analysis. For the D-serine exposure set described, 104 individual samples were processed by aliquoting 1.0 ml of urine into a 2 ml centrifuge tube and centrifuged at 13,000 RPM for 5 minutes at 5°C to remove debris. The supernatant was removed using a 1 mL tuberculin syringe and filtered through a 0.2 μm Polytetrafluaroethylene (PTFE) syringe filter disc prior to aliquot transfer to two Waters Corp. Total Recovery Vials and subsequent duplicate testing.

4.3 Instrumentation and Methods

The LC/MS system utilized for sample analysis was a Waters Q-ToF Micro in line with a Waters Acquity Ultra Pure Liquid Chromatography (UPLC). The source temperature was set to 130°C, a desolvation gas temperature of 320°C and a desolvation gas flow of 600 l/h were employed. The capillary voltage was set at 3.2 kV for both positive and negative ion mode analysis. A scan time of 0.4 s with an inter-scan delay of 0.1 s was used throughout and data was collected in centroid mode. A 1μl aliquot of filtered urine was injected onto a 2.1x100mm, 1.7μm Acquity UPLC BEH C18 column (Waters Corporation) held at 40°C. Retained small molecules were eluted via a linear gradient of 98% A for 2 min, 2–50% B from 2–11 min, 50–98% B over 12-12.49 min, returning to 98% A at 12.5 min and remaining there until completion of the run at 15 minutes at an eluent flow rate of 0.25 ml/min; where A = 0.1% formic acid and B = 0.1% formic acid in acetonitrile. The mass spectrometric data was collected in full scan mode from m/z 80 to 1000 from 0.8 to 15 minutes. Urine samples were run in duplicate and analyzed using MeDDL using spectra from 0.8-12 minutes. For ms/ms data, random urine samples were run using data dependent acquisition with multiple voltages applied. Standards were purchased from Sigma-Aldrich (St. Louis, MO) and run at 1 mg/ml (1ug injection) under the same LC-MS conditions as the samples to validate retention times and ms/ms spectra. Sample analysis for determination of differential metabolites was performed using the MeDDL tool which is described below.

4.4 Algorithms and Implementations

The overall goal of the MeDDL system is to facilitate the analysis of LC-MS experimental results. With this goal in mind, the system is structured to provide a global view of experimental results so a user can quickly identify samples exhibiting interesting or unusual patterns of behavior while still having the option to probe these samples at ever finer levels of detail. MeDDL accomplishes this by allowing the user to search for relationships between subsets of subjects at selected times or treatment levels. The user may ask for subsets which exhibit specific levels of change in the behavior of the response. The user may restrict the fold-change to positive, negative or combined levels of changes. For example, the user can seek all peaks that exhibited a 5 fold positive change between the control subjects and treated subjects at 24 or 48 hours. In addition, MeDDL also allows the user to perform detailed statistical analysis including ANOVA (1-way, 2-way and N-way) among the selected subject groups. The user can optionally perform multiple pairwise comparison tests among the means of groups to determine whether or not all differences- among group means satisfy a user defined level of significance. A Bonferroni correction is applied to compensate for the tendency to incorrectly find a single pairwise significant difference among multiple comparisons.

The MeDDL system is composed of two major subsystems: peak analysis and visualization. Peak analysis encompasses several subsidiary tasks including peak extraction, peak registration and extraction of registered peaks sets. The visualization system takes the information provided by the peak analysis subsystem and combines it with information describing the overall experiment to allow the user to explore the results from the perspective of the experimental parameters.

4.5 Peak Analysis Subsystem

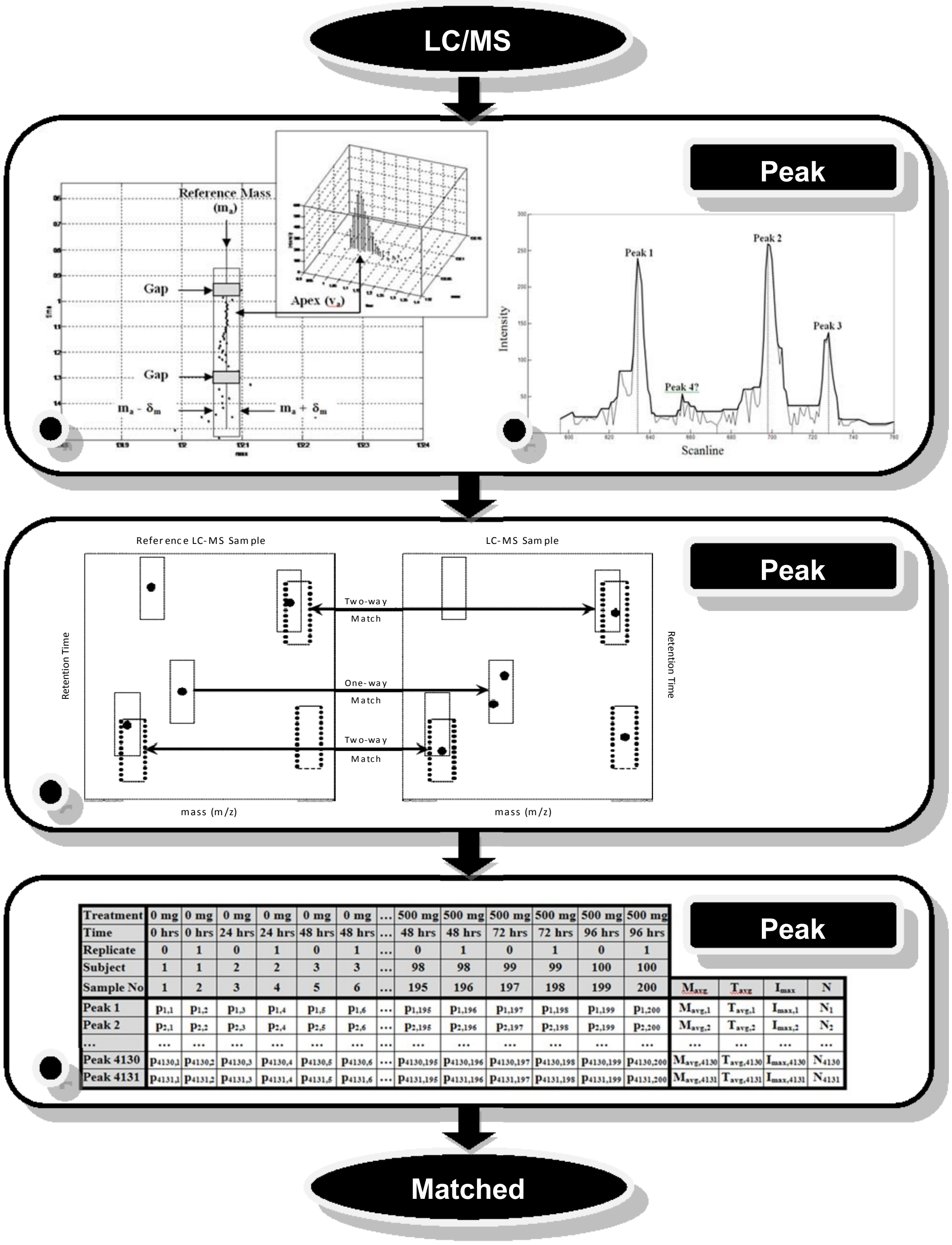
In this section, we present a brief overview of the algorithms that comprise the peak analysis subsystem. As shown in Figure 4, these algorithms include peak extraction, peak registration and peak matching. Each algorithm is described below.

The peak extraction algorithm is composed of two phases. The first phase is designed to form temporal clusters required for chromatographic time alignment and the second phase partitions clusters into individual peaks. A full experiment consists of hundreds of files from different LC-MS sample runs, each of which is identified by subject, treatment, time, and optionally replication indices. A single LC-MS sample is composed of a set of n measurement points P = [p_i | i = 1, 2, 3, … n] of the form:

p_i = ($M_{i,}L_i$, I_i) with components of mass (M_i), scan number (L_i) and intensity value (I_i). Each scan (L_i) also has a corresponding retention time (T_i). Extracted peaks are temporal sequences of similar mass coordinates across multiple scans. A simple example of the algorithm for forming peak clusters is shown in Figure 4A.

The peak extraction process is initialized by selecting a reference point (M_a, L_a, I_a) with a large intensity which will become the apex of the resultant peak bounded by a narrow mass band. The width of the mass band is set by a mass uncertainty parameter (Δ_m) specified by the user. Within this mass-band, points are assembled into cluster sequences in both temporal directions from the initial reference point, accepting only one point per scan. A resultant cluster may grow to lengths spanning many peaks.

The next step, partitioning the cluster into individual peaks, is a difficult design problem, because it must instantiate a peak definition that separates the significant peaks from the noisy and uninteresting ones. Often, partitioning a cluster visually can be difficult, so some ambiguous results are unavoidable. In other words, if it is difficult to resolve peaks visually, it is difficult to automate. In Figure 4B there are many small, jagged, noisy peaks and three or four prominent peaks. The decision to extract 3 or 4 peaks is determined by adjusting a user-accessible control parameter. In this example, a closing operator from mathematical morphology [106] has been employed to filter out unwanted peaks, including the fourth obvious candidate. The horizontal fill-in lines are determined by the size of the structuring element used by the morphological operators to probe the cluster's structure.



Treatment	0 mg	0 mg	0 mg	0 mg	0 mg	0 mg	...	500 mg	500 mg	500 mg	500 mg	500 mg	500 mg				
Time	0 hrs	0 hrs	24 hrs	24 hrs	48 hrs	48 hrs	...	48 hrs	48 hrs	72 hrs	72 hrs	96 hrs	96 hrs				
Replicate	0	1	0	1	0	1	...	0	1	0	1	0	1				
Subject	1	1	2	2	3	3	...	98	98	99	99	100	100				
Sample No	1	2	3	4	5	6	...	195	196	197	198	199	200	M_{avg}	T_{avg}	I_{max}	N
Peak 1	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	$p_{1,4}$	$p_{1,5}$	$p_{1,6}$	...	$p_{1,195}$	$p_{1,196}$	$p_{1,197}$	$p_{1,198}$	$p_{1,199}$	$p_{1,200}$	$M_{avg,1}$	$T_{avg,1}$	$I_{max,1}$	N_1
Peak 2	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$	$p_{2,4}$	$p_{2,5}$	$p_{2,6}$	...	$p_{2,195}$	$p_{2,196}$	$p_{2,197}$	$p_{2,198}$	$p_{2,199}$	$p_{2,200}$	$M_{avg,2}$	$T_{avg,2}$	$I_{max,2}$	N_2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Peak 4130	$p_{4130,1}$	$p_{4130,2}$	$p_{4130,3}$	$p_{4130,4}$	$p_{4130,5}$	$p_{4130,6}$	...	$p_{4130,195}$	$p_{4130,196}$	$p_{4130,197}$	$p_{4130,198}$	$p_{4130,199}$	$p_{4130,200}$	$M_{avg,4130}$	$T_{avg,4130}$	$I_{max,4130}$	N_{4130}
Peak 4131	$p_{4131,1}$	$p_{4131,2}$	$p_{4131,3}$	$p_{4131,4}$	$p_{4131,5}$	$p_{4131,6}$	...	$p_{4131,195}$	$p_{4131,196}$	$p_{4131,197}$	$p_{4131,198}$	$p_{4131,199}$	$p_{4131,200}$	$M_{avg,4131}$	$T_{avg,4131}$	$I_{max,4131}$	N_{4131}

Figure 4: Schematic Overview of the Peak Analysis Subsystem

(A) Sample peak cluster extraction. (B) Sample peak cluster partition. (C) Preliminary peak matching for registration. (D) Matched Peaks. $p_{i,j} = <M_{i,j}, T_{i,j}, I_{i,j}>$ – apex point of extract peak with mass ($M_{i,j}$), register time ($T_{i,j}$) and intensity ($I_{i,j}$). The behavior of each row is summarized by the average mass ($M_{avg,i}$), average registered time ($T_{avg,i}$), average intensity ($I_{max,i}$) and a count of the number of peaks detected (N_i) (from Grigsby et.al. 2010).

4.6 Peak Registration

Peak registration uses only mass and retention time coordinates (M_a,T_a) of the peak apices to achieve the significant data reduction required to work efficiently across many LC-MS samples. Peak registration primarily involves temporal alignment of peaks, although for some instruments the alignment of mass measurements is also required. Initially, one of the images is selected as the reference image and all others are transformed to match it. As illustrated in Figure 4C, this transformation is accomplished by bracketing each peak with a maximum shift window (Δ_m, Δ_t) and identifying matching pairs of peaks. Ideally, these are one-to-one unique matches, meaning each peak has only one unique candidate match. A larger set of candidate matched peaks can be defined by relaxing the criteria so that only the peak in either the reference image or alignment images has a unique matching peak. The set of matched peaks is then used to compute a polynomial transformation that maps retention times of images relative to the reference image. The order of the polynomial is determined by the user (Eq. 1).

$$T_r = \mathrm{polyval}(T, \mathrm{polyCoef}) = c_2 \cdot T^2 + c_l \cdot T + c_o \ldots \qquad \text{Eq. 1}$$

Should the need arise to make adjustments to mass coordinates (Eq. 2); the set would be used to compute a bivariate alignment polynomial.

$$T_r = \mathrm{polyval}(T,\ M,\ \mathrm{polyCoef}) \qquad \text{Eq. 2a}$$

$$M_r = \mathrm{polyval}(T,\ M,\ \mathrm{polyCoef}) \qquad \text{Eq. 2b}$$

4.7 Peak Matching

The set of matching pairs of peaks is used to initialize a matrix of matched peaks (Figure 4D). Each column represents one image and each row contains a set of registered peaks. The peak coordinates (M_a, T_a) are averaged over non-empty images in each row (M_{avg},T_{avg}), producing a synthetic reference image so that the original reference image is no longer required. A number of cycles of the matching algorithm are then used to fill in existing rows and to extend the number of rows by seeding the reference image with peaks from the pool of unused peaks. The coordinates of the seed peaks are used as initial values for (M_{avg},T_{avg}). Each iteration of the matching algorithm produces new alignment polynomials by pairing image peak coordinates with the evolving row averages. A peak matches the row average if its coordinates fall within a (Δ_m, Δ_t) box centered on the row average, where Δ_t is much smaller than the Δ_m used for pre-alignment matching pairs. The final matching step, which attempts to fill in any empty slots in the matching matrix, is accomplished by selecting the raw data point with the maximum intensity in the (Δ_m, Δ_t) acceptance box as a peak substitute.

4.8 Visualization System

The visualization system is based on the Model-View-Controller (MVC) software architecture pattern [107]. The model is composed of a series of relational data tables that include the registered match peak table (Figure 4D), the experimental descriptor table, cluster-peak data tables and raw sample data tables. The user interacts with the model via a graphical interface that supports mouse and keyboard input. The communication between the controller and the model is implemented using callback mechanisms defined in the Matlab programming language. When the user triggers a callback event, the controller notifies the model of the user's action and then possibly alters the state of the model. The view may automatically be invoked by the controller to update some subset of displays as a result of a change in the state of the model or the view may query the model to generate a display based on a user request.

The user interaction with the model is organized into a collection of filters that allow the user structured access to the various components of the data model (Figure 5). These filters are divided into logical categories: data, statistical, chemical, and experimental. The data filters allow access to subsets of data that are restricted by mass, retention time or intensity. The statistical filters allow the user to locate statistically significant patterns of behavior across the entire set of registered peaks. Chemical filters allow the user to remove certain peaks from the analysis based on chemical properties. For example, isotopic peaks or adducts can be automatically filtered to simplify analysis. Finally, experiment level filters allow the user to select items related to the biological experiment such as treatments levels or longitudinal studies for analysis.

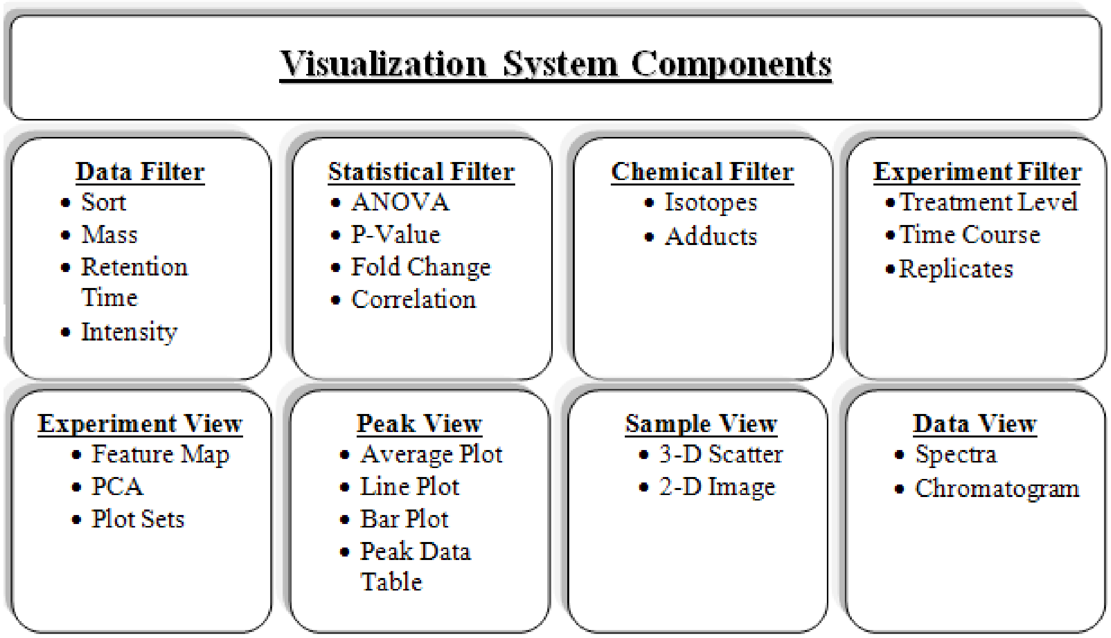
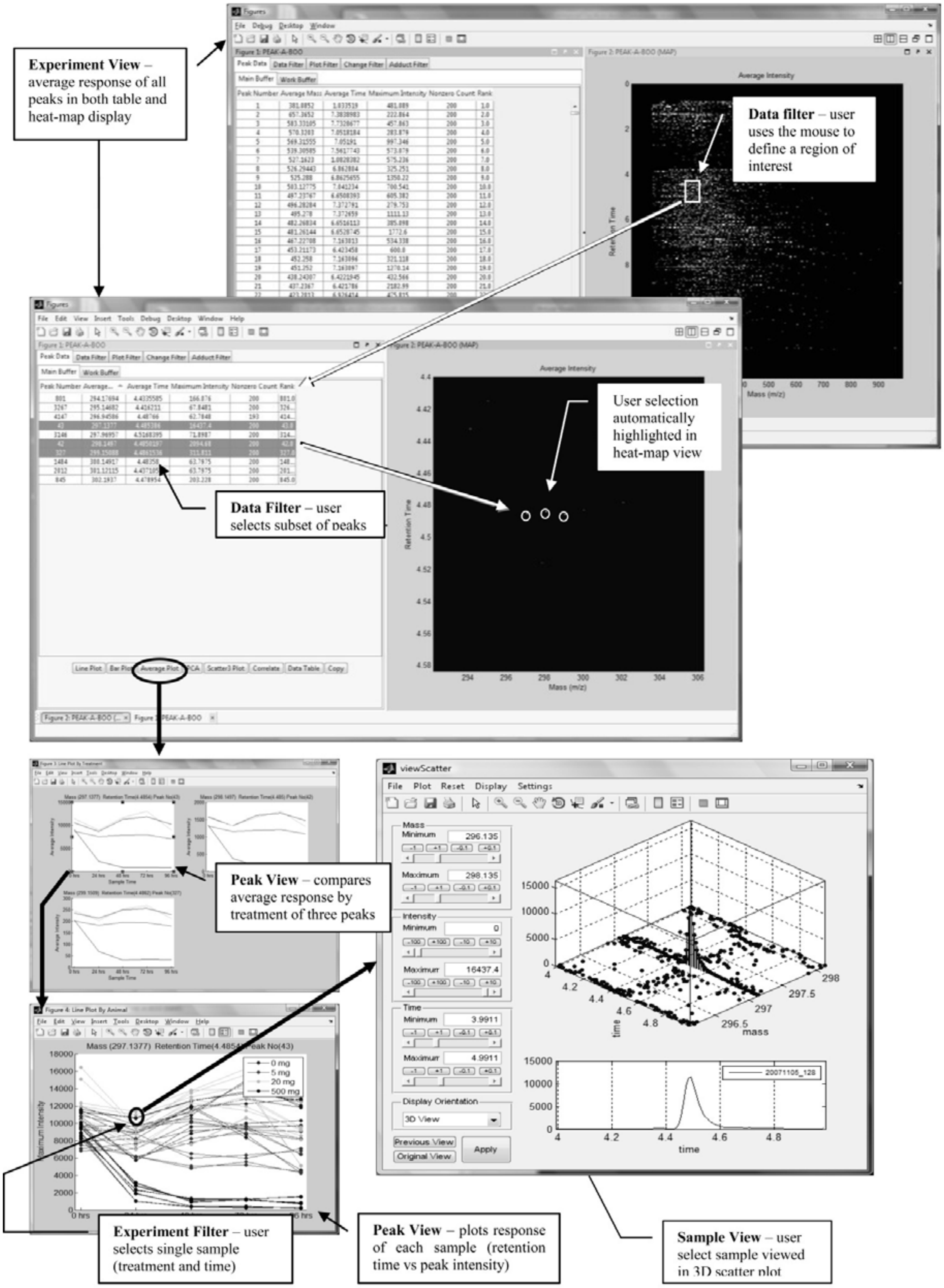


Figure 5: Visualization System Overview
(from Grigsby et.al. 2010)

The filtered data is visualized through a variety of displays. The displays allow a multifaceted view of the data. Figure 6 demonstrates one series of filter-display interactions possible using the visualization system. In this example, the main display opens with a view of all registered peaks stored in a summary table along with a heat-map. Each point in the heat-map represents the location of a registered set of matched peaks. The position of the point in the heat-map denotes the peak-set's average mass and average retention time. These correspond to the values of (M_{avg}, T_{avg}) shown in Figure 4D. The brightness of the point is determined by the value of the most intense peak in the registered set (I_{max} in Figure 4D). As shown in Figure 6, the user can apply a data filter to identify a smaller set of registered peaks for analysis and then alter the view to show line plots summarizing the behavior of several registered peaks. The user can select a single registered peak, plot the behavior of all samples as a function of the experimental parameters (treatment vs time) and select a specific sample for further analysis. Additionally, the user can explore the raw data in a 3D scatter plot with the ability to zoom-in and zoom-out in any specified spectral region.

To illustrate the utility and power of MeDDL in the visualization of large, multi-group experiments, we analyzed data from an LC-MS effort for profiling low level kidney biomarkers in the F344 rat model. Importation of the raw QToF MS data files and subsequent analysis of the aligned and registered peak database by MeDDL identified numerous metabolic changes in the urine of the animals after D-Serine treatment compared to control animals. Registration and processing of the D-serine exposure data (n=208) utilizing peak inclusion criteria requiring each m/z at a given retention time be present in a minimum of 5% of all samples was accomplished in 122 minutes. A smaller sample set of n=20 was similarly registered and processed in 12 minutes to establish scalability, with all analysis performed utilizing a dual core 2.53GHz Central Processing Unit (CPU) with 6 GB of Random Access Memory (RAM). This alignment and registration encompassed all detectable peaks, with absolute intensity values as low as 30 being registered (background level previously established by our group for the QToF Micro).



Peak Number	Average Mass	Average Time	Maximum Intensity	Nonzero Count	Rank
1	381.0852	1.033519	481.089	200	1.0
2	657.3652	7.3838983	222.864	200	2.0
3	583.33105	7.7328677	457.863	200	3.0
4	570.3283	7.0518184	283.879	200	4.0
5	569.31555	7.05191	997.346	200	5.0
6	539.30585	7.5617743	573.879	200	6.0
7	527.1623	1.0828282	575.236	200	7.0
8	526.29443	6.862884	325.251	200	8.0
9	525.288	6.8625655	1350.22	200	9.0
10	503.12775	7.841234	700.541	200	10.0
11	497.23767	6.6508393	605.382	200	11.0
12	496.28284	7.372791	279.753	200	12.0
13	495.278	7.372659	1111.13	200	13.0
14	482.26834	6.6516113	385.898	200	14.0
15	481.26144	6.6528745	1772.6	200	15.0
16	467.22708	7.163813	534.338	200	16.0
17	453.21173	6.423458	600.0	200	17.0
18	452.258	7.163096	321.118	200	18.0
19	451.252	7.163097	1270.14	200	19.0
20	438.24307	6.4221945	432.566	200	20.0
21	437.2367	6.421786	2182.99	200	21.0
22	423.2813	6.926414	425.815	200	22...

Peak Number	Average...	Average Time	Maximum Intensity	Nonzero Count	Rank
801	294.17694	4.4335585	166.876	200	801.0
3267	295.14682	4.416211	67.8481	200	326...
4147	296.94586	4.48766	62.7848	193	414...
43	297.1377	4.485386	16437.4	200	43.0
3146	297.96957	4.5168395	71.8987	200	314...
42	298.1497	4.4850187	2094.68	200	42.0
327	299.15088	4.4861536	311.811	200	327.0
1484	300.14917	4.48358	63.7975	200	148...
2812	301.12115	4.437105	63.7975	200	281...
845	302.1937	4.478954	203.228	200	845.0

Mass
Minimum 296.135
Maximum 298.135

Intensity
Minimum 0
Maximum 16437.4

Time
Minimum 3.9911
Maximum 4.9911

Display Orientation
3D View

Experiment View – average response of all peaks in both table and heat-map display

Data filter – user uses the mouse to define a region of interest

User selection automatically highlighted in heat-map view

Data Filter – user selects subset of peaks

Peak View – compares average response by treatment of three peaks

Experiment Filter – user selects single sample (treatment and time)

Peak View – plots response of each sample (retention time vs peak intensity)

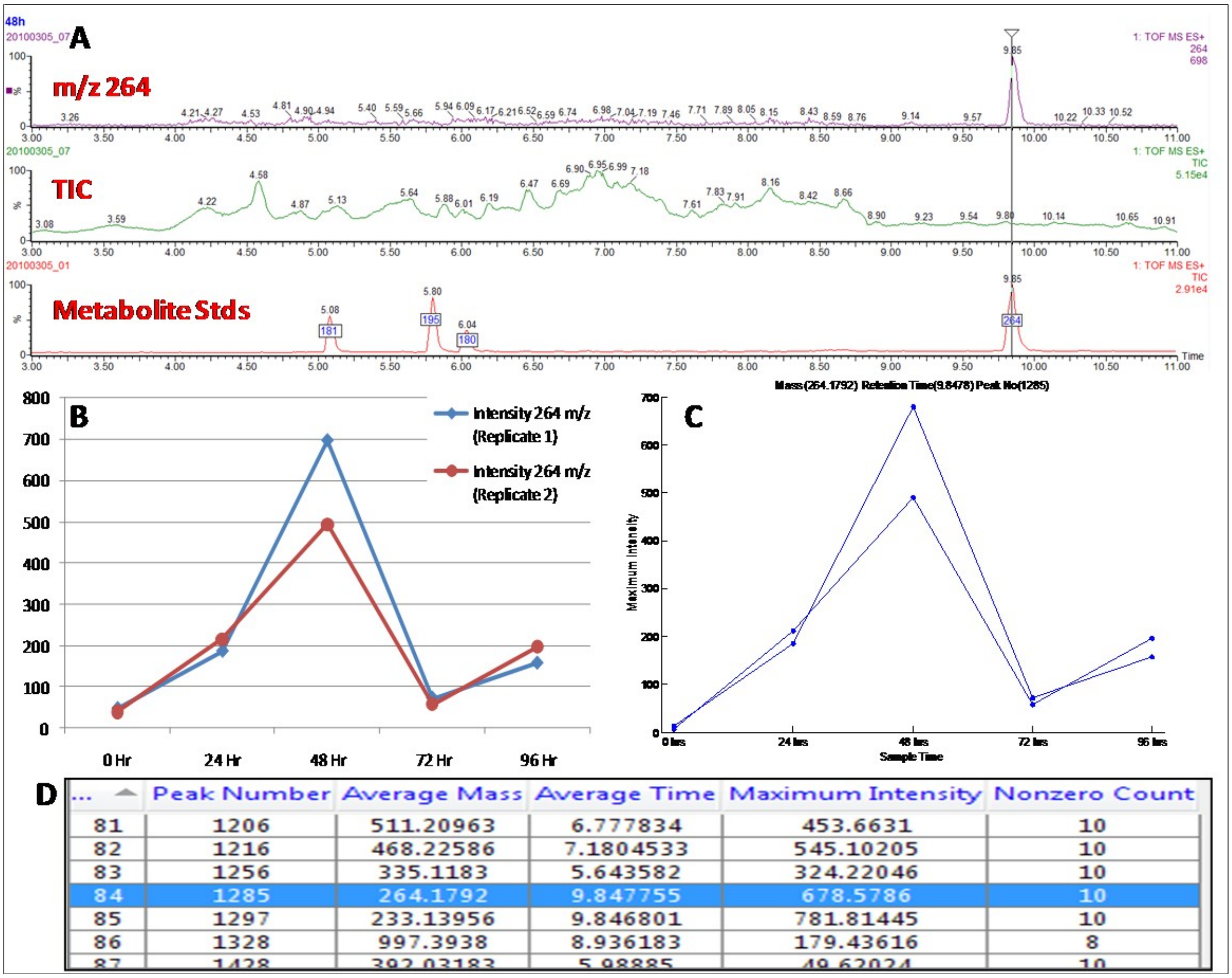
Sample View – user select sample viewed in 3D scatter plot

Figure 6: Example MVC Interaction

(from Grigsby et.al. 2010)

4.9 Results and Discussion

To verify the accuracy of the registration and analysis algorithms, a spiked study of control F344 rat urine was performed (Figure 7A) using a purchased metabolite test mixture of five known compounds (Waters Corporation Metabolomic Test Mix). An artificial dose response was generated as shown below (Table 4) and examination of nortryptyline (m/z 264.1752), the highest intensity standard in this set, via Masslynx (Waters Corporation) generated the response illustrated in Figure 7B (below). Following processing by MeDDL, nortryptyline was registered by the software generating an identical response curve to that manually determined in the vendor supplied instrument control software (Figure 7C). Analysis of the spiked data utilizing the previously described fold change filter for all masses showing a 5-fold change across time for a given dose showed inclusion of nortryptyline (Figure 7D).



...	Peak Number	Average Mass	Average Time	Maximum Intensity	Nonzero Count
81	1206	511.20963	6.777834	453.6631	10
82	1216	468.22586	7.1804533	545.10205	10
83	1256	335.1183	5.643582	324.22046	10
84	1285	264.1792	9.847755	678.5786	10
85	1297	233.13956	9.846801	781.81445	10
86	1328	997.3938	8.936183	179.43616	8
87	1428	392.03183	5.98885	49.62024	10

Figure 7: Dose Response Spectra

A – Spectra from spiked study of control F344 rat urine showing presence of nortryptyline in spiked urine, TIC of spiked urine sample, and TIC of neat test mixture utilized in spike. B – Nortryptyline dose response obtained via Masslynx (Waters Corp.). C - Nortryptyline dose response obtained via MeDDL following alignment and registration. D – Selection of nortryptyline via MeDDL as showing >5-fold change in time versus treatment (from Grigsby et.al. 2010).

Table 4: List of Spiked Standards

	Theophylline	Caffeine	Hippuric Acid	4-Nitrobenzoic acid	Nortryptyline
0 Hr	*0 pg*	*0 pg*	*0 pg*	*0 pg*	*0 pg*
24 Hr	*750 pg*	*750 pg*	*750 pg*	*375 pg*	*281 pg*
48 Hr	*3.75 ng*	*3.75 ng*	*3.75 ng*	*1.88 ng*	*1.41 ng*
72 Hr	*375 pg*	*375 pg*	*375 pg*	*188 pg*	*141 pg*
96 Hr	*750 pg*	*750 pg*	*750 pg*	*375 pg*	*281 pg*

(Waters Corporation Metabolite Test Mix) in F344 control urine utilized in software validation study generating a synthetic dose response (from Grigsby et.al. 2010).

Accuracy in both the ability of the software to perform correlations as well as in peak registration can be demonstrated through correlation of adducts and isotopes in the aligned peak database, which also allows for their easy visualization and elimination as candidate biomarkers.

As described, twenty separate groups (4 doses x 5 time points) totaling 208 samples were analyzed. Following alignment and registration of the D-serine exposure data, more than 4000 isotopic peaks were originally registered and matched prior to automated de-isotoping via MeDDL. During this process the isotopes were identified after peak matching was complete. The location (M_{avg},T_{avg}) of each synthetic peak was used to initiate a search for mono-isotopic peaks. For a given peak, a search was conducted to located mono-isotopic peaks by looking for a peaks at location

(M_{avg} + 1,T_{avg}) (M_{avg} + 2,T_{avg}) and (M_{avg} + 3,T_{avg}). A match was found if a peak was located within the region defined by ($M_{avg} + 1 \pm M_{\varepsilon}, T_{avg} \pm T_{\varepsilon}$) where M_{ε} and T_{ε} is a user specified limit on mass and retention time variation between isotopic peaks. Once a potential isotope is identified, the intensity of the actual extracted peaks (main peak and isotopic peak) in each image is compared to verify that the isotopic peak has a decreasing level of intensity. If all peaks in the set and their corresponding isotopic peaks satisfy this requirement, the isotopic peaks are tagged and can be hidden / removed by the user. A similar process is used to locate doubly and triply charge isotopic peaks and tag them for removal.

One of the novel aspects of the MeDDL peak alignment process is the use of a two-stage process that begins with a rough peak match where only a few isolated peaks are identified between a reference image and each unregistered image. These initial peaks are used to compute a polynomial transformation between the reference image and the unregistered image producing a rough alignment. This is essentially a global process that handles systematic misalignment between images. In the peak matching phase, alignments are refined through a process similar to relaxation labeling [108, 109]. After the rough alignment, a synthetic image is created by taking each image in turn and using every peak in the image as the center of a peak acceptance region. Any peak in any other image captured within the acceptance region is matched to this peak. The average mass and retention time (M_{avg},T_{avg}) across the set of peaks is computed as (M_{avg},T_{avg}). Initially, the acceptance region is very small so peaks that were not well-aligned by the polynomial function will not be matched. The relaxation process slowly opens the size of the acceptance region to attempt to draw in one peak from each image. Each time a new peak is captured within the acceptance region, the average (M_{avg},T_{avg}) is recalculated. Thus, the acceptance region gradually shifts and coalesces to maximize the number of matched peaks across all images at the final value of (M_{avg},T_{avg}). This combination of global alignment / local refinement allows the matching to respond to both systematic misalignments as well outliers that appear as random variations within individual images. Additionally, although MeDDL accurately aligned all data generated by our laboratory, the retention time variation observed with the Waters Corp. Acquity UPLC was minimal (<0.25 min). As such, alignment was achieved through the use of a 2nd order polynomial and our two stage peak alignment process. During development, our group evaluated use of higher order functions; however, we deemed it unnecessary for our use. This can be easily modified to be a user editable feature through the software interface if necessary for other chromatography systems.

During the analysis of the exposure data, two of the primary tools included in the MeDDL platform were utilized by our group, principal component analysis [110] (PCA) and a novel fold change filter. The design of the fold change filter analysis is based on a multilevel statistical model that views the behavioral response (intensity) of each synthetic peak as a normally distributed random with the added assumption that the behavior of peaks within individual images is correlated. Based on this underlying statistical model the system is designed to handle longitudinal data sets consisting of subjects exposed to multiple levels of treatments. The statistical models are designed to allow the user to perform statistical tests for significant differences between treatment levels, significant differences between treatment time points, or significant differences between any combination of treatment levels or time points. In future applications, if other analysis tools become required, MeDDL is easily expandable through its use of the MVC software architecture previously described. This software architecture allows a programmer to extend the functionality of the system as follows (1) add a new choice to any pull-down menu in system menu, (2) install a new callback for the menu item that invokes a user defined function, and (3) create a new user function (userFunction.m). The user code added to userFunction.m has full access to all Matlab libraries (e.g. image processing, signal processing, pattern recognition, statistics, etc.) and full access to the summary data describing the matched peaks, the full description of every peak in a matched set and the raw data for every image. This allows a programmer / user to add new functionality to the system without altering the existing functionality.

PCA was performed for all groups of study animals and is shown in Figure 8. The PCA plot demonstrates clear separation between sample dosage and time groups with the majority of metabolomic changes in urine observed at 24, 48 and 72 hours post treatment with 500 mg/kg D-Serine. The number of peaks that undergo at least a two-fold change is 19 times higher for the 500 mg/kg dose than the 5 mg/kg dose, with the changes literally disappearing at 96 hours, most likely indicating kidney recovery.

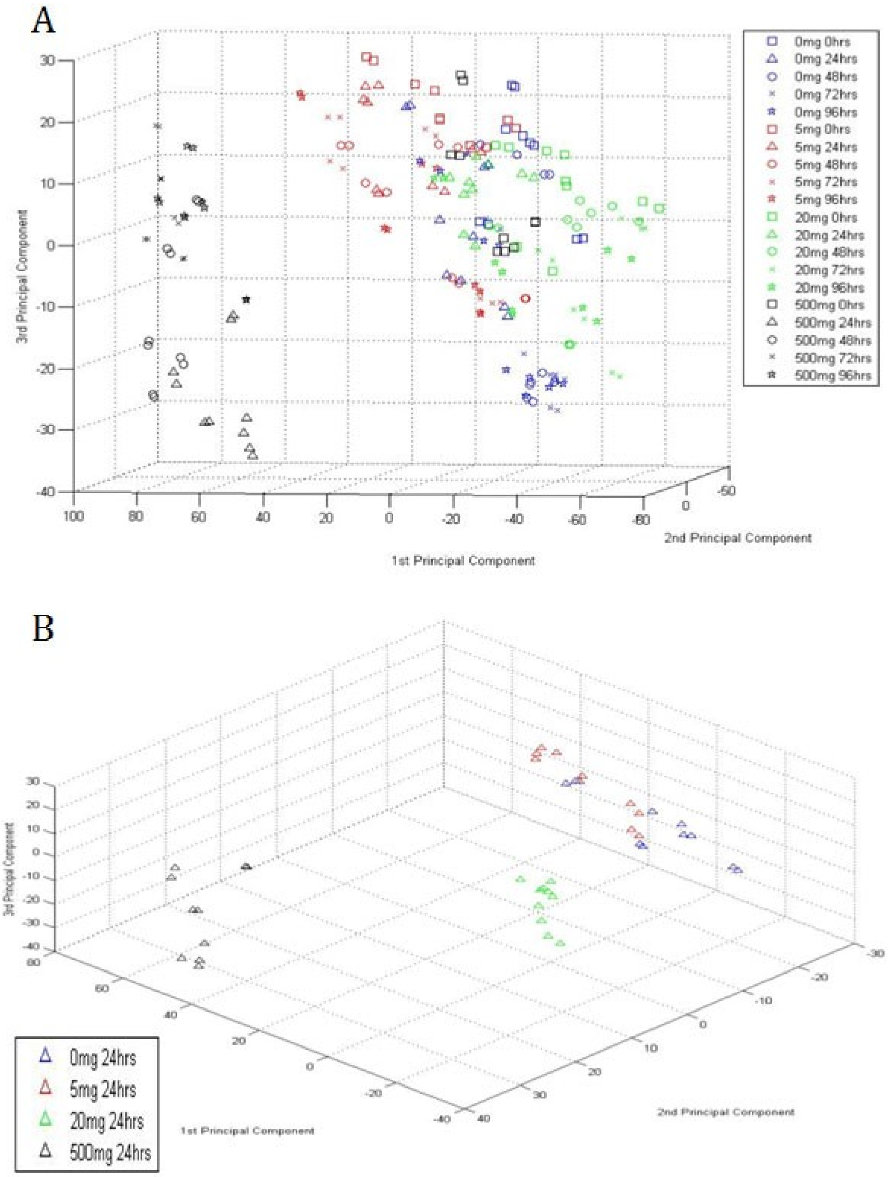


Figure 8: Dose Response PCA

A - PCA of LC/MS data for all experimental animal groups of the study. Legend is on the right of the figure. B - Principal component analysis of LC/MS data for 0, 5, 20 and 500 mg/kg doses at 24 hours only (from Grigsby et.al. 2010).

At 24 hours post-dosing for the 500 mg/kg group, as many as 426 peaks show a greater than two-fold change with the peak intensity cut-off set to a minimum of 100. Although the two-fold increase was established based on our goals of identifying differential yet detectable metabolite biomarker profiles, the fold change filter incorporates detailed statistical analysis including ANOVA (1-way, 2-way and N-way) among the selected subject groups. The user can optionally perform multiple pairwise comparison tests among the means of groups to determine whether or not all differences among group means satisfy a user defined level of significance. A Bonferroni correction is applied to compensate for the tendency to incorrectly find a single pairwise significant difference among multiple comparisons. Further, five metabolite peaks exceeded 100-fold change with the same intensity threshold. It is worthy to note that a number of peaks exhibit a statistically significant change while their intensities are relatively low, with most of these peaks demonstrating negative changes in our analysis. Examples of negative and positive changes are shown in Figure 9. We have excluded isotopic peaks in our data analysis; however, some percentage of differentiated peaks can be attributed to adduct acquisition by metabolites as well as water loss. Thus, the difference of 18 mass units between peaks 1569 and 1598; 952 and 246; 1642 and 1532; 1697 and 1664; and 3277 and 42 strongly suggest a water loss with each set of ions eluting from the column concurrently.

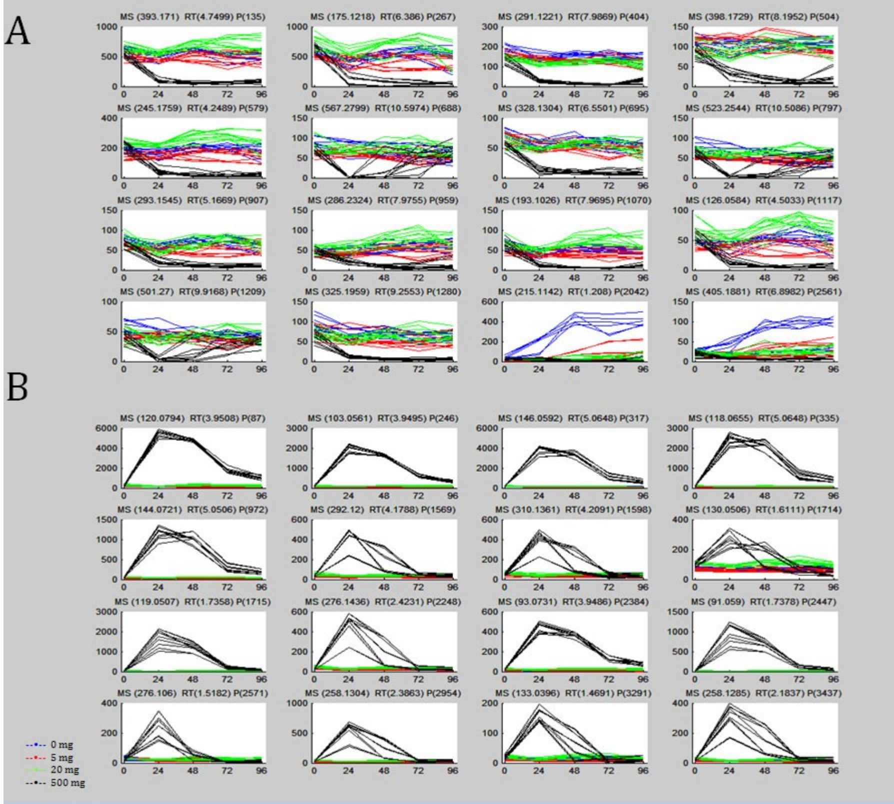


Figure 9: Examples of Selected Peak Plots of Negative

(A) and positive (B) changes after 500 mg/kg D-Serine exposure from Grigsby et.al. 2010)

Preliminary metabolite identification was performed, with a list of potential metabolites shown in Table 5. Purchased metabolite standards were run under the same LC conditions followed by MS/MS as selected samples. Matching retention times and MS/MS fragmentation data generally indicate the conclusive identification of a metabolite, which has been demonstrated in a candidate biomarker identified in this study, 3-indolylacetic acid. The MS/MS spectrum of 3-indolylacetic acid along with a spectrum from rat urine samples is shown in Figure 10. This figure also demonstrates the ability of the MeDDL to automate spectral normalization. In MeDDL, normalization begins by computing the mean (m_1, m_2, m_3,…) and standard deviation (s_1, s_2,s_3, …) of the values of all subjects in each treatment group at the first time point. The peak intensity value $p_s(t)$ for each subject s in group j at time t is then normalized as: $(p_i(t) - m_j) / s_j$. This effectively shifts all the plots so the mean value of the first time point in each treatment group is zero (see Figure 10C).

Table 5: List of Potential Metabolites

m/z	Retention Time	Treatment x Time p-value
521.2412	9.071536	3.31E-05
523.2543	10.508634	8.04E-05
491.2421	7.5071826	1.44E-04
501.2701	9.916751	1.74E-04
714.1853	1.6265737	2.22E-04
611.3051	10.67593	2.60E-04
609.2873	9.332852	5.72E-04
567.2799	10.59735	7.30E-04
613.3242	11.733243	9.17E-04
383.1925	4.111277	0.00169446
779.352	1.4149536	0.00176583
290.1256	1.469668	0.00190476
655.3292	10.74838	0.00203854
435.159	10.100951	0.00356158
589.3217	10.138795	0.00407047
479.2284	10.405753	0.0055822
701.3726	11.777291	0.00567922
326.1946	9.491132	0.00787954
553.2582	4.0478706	0.01769215
633.3461	10.232473	0.01919617
212.1025	1.1299926	0.02197477
330.0621	2.3208911	0.02478929
533.1005	4.628623	0.04081653
511.2622	4.253147	0.04097449
290.1258	1.7111521	0.04116634

List of potential metabolites (shown as m/z) identified in urine of rats after 500 mg/kg D-Serine exposure at 24 and 48 hours after the exposure. Fold change filter set at ten-fold and higher (from Grigsby et.al. 2010).

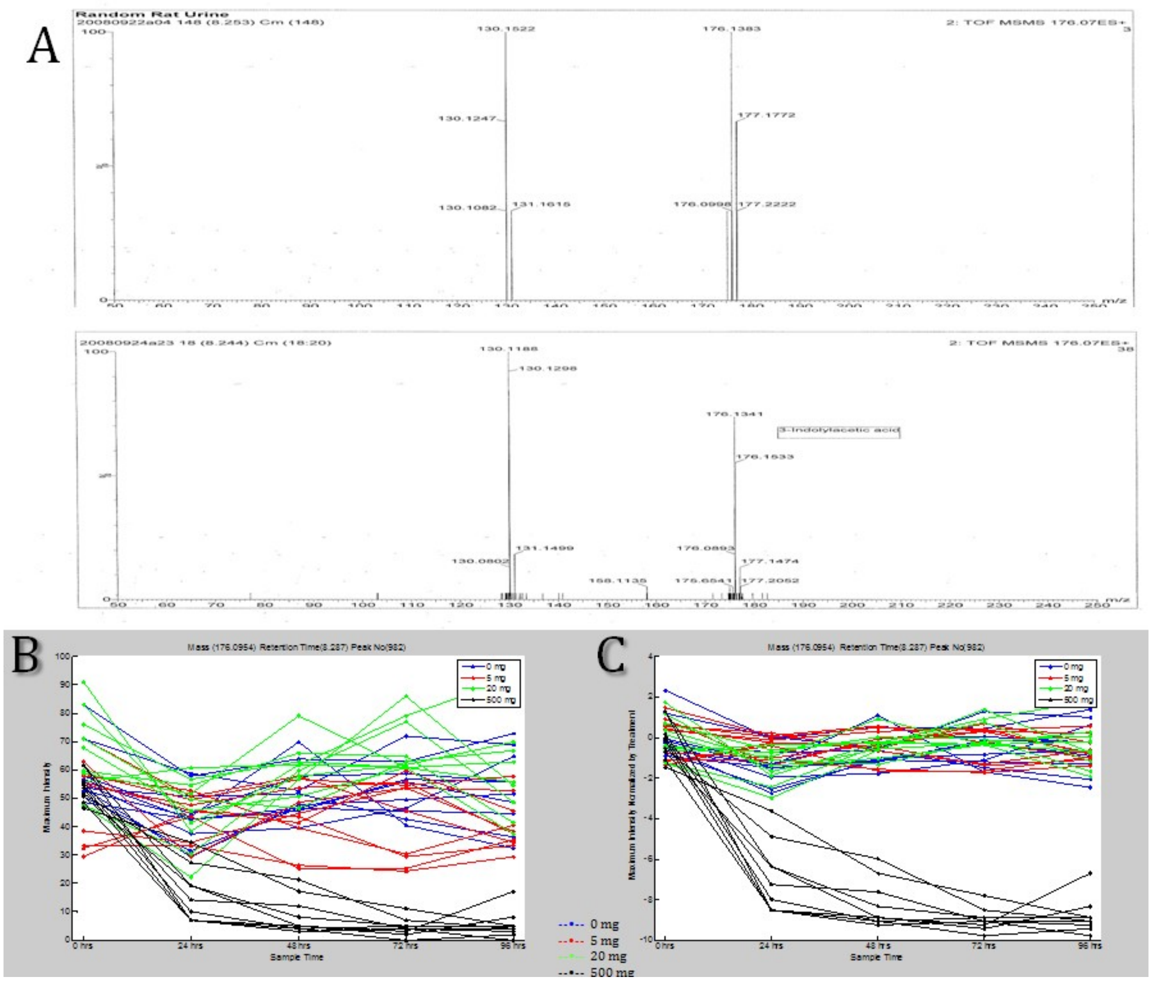


Figure 10: Identification of a Selected Metabolite

Retention times and MS/MS fragmentation of 3-Indolylacetic acid is shown for standards along with corresponding ms/ms from D-Serine urine samples in A. Not normalized (B) and normalized (C) plots show changes for 3-Indolylacetic acid throughout the course of D-Serine study (from Grigsby et.al. 2010).

4.10 Conclusions

The data clearly demonstrate dramatic changes in the urinary metabolic profile in response to the kidney toxicant, D-Serine. A list of potential metabolites corresponding to masses identified in urine of rats is presented. D-Serine metabolomic profiling demonstrates that most changes occur between 24-72 hours. The most dramatic changes occur at the 24-hour time point after exposure to 500 mg/kg D-Serine. The data suggests that near-normal kidney function resumes at 96 hours.

Although the ability to visualize the experiment at all levels may constitute the authors' ideal for biomarker discovery and differential metabolite analysis, we feel it adds considerably to this effort by allowing the user to differentiate metabolite profiles in a large time-dose study while maintaining the ability to focus on individual metabolites and spectra for subsequent identification. Additionally, although the tool performed quite well for LC/MS dose response study analysis, some early limitations were identified in the prototype version of the tool, principally the ability analyze data obtained via GC/MS. GC/MS data presents a specific set of challenges, primarily due to differences in the ionization technique utilized. To address these challenges as well as implementing several marked improvements to the tool, significant changes to the prototype software were completed, as described below.

5.0 GC/MS FUNCTIONALITY AND EXPANDED TOOLSET

5.1 Overview

As stated in Section 1, a growing body of discoveries in molecular signatures has revealed that VOCs, the small molecules associated with an individual's odor and breath, can be monitored to reveal the identity and presence of a unique individual, as well their overall physiological status. Given specific analysis requirements for differential molecular profiling via gas chromate-graphy/mass spectrometry, our group has expanded on the prototype MeDDL tool to allow for processing of VOC data [89]. Preliminary analysis via the MeDDL toolset generally identifies a moderately large number of differential, registered peaks, which, depending on filtering and comparison parameters could be in the range 50-100 peaks separating the conditions of interest. This initial, down-selected subset of peaks is typically too large for incorporation into a portable, electronic nose based system in addition to including VOCs that are not amenable to classification; consequently, it is also important to identify an optimal subset of these peaks to increase classification accuracy and to decrease the cost of the final system. In this chapter, we will discuss an approach to how this differential peak subset and their corresponding intensities are used as features for classification.

As the first illustration of our approach to these studies, we present the below urine based VOC comparison of the two parental strains of the BXD mouse model [111], C57 and DBA. For this comparison, we demonstrate the expanded MeDDL functionality, to include machine learning tools via a classifier similar to a K-Nearest Neighbor (KNN). This modified KNN classifier is used in conjunction with a genetic algorithm (GA) that simultaneously optimizes the classifier and subset of features. The GA utilizes Receiver Operating Characteristic (ROC) curves to produce classifiers having maximal area under their ROC curve. Using this approach, experimental results shown below on over a dozen recognition problems show many examples of classifiers and feature sets that produce perfect ROC curves.

5.2 Data Filtering

Following the spectral registration and alignment previously described [100], the data was analyzed using several of the principal analytical methodologies included in MeDDL: unsupervised clustering via principal component analysis [110]; differential down selection of peaks through combination of a set of logical filters; and utilization of machine learning based tools for significant VOC “feature” identification.

MeDDL was originally created for the analysis of LC/MS data. The ionization techniques generally employed for LC/MS are termed “soft” and impart low energy to eluting ions, resulting in fairly simple mass spectra: often comprised of just the ionized analyte, or “parent” ion. Modifications to the original implementation of MeDDL were required to aid in the analysis of the more complex mass spectra in GC/MS resulting from the “hard ionization” induced by the Electron Impact (EI) fragmentation process in the mass spectrometer’s ion source. A reductionist approach for this analysis was required for the efficient determination of changes observed between sample groups. To address this issue, we created a supplementary time-binning filter allowing the analyst to specify both a time window and lower bound threshold of peak intensities. The comparison then proceeds as follows: an averaged, composite image of each user-defined comparative group is generated (i.e. the surface obtained from samples comprising each comparative group); the most intense peak from all groups is evaluated across all aligned images using a 0.1 minute window and 100,000 absolute (total ion count) threshold;

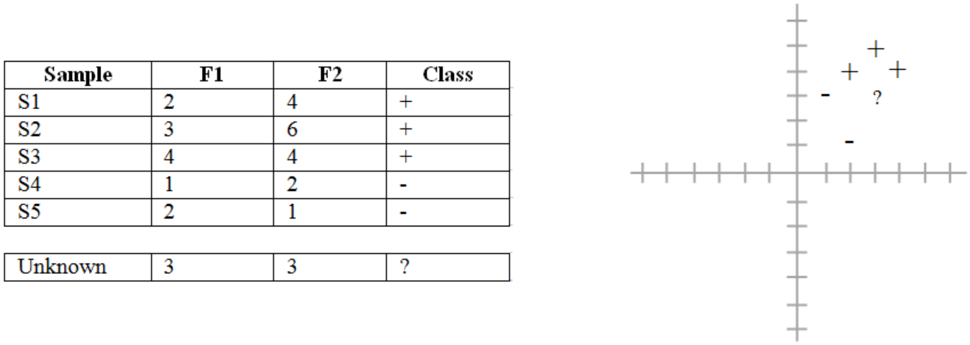
once the comparison is completed, this “time slice” based upon the peak apex ± ½ of the specified time window is removed from further analysis and the next most intense set of peaks are compared. An additional filter applied in the differential analysis of groups in this study included a fold change filter limiting results to only those peaks which demonstrated at least 2 fold or greater change in intensity between strains. It must be noted that although the MeDDL tool contains a wide variety of implemented statistical filters for feature down-selection, we limited their use to only the 2 filters listed to allow for optimal feature selection by the classifier. Once both of these filters were applied to the grouped, global data set, a Boolean "AND" was added to the resulting filtered peak sets to identify the logical intersection, an approach similar to that used in generation of a Venn diagram. These reduced data sets were then used for further classification described below.

5.3 Classification

The filtered, numerical data sets, or feature vectors, produced by the preprocessing described in the previous section must be used to perform classification on unknown samples for optimal results. However, performing classification with these features still presents several problems. First, the filtered features include noisy, irrelevant features, despite the preprocessing steps taken to identify features that have both intra-class similarity and inter-class dissimilarity. Second, the set of filtered features include those that are highly correlated and therefore are redundant. These two observations suggest the classification system should produce a classifier, but should also down select the incoming feature set to a small set of cooperative features that are amenable to classification.

5.4 Modified KNN Classifier

The basic KNN is a two-class classifier that is often used in situations where the data distributions are generally unknown [112]. KNN training is performed by using all samples of the training data as labeled prototypes. Unknown samples are classified by comparing the distance of the unknown sample to the k nearest prototypes, where k is a small user-defined integer (e.g., 3). In binary classification (i.e., -two class classification), choosing an odd value for k avoids a potential tie vote. The method of computing distance with N-dimensional data is commonly done in two different ways: Euclidean distance and L1 norm, or Manhattan/Minkowski distance formula using $p = 2$. This work uses Euclidean distance, but the L1 norm appeared to provide similar results. The three nearest prototypes then vote on the unknown’s class label. Figure 11 illustrates this process in two dimensions. In this sample, the training data contains 5 samples, which includes 3 positive samples and 2 negative samples. The three closest samples to the unknown are S1, S3, and S4, with the majority those samples being positives; consequently, the unknown would be labeled as positive.



Sample	F1	F2	Class
S1	2	4	+
S2	3	6	+
S3	4	4	+
S4	1	2	-
S5	2	1	-
Unknown	3	3	?

Figure 11: Example Data and Plot of Data

(from Grigsby et.al. 2012)

One unique objective of this work was to develop a classifier that has one or more parameters that control the classifier's behavior. For example, it may be important to correctly classify positives, with an increased tolerance for false alarms. Conversely, it may be deemed acceptable to miss a couple positives, if the increased number of false alarms is kept small. The k parameter in the KNN classifier does not provide such a parameter. k simply denotes the number of voters and does not provide a way to increase/decrease the sensitivity toward the class boundaries. Further, the number of prototypes is typically quite small in biological studies and therefore modulating the number of voters would have limited utility.

For an appropriately configurable classifier, a ROC curve visually illustrates the possible tradeoffs between the rates of true positives and false positives. Figures 12a and 12b illustrate a typical ROC curve and the perfect ROC curve. Figure 12a depicts the tradeoffs of a hypothetical classifier. The figure shows that the classifier has a parameter that can allow it to obtain a 0.75 true positive rate, while simultaneously having a false alarm rate of 0.25. Should the operational situation require 0.9 rate of recognizing true positives, the rate of false alarms would reach a predicted level of approximately 0.75. ROC curves are monotonically increasing. The perfect classifier would obtain a rate of 1.0 for positives with a false alarm rate of 0.0. This perfect ROC curve is shown in Figure 12b.

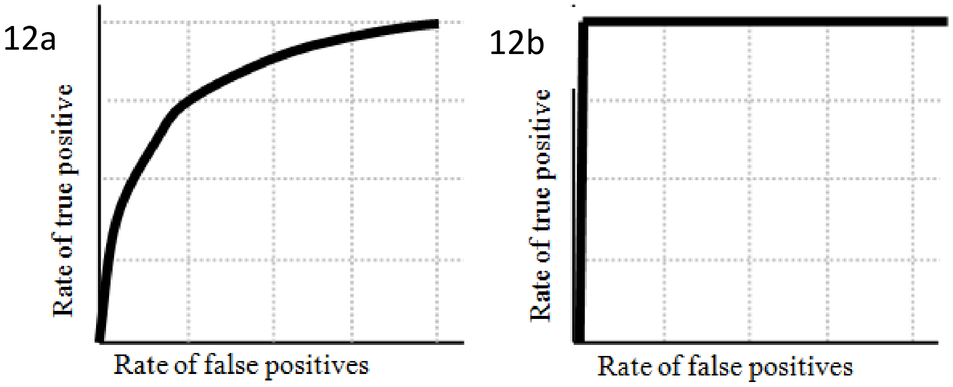
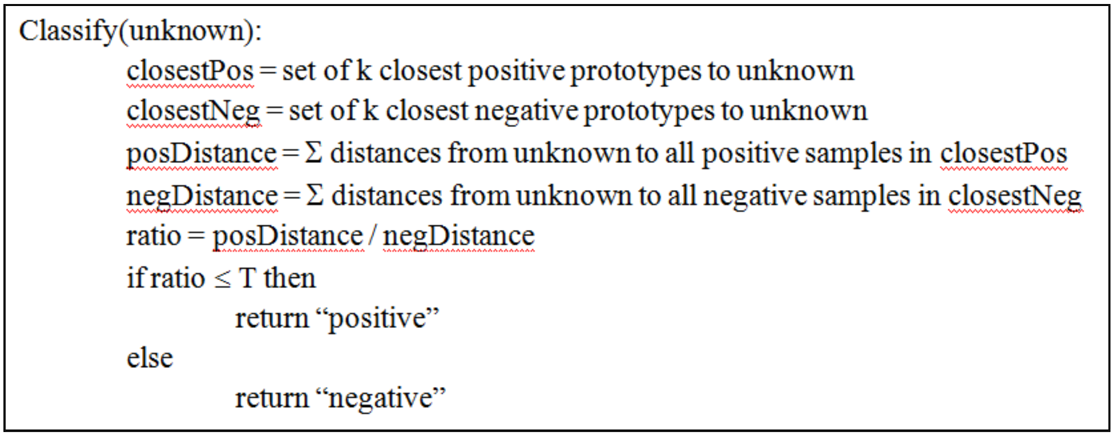


Figure 12: ROC Curves

Figure 2a shows a typical ROC curve. Figure 2b shows the perfect ROC curve (from Grigsby et.al. 2012).

To provide for an adjustable parameter, the KNN's decision rule is modified (Figure13). Whereas, the basic KNN's decision rule is to count the votes to the nearest k prototypes, the modified decision rule uses the distance value to influence its decision. This is approach assumes that being closer to a prototype indicates that it is more likely to be of that category. The definition for the modified KNN decision rule is as follows, where T is the configurable parameter and k is an integer > 0. The classification rule takes the ratio between the total distances to the closest positive prototypes and the closest negative prototypes. If the unknown happens to be a positive, it is expected that posDistance would be small and negDistance would be large, producing a small value for ratio. By adjusting T to a small value, the criteria for declaring "positive" becomes more stringent, in that the unknown's distance from the positives must be quite small while simultaneously its distance from the negatives must be relatively large. Conversely, setting T to a large value allows more samples to be classified as positives. In the extreme case, T = infinity, all unknown samples will be classified as positives.



```
Classify(unknown):
    closestPos = set of k closest positive prototypes to unknown
    closestNeg = set of k closest negative prototypes to unknown
    posDistance = $\Sigma$ distances from unknown to all positive samples in closestPos
    negDistance = $\Sigma$ distances from unknown to all negative samples in closestNeg
    ratio = posDistance / negDistance
    if ratio $\leq$ T then
        return "positive"
    else
        return "negative"
```

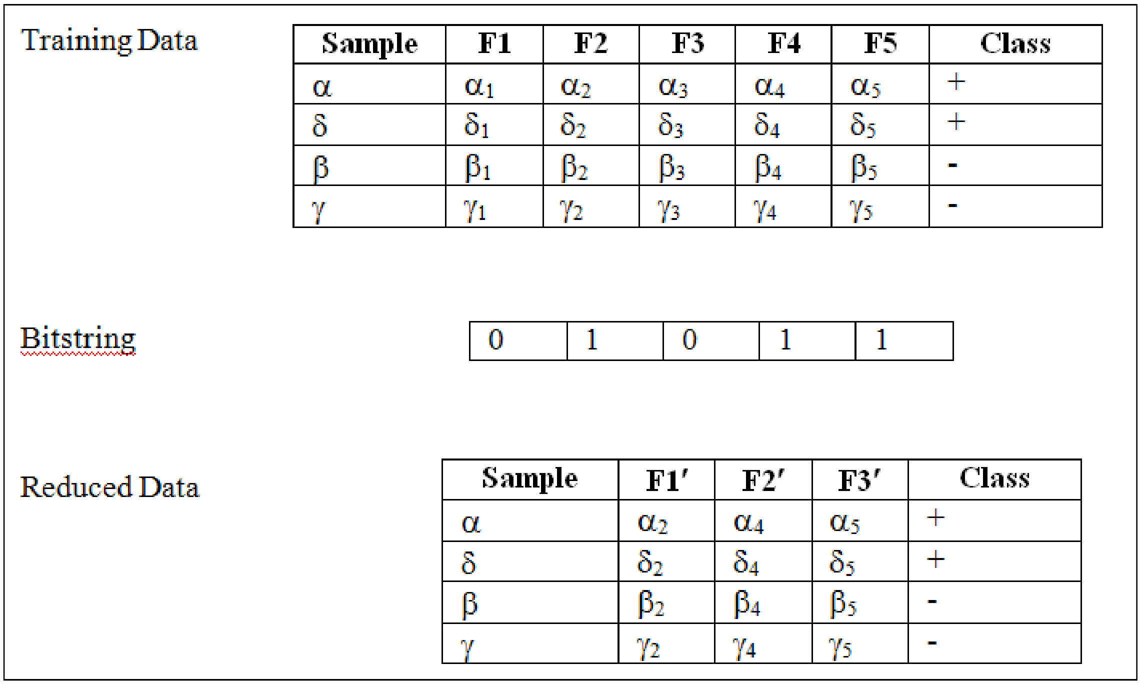
Figure 13: Modified KNN Pseudocode

(from Grigsby et.al. 2012)

5.5 Learning Algorithm for Feature Selection

After preprocessing, the set of filtered features is sent to the classification system. As mentioned above, the potential exists for reducing this set to an even smaller number. Ideally, this reduction would produce a less costly system and produce a subset of features that are more effective than using the entire set as a whole. The ideal subset would contain features with general properties such as: mutual independence, inter-class dissimilarity, and intra-class similarity. Rather than applying more filters to achieve this, our approach is to use the modified KNN classifier to assess the quality of a feature subset; where good subsets will provide good classification and poor subsets will not be very accurate.

The process of selecting a subset from a large set uses a sequence of 0's and 1's to represent the subset. Here the bit positions containing a 1 or 0 indicate features to be included or excluded. Figure 14 shows a diagram illustrating how one bitstring is used to down-select the features and how that down-selection affects the resulting data set that is fed to the KNN learning algorithm. In this example, the bitstring happens to have three on-bits located at positions 2, 4, and 5, indicating that only features 2, 4, and 5 are used and features 1 and 3 are ignored. The down-selected data is then used to form the modified KNN.



Training Data

Sample	F1	F2	F3	F4	F5	Class
α	α_1	α_2	α_3	α_4	α_5	+
δ	δ_1	δ_2	δ_3	δ_4	δ_5	+
β	β_1	β_2	β_3	β_4	β_5	-
γ	γ_1	γ_2	γ_3	γ_4	γ_5	-

Bitstring

0	1	0	1	1

Reduced Data

Sample	F1′	F2′	F3′	Class
α	α_2	α_4	α_5	+
δ	δ_2	δ_4	δ_5	+
β	β_2	β_4	β_5	-
γ	γ_2	γ_4	γ_5	-

Figure 14: Reduction of Training Data

The topmost figure shows the entire set of data. The middle figure shows one bitstring produced by the GA. The bottommost figure shows the training data without the excluded features (from Grigsby et.al. 2012).

A GA is a natural learning algorithm to apply to this problem [113, 114] since it operates on a bitstring. The reader is referred to the text by Goldberg [115] for a more complete treatment of GAs. For our purposes, it suffices to say that the GA is a method for optimizing a sequence of 0's and 1's. In order to achieve this, the GA requires a method for evaluating the quality of the sequence. By assigning a numeric score to a sequence, and many other sequences, the GA navigates the search space to find sequences that are better than the ones it is currently is examining.

Leave-One-Out (LOO) cross validation [112, 116] is a common method for estimating the quality of a classifier using only training data. LOO iterates over all the training samples, where each sample is temporarily removed from the training set. This smaller set is then used to train the classifier, which is then applied to the sample that was held out. Ideally, the classifier will correctly classify the sample. By repeating this process over all training samples, it is possible to assess the generality of the learning technique. If the LOO algorithm shows solid performance over a large percentage of the samples, it can be assumed that the learning technique generalizes to truly unknown samples.

On each iteration of the LOO algorithm, the bitstring in question ultimately results in a KNN that is used to classify the sample temporarily removed. Instead of classifying the sample, the ratio between posDistance and negDistance is recorded. The set of ratios can be used to create a ROC curve that predicts the final system's ROC curve, where the final system refers to the modified KNN that is obtained by using all of the training data. The area under the predicted ROC curve is used as the bitstring's evaluation score. Naturally, a score of 1 corresponds to a perfect ROC curve, which indicates that the feature set forms an effective KNN classifier.

5.6 Materials / Methods for Machine Learning

Animal use in this study was conducted in accordance with the principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 1996, and the Animal Welfare Act of 1966, as amended. BXD mice parental strains (DBA and C57) utilized for this study were singly housed in metabolic cages which are approximately nine cm in diameter and urine and feces were separated and isolated. Individual mouse urine samples were collected using 1 mL disposable transfer pipettes (Thermo Fisher Scientific) and placed in 2 mL Eppendorf Snap-Cap Microcentrifuge Safe-Lock tubes. The urine was then stored frozen at -80°C and thawed on ice prior to analysis. For the BXD VOC baseline set described, 170 individual samples representing the two parental strains (C57 N = 81, DBA N = 89), and six additional test samples (C57 N = 3, DBA N = 3) were processed by aliquoting 200 uL of urine into a 10 ml crimp-top headspace vial (National Scientific). The vials were immediately crimped with Red PTFE/white silicone crimp seals (Fisher). The bench-top GC/MS system utilized for sample analysis was a Thermo Fisher Trace GC Ultra gas chromatograph interfaced to a Thermo Triplus autosampler configured for automated SPME headspace sampling and in-line with a Thermo DSQII single quadrupole mass spectrometer. Collection of organic volatiles from the urine was accomplished using a two cm CAR/DVB/PDMS SPME, Supelco supplier, inserted by the Triplus autosampler into the head-space of the sample vials. The headspace samples were incubated at 60°C for 15 minutes, followed by extraction at 60°C for 30 minutes and automated direct injection. Volatiles gathered by the SPME fiber were analyzed through desorption of the fiber by heating to elevated temperature and separation with a Restek Stabilwax 30m, 0.25mm ID column. Helium was used as the carrier gas at a flow-rate of 1.5 ml/min. A narrow bore SPME injector liner (0.75 mm I.D.) was used (Thermo). The following conditions were utilized for sample analysis: desorption for 2 min via a PTV injector held at 230°C; oven temperature program 50°C (4 min); 5°C/min to 230°C; hold 30 minutes giving a total run time of 70 minutes. The DSQII MS transfer line was held at 230°C and the instrument was operated in positive scan mode from 41 to 400 amu. The raw data was collected in centroid mode and the resulting chromatograms and mass spectra (raw files) were then converted to common data format (CDF) and subsequently analyzed through MeDDL. Due to the fact that SPME extraction is a competitive process leading to mutual displacement from the adsorption sites between different analytes or analytes and matrix constituents, the results of this study as described report data semi-quantitatively based on relative peak heights.

5.7 Machine Learning Results

A total of 170 BXD parental urine samples (DBA and C57 “teaching set”) were collected and analyzed over a four month period with the six unregistered “unknown”, test samples utilized below acquired over 12 months later. Following GC/MS analysis, CDF conversion, and MeDDL registration, the samples in the “teaching set” were filtered for a two-fold change and time binned (0.1 min window, 100K absolute threshold minimum cutoff). The filter results are shown in Table 6, with peakset 1 comprising all registered peaks, peakset 2 comprising time binning, peakset 3 comprising fold change, and peakset 4 the resultant intersection of the two applied filters. This subset of 52 VOC features, or peaks, were first screen by PCA (Figure 15) to demonstrate group separation prior to analysis by the hybrid GA classifier. Principal component analysis is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This technique is often difficult in usage to identify the individual subset of features responsible for group separation, but is quite useful as a screening technique as in Figure 15.

Table 6: BXD Parental Strain Peak Registration and Peakset *(PkSet)* Filter Results

	Subset	Parameter	Size
PkSet 1	All Peaks		2845
PkSet 2	Time Binning	Delta T: 0.1 Min Int: 1000000	293
PkSet 3	Fold Change	2	500
PkSet 4	PkSet 2 AND PkSet 3	Boolean AND	52

(from Grigsby et.al. 2012).

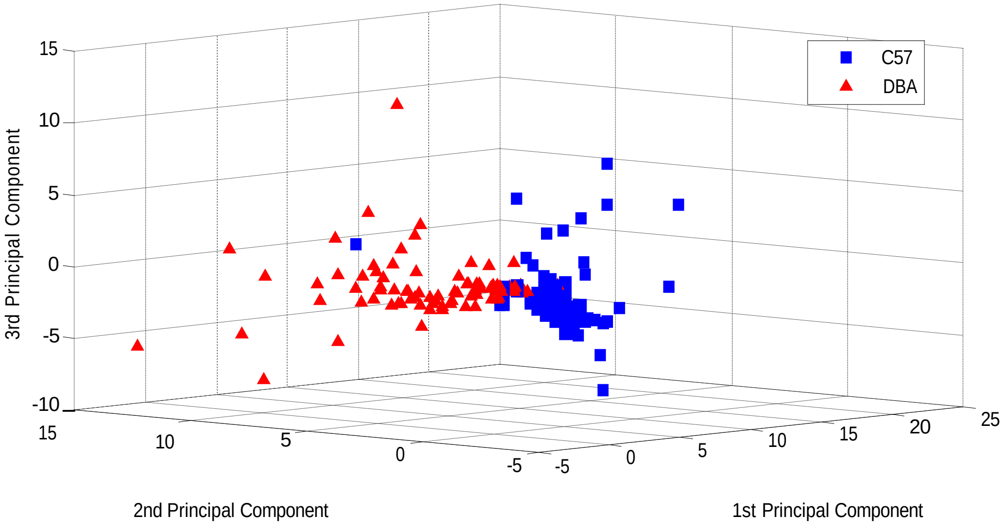
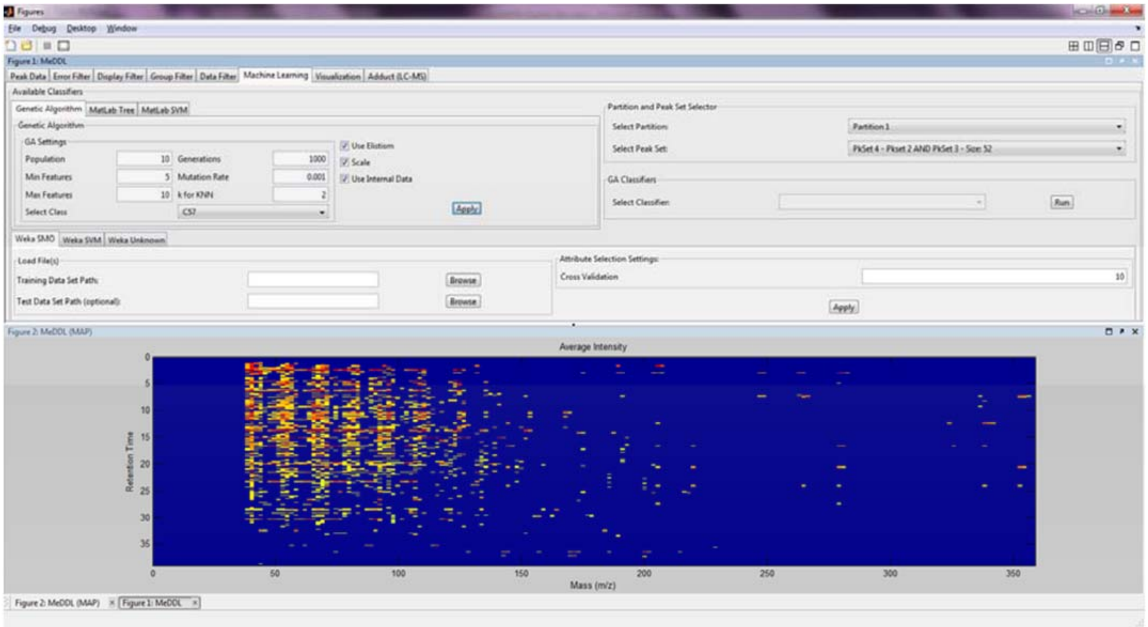


Figure 15: PCA of C57 and DBA Filtered Intersect *(peakset 4)* Results
(from Grigsby et.al. 2012)

MeDDL offers users the ability to utilize several different types of classification methods and separately store the resulting output for classification of additional, unregistered unknowns. These methods use a combination of pre-coded Matlab classifiers, Waikato Environment for Knowledge Analysis (WEKA) classifiers, and the novel, in-house developed hybrid GA classifier, implemented in Java and Matlab, described in this study (Figure 16). The internal data classification allows users to teach the classifiers from peak sets generated using the tool. The external data classification is currently designed to process both CDF files and Comma Separated Value (CSV) files. All classification methods support classifying intensities or ratios of intensities though application of appropriate data filters.


Figures
File Debug Desktop Window
Figure 1: MeDDL
Peak Data | Error Filter | Display Filter | Group Filter | Data Filter | Machine Learning | Visualization | Adduct (LC-MS)
Available Classifiers
Genetic Algorithm | MatLab Tree | MatLab SVM
Apply
Run
Weka SMO | Weka SVM | Weka Unknown
Browse
Apply
Figure 2: MeDDL (MAP)
Average Intensity
Retention Time
Mass (m/z)
0
5
10
15
20
25
30
35
0
50
100
150
200
250
300
350
Figure 2: MeDDL (MAP)
Figure 1: MeDDL

Genetic Algorithm

GA Settings

Population 10 Generations 1000

Min Features 5 Mutation Rate 0.001

Max Features 10 k for KNN 2

Select Class C57

☑ Use Elistism

☑ Scale

☑ Use Internal Data

Partition and Peak Set Selector

Select Partition: Partition 1

Select Peak Set: PkSet 4 - Pkset 2 AND PkSet 3 - Size: 52

GA Classifiers

Select Classifier:

Load File(s)

Training Data Set Path:

Test Data Set Path (optional):

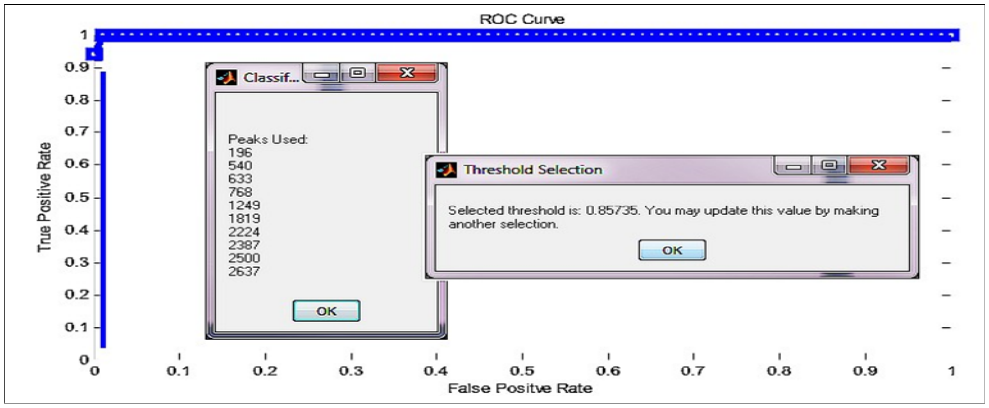
Attribute Selection Settings:

Cross Validation 10

Figure 16: MeDDL Tool Machine Learning Implementation and GA Settings

(from Grigsby et.al. 2012).

In testing the hybrid GA for this study (Figure 17), setting k = 2, minimum features = 4, and maximum features = 10 provided both perfect classification of C57 versus DBA for both the 170 teaching samples as well as the 6 "unknown" external samples. Reverse classification (DBA versus C57) using these same settings resulted in 2 mis-classifications of the "unknowns" illustrating the need to optimize the GA settings for each classifier result.



Peaks Used:
196
540
633
768
1249
1819
2224
2387
2500
2637

Selected threshold is: 0.85735. You may update this value by making another selection.

Figure 17: Hybrid GA Results

Vertical line is user adjustable slider to determine T threshold values (from Grigsby et.al. 2012).

Results of the hybrid GA classifier were comprised of 10 VOC "features", which is the maximum features size allowed by the GA settings. An example of one of the selected VOCs is shown in Figure 18. In an focused biomarker study, each resultant peak would then be preliminarily identified through comparison to the National Institute of Standards and Technologies (NIST) 08 database and Wiley libraries and verified though expert, manual spectral analysis and comparison with purchased standards.

5.8 GC/MS Functionality Conclusions

The MeDDL platform has been markedly improved from the original version, with streamlined analysis of multi-group comparisons through the addition of a more intuitive interface, the ability to dynamically alter group definitions and group comparative displays, and the creation of definable, group comparative graphics. These changes in combination with the addition of machine learning approaches greatly enhance the capability of the tool and future applicability to studies requiring biomarker discovery for sensor and diagnostics applications. In the following chapter, several other examples of the universal utility of the tool in support of various GC/MS differential profiling studies will be described.

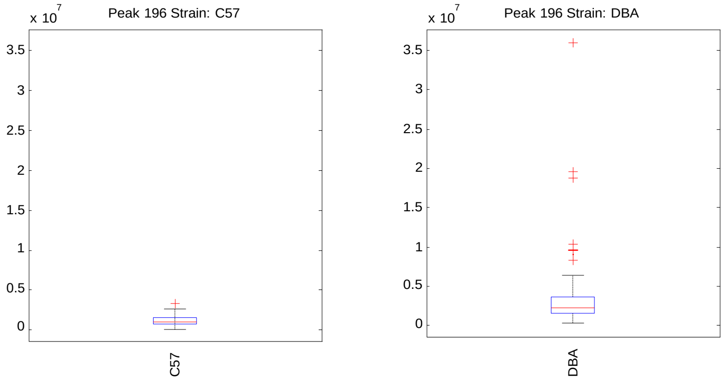


Figure 18: Boxplot of Hybrid GA VOC Feature Output Selected by Classifier

(from Grigsby et.al. 2012)

6.0 GC/MS DATA PIPLELINE

6.1 Overview

To best demonstrate that the application of the comparative algorithms, logical operators, and statistical filters implemented in the developed software tool allow for the efficient analysis of GC/MS based VOC profiling studies, a variety of studies were analyzed and presented below. Some of these studies in particular presented unique data handling challenges and aided in identification of needed enhancements required in future, "real-world" applications described in the following chapter. Analysis of these separate datasets also help provide validation of implemented algorithms with examples provided. Experimental examples of application of the volatile analysis pipeline developed, illustrated in Figure 19, and described below include: characterization of both human and murine urine as it ages, human markers of age and ethnicity in axillary odors, and characterization of the binding between volatile ligands and murine MUPs. GC/MS Experimental Example 1: Differential Binding Affinities Between Volatile Ligands and Urinary Proteins Due to Genetic Variation in Mice

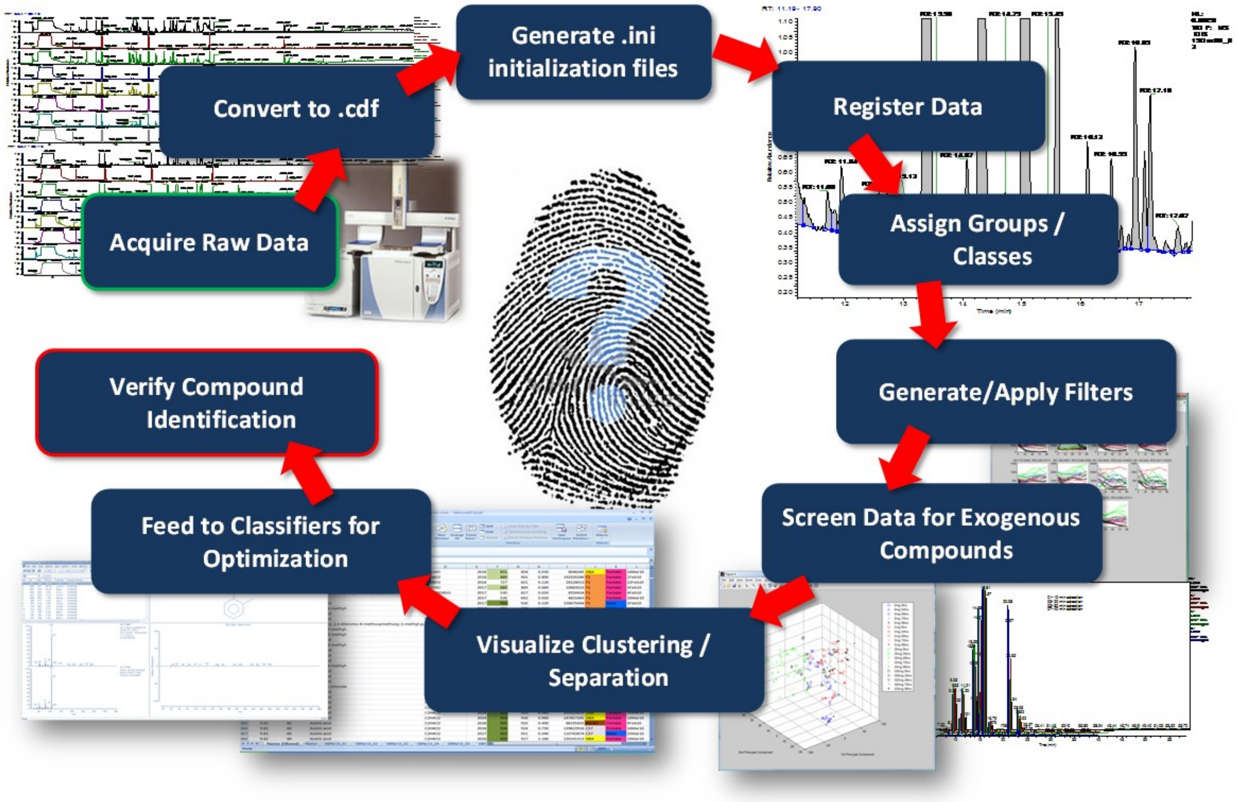

Acquire Raw Data
Convert to .cdf
Generate .ini initialization files
Register Data
Assign Groups / Classes
Generate/Apply Filters
Screen Data for Exogenous Compounds
Visualize Clustering / Separation
Feed to Classifiers for Optimization
Verify Compound Identification

Figure 19: GC/MS Data Pipeline Utilizing the MeDDL Tool

Metabolite Differential and Discovery Lab (MeDDL) Tool

A collaborative study was completed using the MeDDL tool with scientists from Monell Chemical Senses Center (Philadelphia, PA) to investigate the composition of bound and unbound VOCs on the mouse MUPs of various inbred species. This effort was part of an investigation on the nature of murine volatile and pheromone based signaling and is described in detail in Kwak, et. al., 2012 [117]. In short, a comparison of the binding affinities in pooled male urine samples from three different inbred mouse strains (B6, BALB/b and AKR) was completed by measuring the release of volatile ligands before and after denaturation of the MUPs via SPME based headspace concentration and GC/MS analysis. The sample set analyzed consisted of pooled urine samples collected over a 10-day period, one from each of the individual mice in the experiment (N = 8 B6, 7 AKR, and 6 BAL/b mice respectively). Raw spectral data of both intact and denatured murine urine was supplied by Monell along with associated meta-data of the study group samples. A sampling of the raw spectral data was first reviewed for chromatographic and spectral reproducibility across the major urinary peaks, and then for instrumental absolute baseline and peak intensities to be used for spectral registration and alignment. All data was then converted to netCDF format and aligned/registered via the MeDDL tool, resulting in a total of 1895 peaks.

Following registration and alignment, the GC/MS data were analyzed by PCA and the fold change filter as described in section 4. Screening for group separation via PCA, shown in Figures 20a-b demonstrate clear separation between each of the six different groups as well as between intact and denatured groups, indicating that both strain difference and protein denaturation are important contributors to the unique volatile profiles. To identify those peaks responsible for the separation observed via the PCA, we used a combination time-binning (due to EI fragmentation), tests for fold-change and statistical significance, and absolute intensity using the below parameter values. We restricted the fold-change to only those time windows which demonstrated two fold increased levels of change upon protein denaturation (positive change) within a specified time slice of 0.1 minutes and an intensity threshold of 300,000 absolute intensity (total ion count). A test for significance between all comparisons was also completed using N-way ANOVA ($P \leq 0.1$), which included Bonferroni correction to compensate for the tendency to incorrectly find a single pairwise significant difference among multiple comparisons. Results obtained from the three strain specific pair-wise comparisons of intact to denatured samples displayed the increased release in 49 peaks induced by protein denaturation in AKR, 26 in B6, and 36 in BALB/b, respectively. Figure 21 illustrates the changes in the release of ligands upon denaturation. GC/MS Experimental Example 2: Changes in Volatile Compounds of Mouse Urine as it Ages: Their Interactions with Water and Urinary Proteins

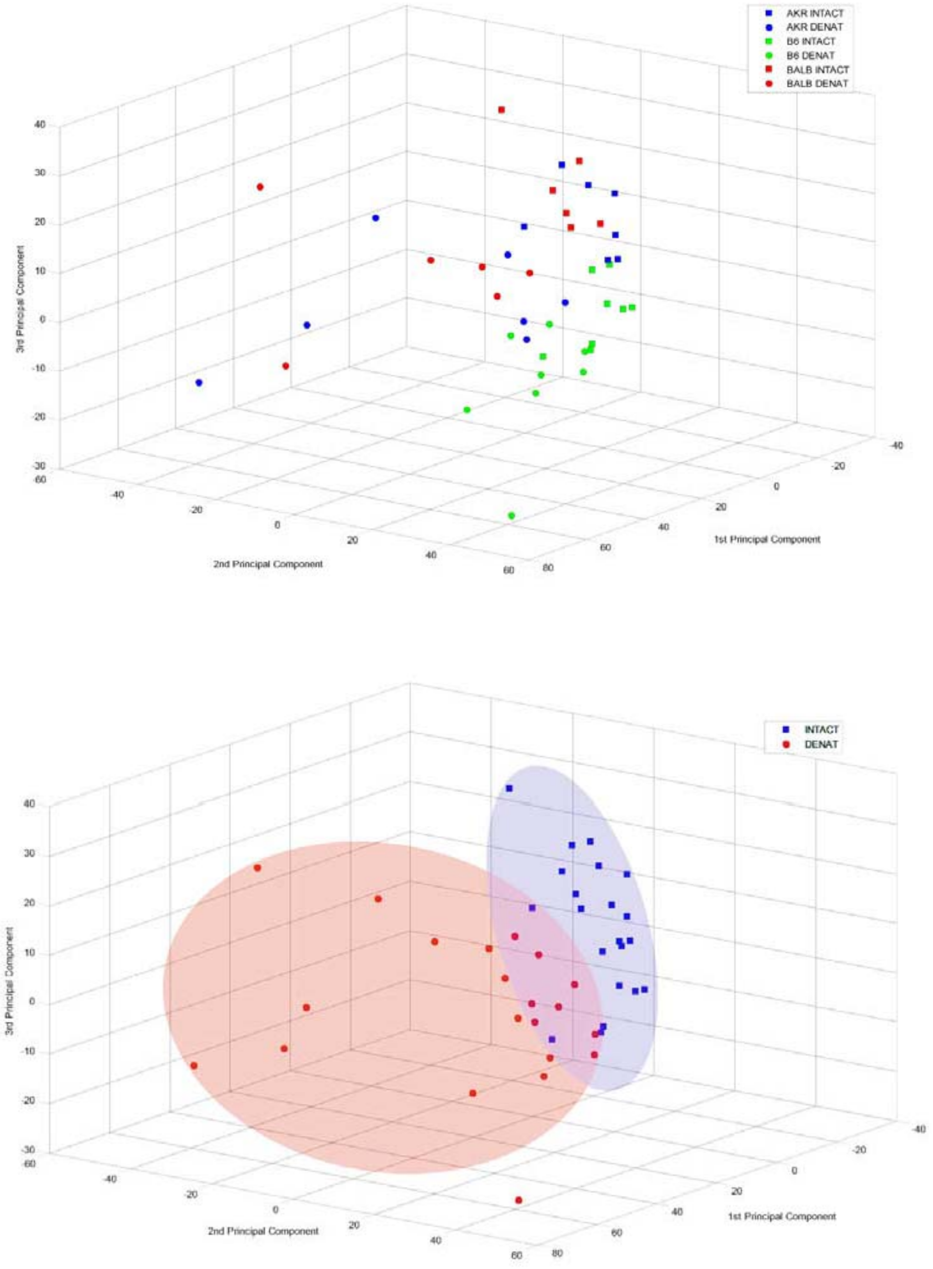


Figure 20: PCA of GC/MS Data

The six different groups (a) and for the intact and denatured groups (b) from Kwak et.al., 2012.

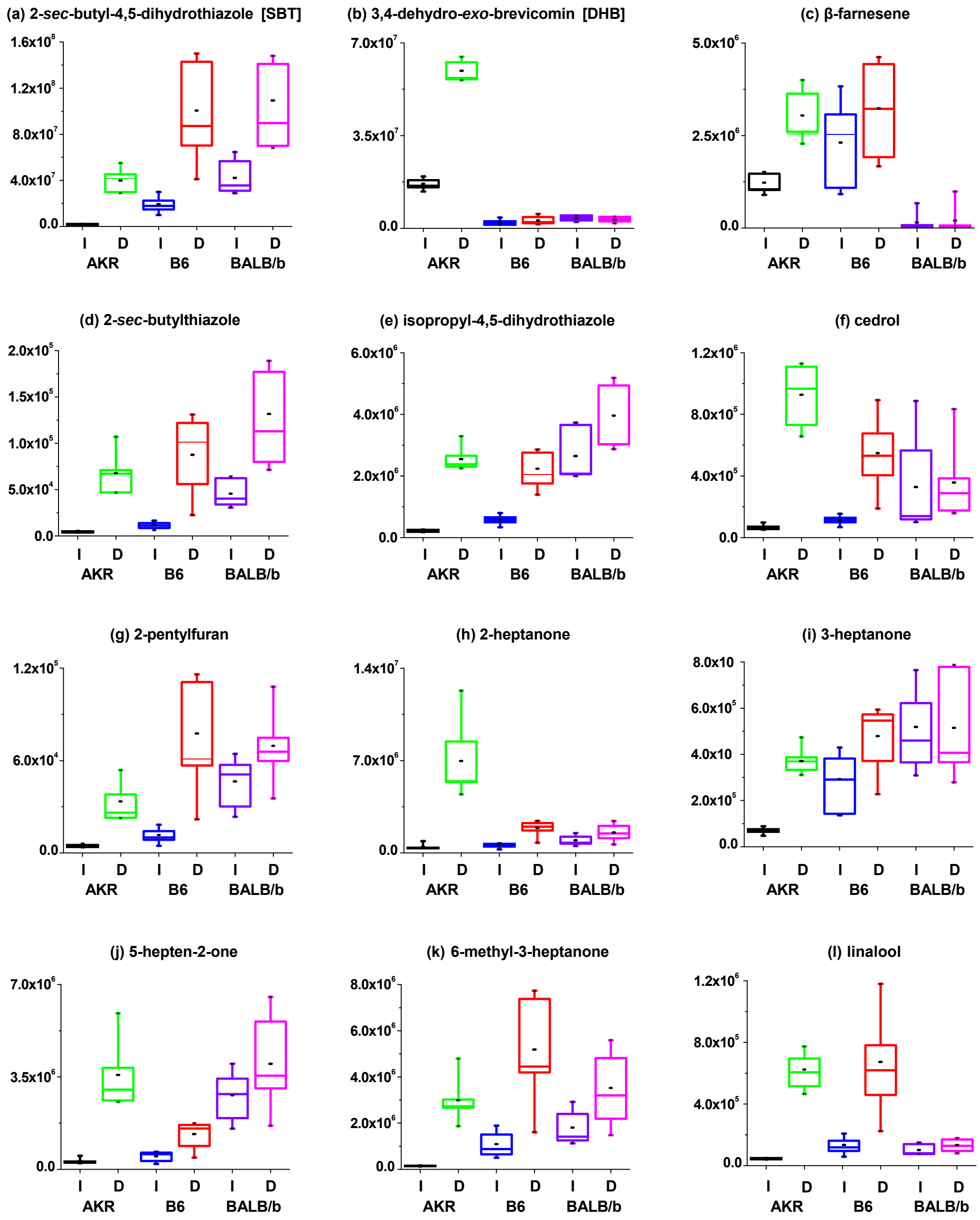


Figure 21: Changes in the Release of Ligands upon Protein Denaturation

Changes in the release of ligands upon protein denaturation in the urine samples derived from different mouse strains. The y axis indicates the absolute intensity of the base peak ion in each ligand. I: intact urine; D: denatured urine. The degree of release in each ligand was distinctive in each strain.

A collaborative study was completed using the MeDDL tool with scientists from Monell Chemical Senses Center (Philadelphia, PA) to investigate changes in the volatile composition of mouse urine as it ages. This effort was part of an investigation on the nature of murine volatile and pheromone based signaling and is described in detail in Kwak, et. al., 2013 [118]. In short, the amount of water in an aqueous sample influences releases of VOCs from the sample. As the sample dries, evaporations of water-soluble VOCs accelerate, whereas the loss of water may render some VOCs to bind to solid surfaces, preventing them from being released into the air. A number of studies measured the loss of VOCs as male mouse urine aged. Some VOCs were removed rapidly, whereas others were released slowly. However, the previous studies did not clearly demonstrate whether the gradual releases of the VOCs were due to their binding to MUPs and/or due to the loss of water as urine became dried. In addition to the roles of water in the release of VOCs mentioned above, the loss of water in urine may alter the structure of MUPs, losing their ability to retain volatile ligands. Here, we investigated the effect of water loss on the releases of VOCs while mouse urine dried, and determined whether the ligand-binding ability of MUPs in the dried urine remains active. Using similar sample collection and GC/MS methodologies utilized in Kwak, et. al. 2012, a data set comprised of 24 chromatograms were generated from a pooled collection of 4 - B6 male mice and 3 - B6 female mice acquired over several days by analysis of 12 aliquots of each sex pool. These 24 aliquots (12 for male and 12 for female) were separated into 6 condition groups, giving an N=3 per treatment. Treatment groups consisted of "intact", "aged", "aged + water", and "aged + water + GdmCl" with each treatment described in detail in Kwak, et. al., 2013.

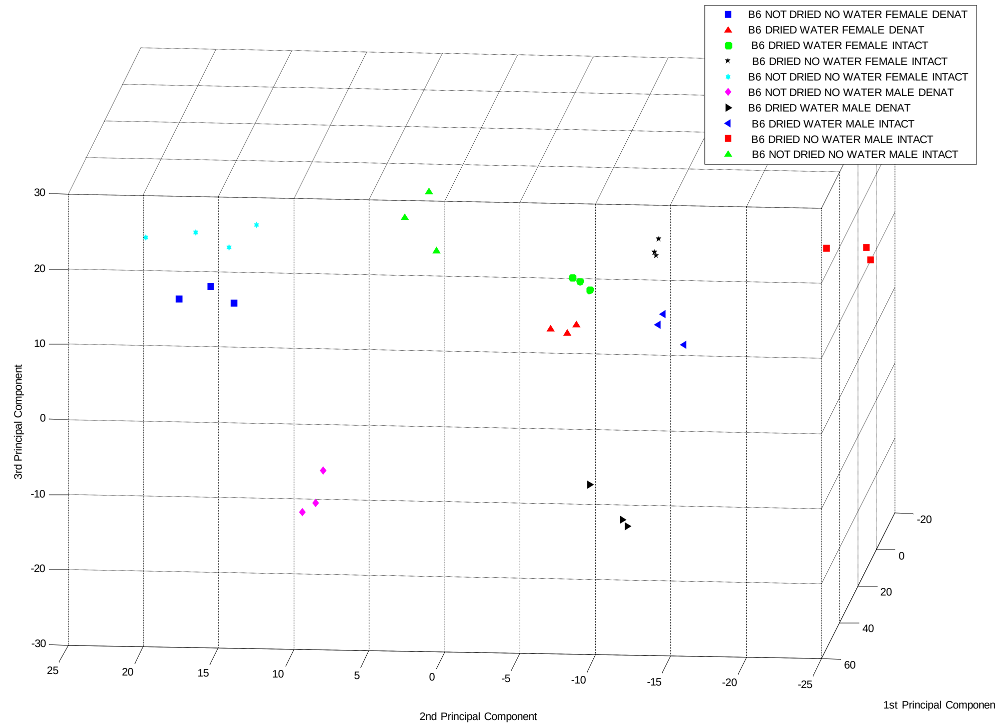


Figure 22: PCA of Treatment Groups

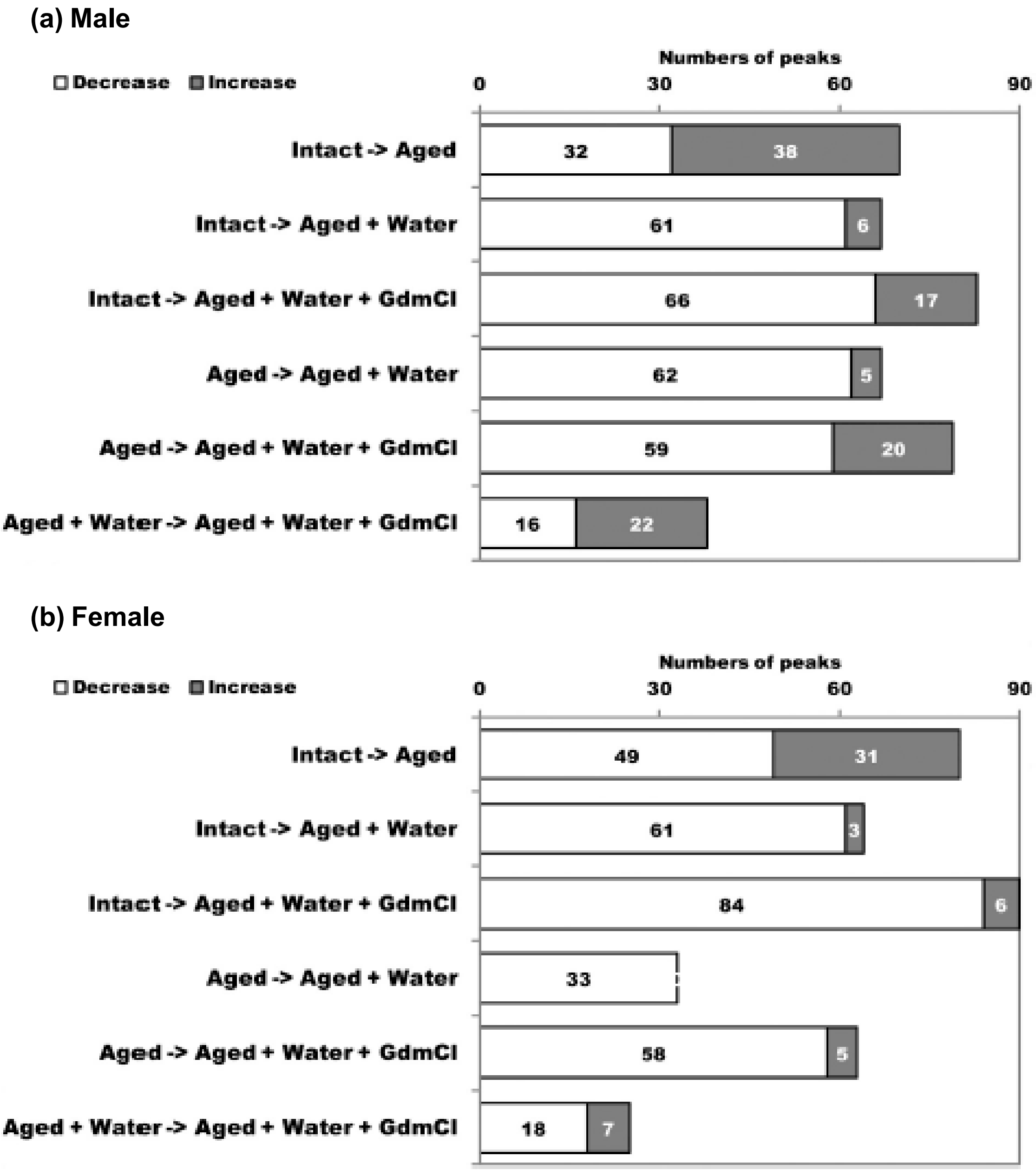
PCA of the treatment groups from showing clear feature separation

Raw spectral data was supplied by Monell along with associated metadata of the study group samples. A sampling of the raw spectral data was first reviewed for chromatographic and spectral reproducibility across the major urinary peaks, and then for instrumental absolute baseline and peak intensities to be used for spectral registration and alignment. All data was then converted to netCDF and aligned/registered via the MeDDL tool, resulting in a total of 1166 peaks. Following registration and alignment, the GC/MS data were analyzed by PCA and the fold change filter as described in Section 4.0. Screening for group separation via PCA, shown in Figures 22, shows clear separation between each of the treatment groups, indicating that sex, MUP denaturation, and hydration state are important contributors to the unique volatile profiles. To identify those peaks responsible for the separation observed via the PCA, we used a combination time-binning (t = 0.1 min) inclusive of only those peaks >200 K absolute intensity and tests for fold-change (≥ 2). Once each of these filters was applied to the grouped, global data set, a Boolean "AND" was added to the resulting filtered peak sets to identify their logical intersection. fold-change (≥ 2). Once each of these filters was applied to the grouped, global data set, a Boolean "AND" was added to the resulting filtered peak sets to identify their logical intersection, resulting in 142 compounds for further investigation (see Figure 23a-b), which is described elsewhere (see Kwak et.al. 2013).

GC/MS Experimental Example 3: Changes in Volatile Compounds of Human Urine as it Ages: Their Interaction with Water

A collaborative study was completed using the MeDDL tool with scientists from Monell Chemical Senses Center (Philadelphia, PA) to investigate changes in the volatile composition of human urine as it ages. The urinary odors emitted from toilet facilities and from individuals suffering from either incontinence or metabolic disorders are perceived as unpleasant. As anecdotal reports suggest that the odor of aged urine is different from that of fresh urine, using techniques described above for murine urine this study sought to identify the specific, differential VOCs released from aged human urine. As described in Kwak et.al. 2013 [119], the urine samples analyzed consisted of a pooled sample collected from 6 adults (3males and 3 females). From this pooled sample, six 1 mL aliquots were prepared, with each aliquot placed in a 60 mL glass jar and capped.

Three of the prepared aliquots were analyzed immediately via SPME based headspace concentration and GC/MS analysis and served as the "Intact" study group. The remaining three aliquots were left uncapped in a ventilated chemistry hood for 24 hours until nearly dried and served as the "Aged" study group. Raw spectral data was supplied by Monell along with associated metadata of the study group samples. As described above, a sampling of the raw spectral data was first reviewed for chromatographic and spectral reproducibility across the major urinary peaks, and then for instrumental absolute baseline and peak intensities to be used for spectral registration and alignment. All data was then converted to netCDF and aligned/registered via the MeDDL tool, resulting in a total of 384 peaks. To identify the differential VOCs, the following filter settings were used: the fold change filter was limited to those peaks with two-fold or greater change in intensity and the time binning filter parameter was set using 0.1 min bins and inclusive of only those peaks >10K absolute intensity.


(a) Male

(b) Female

Figure 23: The Number of Binned Peaks which Displayed Two-Fold or Greater Change
absolute intensity changes in each pairwise comparison (from Kwak et.al., 2013)

Once each of these filters was applied to the grouped, global data set, a Boolean "AND" was added to the resulting filtered peak sets to identify their logical intersection. Results of the time binning filter generated 103 discrete peaks and combination with the fold filter (Intact vs Aged) resulted in 58 differential VOCs. The 58 peaks were identified and a list of the identified VOCs is provided in Table 7 (from Kwak et.al. 2013).

Table 7: A List of the Volatile Compounds whose Levels Changed after the Urine Samples were Aged [119].

Retention time (min)	Base ion	Compound ID	Fold change (Intact -> Aged)
1.32	42	*trimethylamine**	5.05
2.15	43	acetone	29.66
2.90	43	2-butanone	17.43
3.22	45	ethanol	11.71
4.06	43	2-pentanone	51.83
4.42	43	2-methyl-3-pentanone	13.17
4.62	58	2-hexanone	8.19
5.59	43	3-hexanone	11.04
6.15	57	3-heptanone	32.95
6.50	43	3-ethyl-2-pentanone	4.12
6.88	43	3-methyl-2-hexanone	3.74
7.39	43	4-heptanone	145.33
8.86	43	2-heptanone	12.88
9.36	43	3-methyl-2-heptanone	12.81
9.57	57	6-methyl-3-heptanone	11.69
9.82	69	3-methylcyclopentanone	14.45
9.92	71	2-methyl-4-heptanone	5.42
12.40	43	4-nonanone	2.42
12.76	83	3-ethylcyclopentanone	22.58
13.02	72	3,5-dimethyl-2-octanone	21.16
13.92	58	2-nonanone	11.55
15.68	43	*4-hydroxy-2-pentanone**	2.17

15.83	112	p-menthan-3-one	80.09
16.06	60	*acetic acid**	7.01
18.44	45	*propylene glycol**	14.58
20.45	74	*2-methylbutyric acid**	11.19
20.50	60	*3-methylbutyric acid**	3.26
20.97	145	3,6-dimethylbenzofuran	15.88
21.39	110	p-menth-1-en-3-one	32.04
21.48	107	a terpene	31.94
21.80	59	*acetamide**	2.61
22.77	42	*caprolactone**	4.41
24.58	79	*dimethyl sulfone**	11.96
26.78	71	*pentolactone**	7.34
27.47	107	p-cresol	6.59

(indicates the compounds whose absolute intensities increased after the samples were aged.)*

GC/MS Experimental Example 4: Human Volatile Markers of Age and Ethnicity in Axillary Odor profiles

A collaborative study was completed using the MeDDL tool with scientists from Monell Chemical Senses Center (Philadelphia, PA) to investigate human volatile markers of age and ethnicity in axillary odor profiles. Volatile organic compound biomarker discovery in humans is especially challenging due to significant variances in diet, environmental conditions/exposures, and genetics. Thus, creation and validation of outlier filtering and normalization approaches for spectral data is required for the differential analysis of many studies involving human subjects, with four approaches evaluated for this study: Total Ion Current (TIC) based normalization; "Olympic average" based normalization; outlier filtering; and "Group Distribution" based filtering. Please note that these are just a few of the filter types available in MeDDL, with a more complete listing available in Appendix A. In short, TIC normalization involves summing the intensities of all peaks contained in each file of the spectra and setting that sum as equal across all files. This approach, however, can be skewed by the presence of high intensity outliers, thus "Olympic average" based normalization was implemented. In "Olympic average" based normalization, user defined upper and lower percentiles are calculated across all of the files using user defined values. This will generate in two vectors containing percentiles, each with a size of n by 1, where n is the number of features. These percentile vectors are used to normalize the intensity values for all ions across each file. The difference of the original intensities in a file and the lower percentile value is divided by the difference of the upper and lower percentiles. This technique will scale down the intensities by several orders of magnitude, but should be less susceptible to the presence of low/high intensity outliers. Next, the outlier filter, using an operator defined threshold, removes those peaks that are outside of the range of the n number of standard deviations from the mean. The user specifies the number of standard deviations to accept and data outside of this range will not be shown throughout MeDDL (i.e. this is a display filter, not a data exclusion filter). It should be noted that as this filter replaces excluded data points from the display with Not a Number's (NaNs), some comparisons and plots, such as PCA, cannot be generated using the modified data set. The final filtering technique evaluated, and the one utilized for study, is a similar mean distribution filter to the "outlier" display filter labeled "group distribution" as currently implemented in MeDDL. In this filter, the means and standard deviations are calculated across all files for each peak. Each value is checked to ensure that it is within the user-defined number of standard deviations from the mean, and if the peak is more standard deviations away then the given threshold, the peak is excluded from the final peak set.

In brief, for sample generation by Monell scientists, body odors were collected from 40 female subjects consisting of N=10 of four ethnic and age group (Young Caucasian, Young Asian, Older Caucasian, and Older Asian) under a protocol approved by the University of Pennsylvania Institutional Review Board. Subgroups of these extracts were combined to form super-donors to eliminate sensory panelists focusing on individual donors. Prior to skin extraction, subjects were screened via collected SPME samples prior to formation of super-donors to insure that no donor was an outlier (e.g., unusual VOCs such as bromoform from swimming pool water). The raw GC/MS data from these collected samples was supplied by Monell along with associated metadata of the study group samples. As described above, a sampling of the raw spectral data was first reviewed for chromatographic and spectral reproducibility across the major peaks, and then for instrumental absolute baseline and peak intensities to be used for spectral registration and alignment. All data was then converted to netCDF and aligned/registered via the MeDDL tool to look at inter-age and inter-ethnicity variation in skin VOCs.

A total of 2960 peaks were registered. To identify the differential VOCs, the following filter settings were used: the fold change filter was limited to those peaks with two-fold or greater change in intensity and the time binning filter parameter was set using 0.1 min bins and inclusive of only those peaks >100K absolute intensity. A minimum threshold hold filter was also applied to include only those peaks exceeding an absolute intensity threshold as low level, trace compounds were not considered relevant to the funding sponsor (deodorant manufacturer). Next, the group distribution filter, described above, as applied and excluded all peaks > 3 standard deviations from the group mean. Once each of these filters was applied to the grouped, global data set, a Boolean "AND" was added to the resulting filtered peak sets to identify their logical intersection. Results of the each of the respective filters are shown below, and the resulting intersection indentified 55 differential VOCs.

PkSet 1 - All Peaks - Size: 2960
PkSet 2 - P-Value <= 0.1 - N-way - Size: 2063
PkSet 3 - Time binning - Delta T: 0.1 Min. Int.: 100000 - Size: 419
PkSet 4 - Fold Change of 2 - Size: 2123
PkSet 5 - Group Intensity Filter of 300000 - Size: 785
PkSet 6 - Group Distribution Filter of 3 - Size: 332
PkSet 7 - Pkset 2 AND PkSet 3 AND PkSet 4 AND PkSet 5 AND PkSet 6 - Size: 55

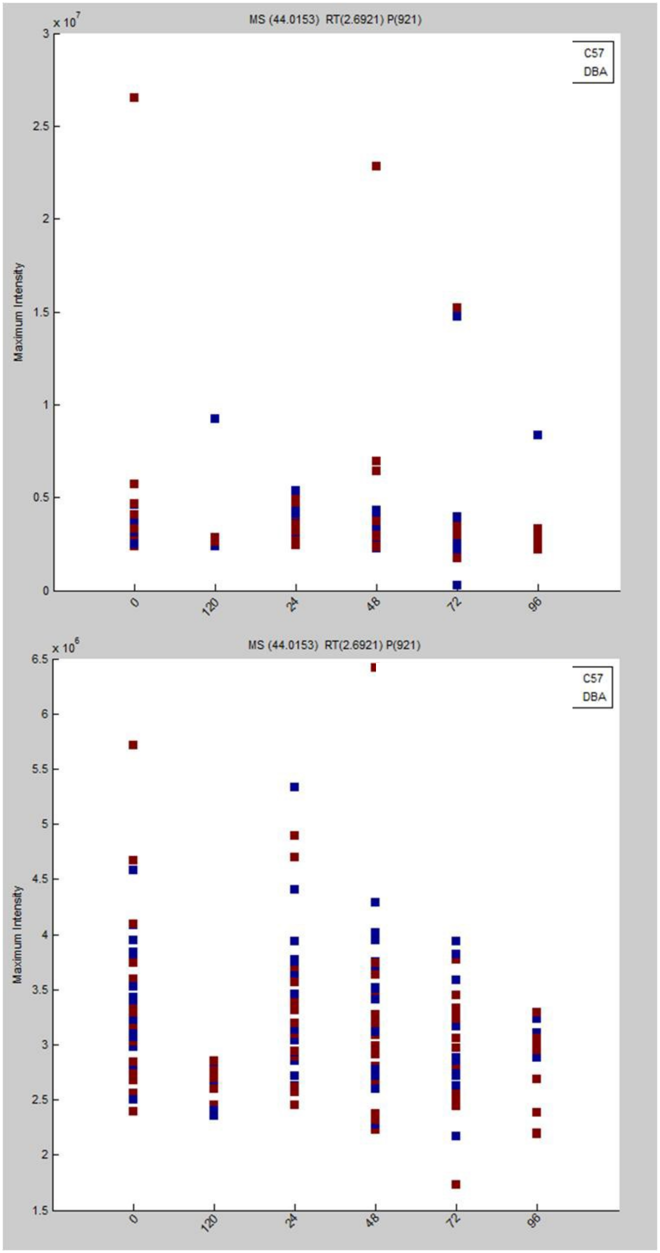


Figure 24: Outlier Filter Example

The figure on the top was generated using the original data and the figure on the bottom was generated after the outliers were removed. A setting of 1 standard deviation was used in this case. Notice the difference in scale between the two images

Results of principal component analysis and unsupervised hierarchical clustering of this filtered data set are shown in Figure 25 and demonstrated definitive separation between Asian and Caucasian older age group females.

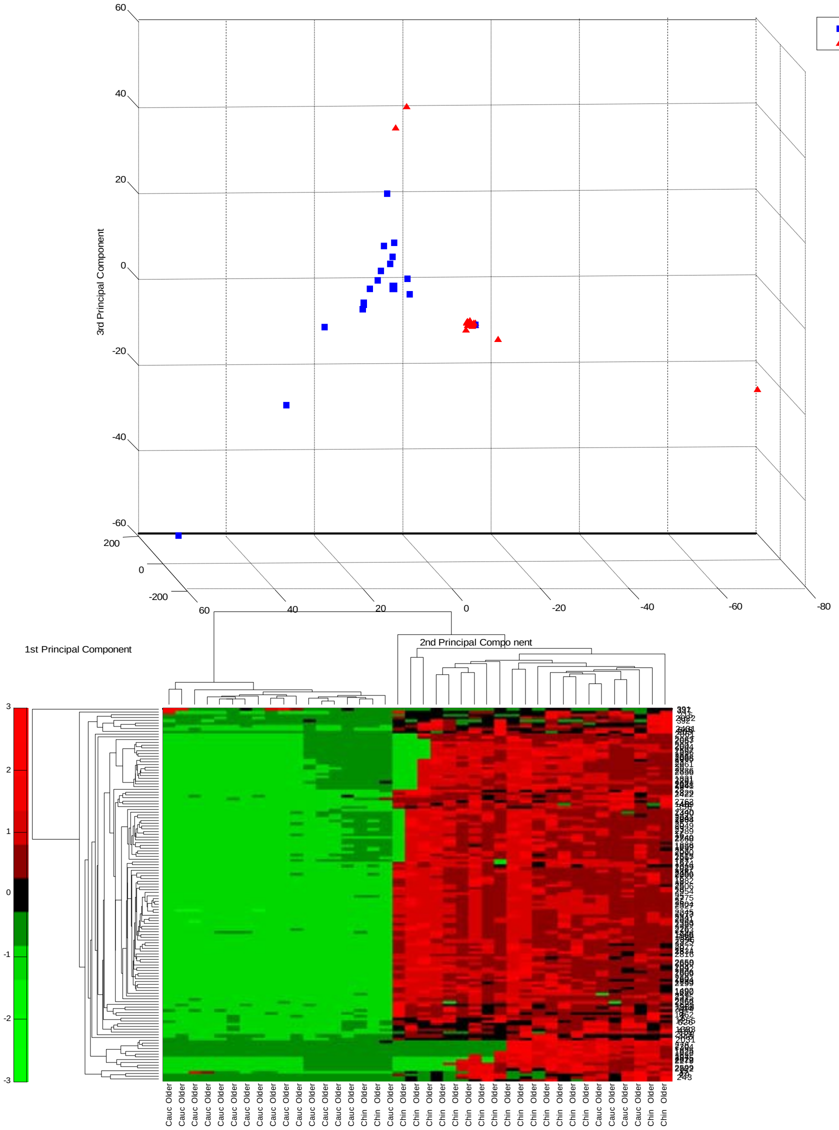


Figure 25: Filtered Data from Age and Ethnicity-Based Skin VOC Differences

6.2 Conclusions

Given the unique requirements for the various GC/MS based biomarker studies described above, comparative analysis, via MeDDL has facilitated efficient of each and successfully demonstrated not only application and utility of the expanded GC/MS functionality but also the versatility of the approach. This described methodology is representative of our analytical pipeline and is applicable to a wide range of VOC and small molecule based differential profiling and biomarker discovery applications such as human performance monitoring, odor based biometrics, medical diagnostics, targeted materials detection, and environmental health and safety investigations, some of which will be addressed in the following section.

7.0 COMPREHENSIVE SAMPLING PIPELINE

7.1 Overview

As addressed earlier, studies which require complete characterization of an unknown, whether it is biologically derived, such as human breath, or environmental, as in the case of fighter cockpit air, necessitate varied approaches to collection for accurate and broad-spectrum capture of all potential molecular species present. To address study requirements for work performed by our laboratory and to complete the end-to-end analytical pipeline which was the focus of this research, several "real-world" sample collection methods were developed and implemented for human breath capture as well as characterization of the in-flight, pilot air supply in operational combat aircraft, one of which is presented below. As covered in chapter 1, a variety of methods are available for volatile sampling. To be able to both quantitate obtained samples and to increase the likelihood of isolating potential environmental contaminants, thermal desorption tube based approaches were selected with follow-on analysis being completed utilizing a modified version of EPA Method TO-17 for monitoring VOCs [120], described below.

7.2 In-Flight Air Quality Sampling

Our laboratory supported investigation of reported issues surrounding the On-Board Oxygen Generating System (OBOGS) utilized on the F22 and other USAF fleet aircraft. As part of this effort, a technical solution was developed by our group to sample and analyze both the cockpit ambient air and OBOGS product air. This involved placing TD tubes in line with the summa canisters already planned for placement into the OBOGS product line and into the cockpit air. This solution allowed for a much more complete picture of the chemical make-up of contaminants in the oxygen supply and cockpit air. The TD tubes provided the ability to capture and analyze heavier molecular weight chemicals (semi-volatiles) that the summas cannot be analyzed for. The summas were still necessary to test for very light organics such as freons and inorganics such as oxygen (O_2), carbon monoxide (CO), and carbon dioxide (CO_2). The summas, when timed with a critical aperture, also provided the airflow through the TD tube so that a known volume was collected for concentration quantification while avoiding the need for additional connections and equipment to pull air through the tubes (see Figure 26). Empirical validation of this sampling strategy showed that a 1 hour collection using the summa canister as a passive pump in line with the TD tube resulted in a 400ml total volume collection for both. This value was used to establish corresponding calibration curves for quantitation of detected compounds.

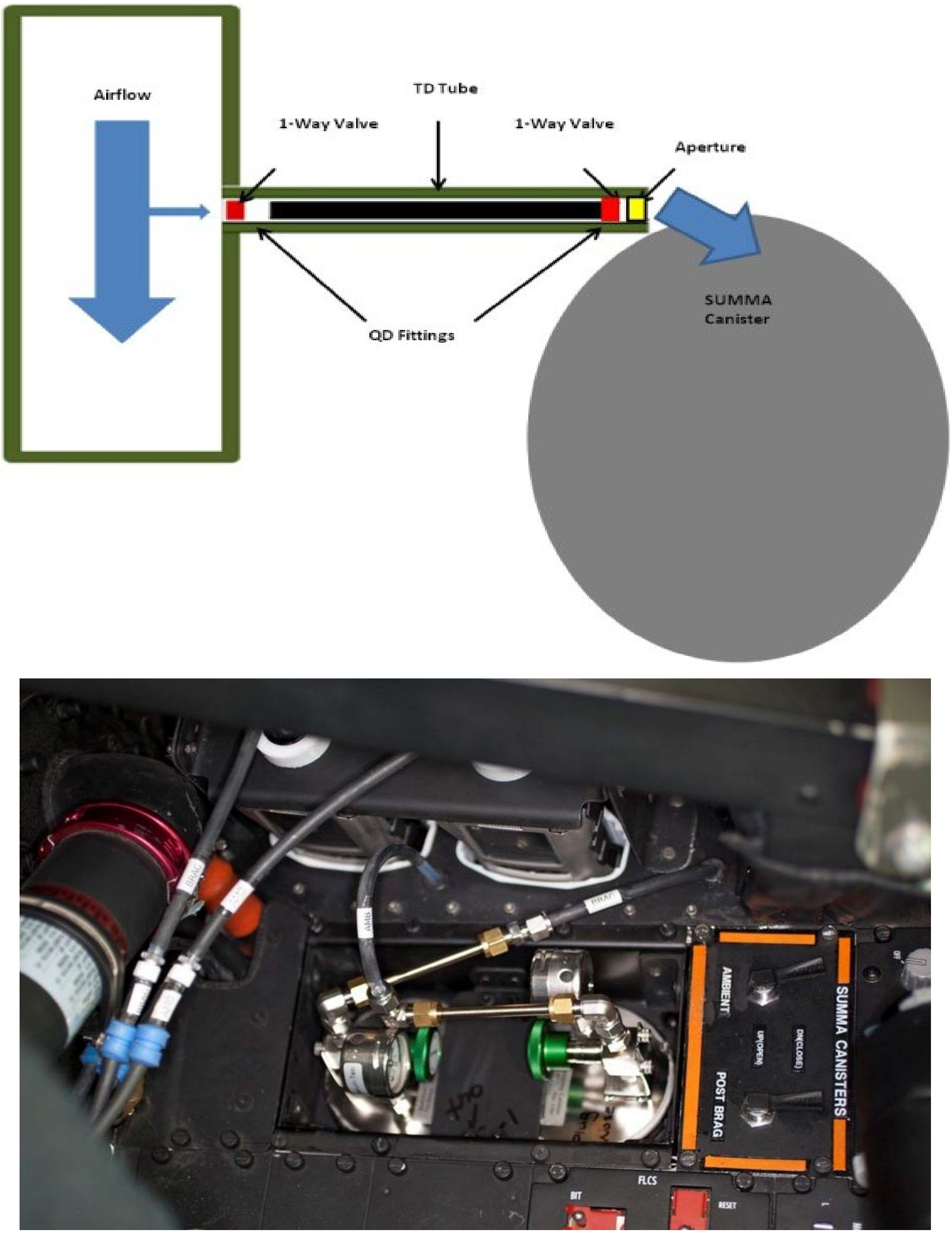


Figure 26: Air Sampling Strategy for In-Flight Collection

Top figure describes airflow through the TD tube and summa canister. Bottom figure shows engineered solution for collection of both cockpit ambient air and OBOGS product supply to pilot.

As stated above, all TD tubes were analyzed using a modified version of EPA method TO-17 for the determination of toxic organic compounds in ambient air using active sampling. GC/MS analysis conditions are listed as follows. TD100 TD auto-sampler parameters: cold trap low temp.: 15°C; Tube desorption temp.: 310°C; tube desorption time: 10 min; trap purge time: 1.0 min; cold trap high temp: 315°C for 5 min; split ratio: split-less; trap heating rate: 40°C/s (MAX); TD flow path: 160°C. Thermo GC Ultra parameters: carrier gas: He; DB-624 column: 60 m x 0.32 mm x 1.80 µm; constant pressure mode: 10 psi; temp. program: 40°C (1 min), 10°C/min to 240°C (20 min). Thermo ISQ conditions: MS source temperature: 275°C; transfer line temperature: 230°C; full survey scan mode mass range: 35 to 550 amu.

Daily quality assessment/control procedures involved running a blank with internal standard and passing tune criteria for bromofluorobenzene (BFB). The internal standard is a 4 component mix composed of two internal standards (1,4 difluorobenzene and chlorobenzene-d5), a surrogate (bromochloromethane), and a tuning compound (bromofluorobenzene). Additionally, prior to processing samples, a laboratory Control Sample (LCS) and Calibration Check Verification (CCV) were completed at the mid-point of the calibration curve. The mid-point calibration curve was 25ppbv. The LCS criterion is ± 30% of 25ppbv (17.5 – 32.5 ppbv).

A two-pronged approach was used for data analysis. First, a commercial Thermo Scientific software product, Tracefinder Environmental, Food, and Safety (EFS) version 2.1 was used to store calibration results and generate quantitative/qualitative reports of TO-17 chemicals detected on the sorbent tubes. This is performed by either matching sample peaks against a user generated library and calibration curves using a combination of retention time, quantitative ion, and confirming ions, or as an unknown/tentatively identified compound via automated NIST11 search. The limit of detection (LOD) established for most of the TO-17 compounds was established by our laboratory at 2ppbv. All reports generated by Tracefinder were exported and summarized into Excel spreadsheets for data consolidation using an in-house Visual Basic software macro.

The second approach to data analysis was completed through use of the MeDDL tool (described above). Following registration and alignment, the defined experimental result groups were processed through a logical combination of the three following data filters: fold change limiting results to only those peaks which demonstrated at least 2 fold or greater change in intensity; N-Way ANOVA with only those peaks having $p < 0.1$ significance between groups; and the time binning filter using 0.1 minute bins inclusive of only those peaks > 200K absolute intensity. Once each of these filters was applied to the grouped, global data set, a Boolean "AND" was added to the resulting filtered peak sets to identify their logical intersection. These reduced data sets were then used for further manual, statistical, and machine learning comparisons and subsequent compound identification.

A total of 116 sorties were flown using this sampling strategy with 65 sorties completed at Langley/Hunter Air Force Base (AFB) and 51 sorties at Joint Base Elmendorf Richardson (JBER). Note that the majority of the 65 sorties were flown from Langley AFB. Each sortie generated a set of summa canister and thermal desorption tube data for both cockpit and OBOGS air samples (Figure 26). Summa canister samples were analyzed using EPA TO-15 method by Columbia Analytical (ALS-Columbia) located in Simi Valley, CA. Although specific data is not available for public release, these investigations identified trace amounts of chemicals in the system, none of which were above known hazardous concentrations, and were what was expected based on laboratory testing of the system's ability to remove contamination. (Figure 27) Follow on studies are planned to determine if the chemicals and concentrations measured are typical among different fighter airframes vs. those collected and will provide a further understanding of fighter aircraft air quality baseline.

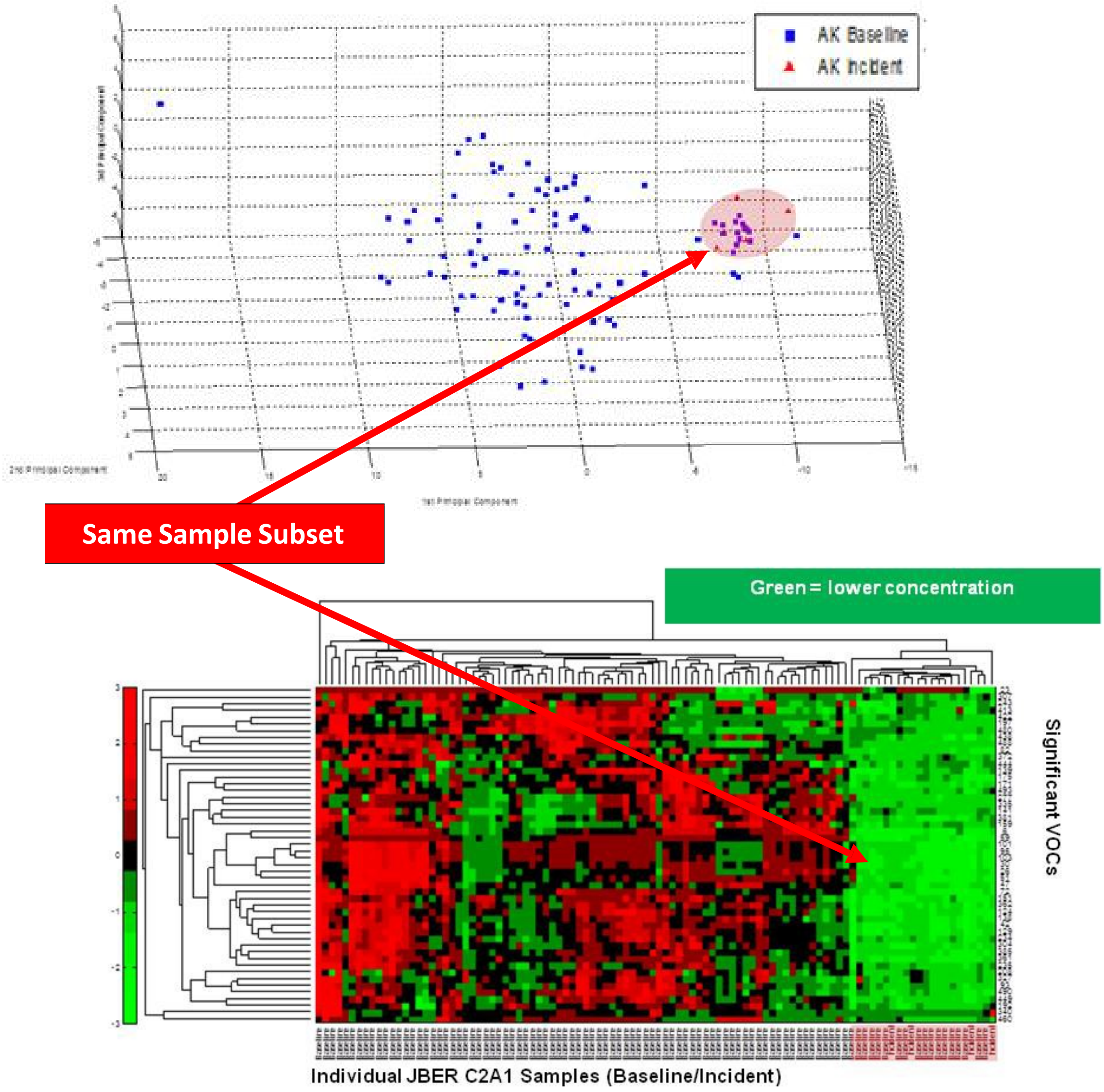


Figure 27: Example of In Flight Data Analyzed using the MeDDL Tool

Samples consisted of 414 charcoal filters analyzed by thermal desorption (modified EPA TO-17).

8.0 CONCLUSIONS AND FUTURE DIRECTION

8.1 Informatics Pipeline

Through a combination of the base MeDDL registration and alignment algorithms and the described additional functionality, MeDDL now offers the analytical chemist the potential for visualizing data in new ways, providing novel insight into the experimental results, and expediting LC/MS and GC/MS based biomarker discovery. Modifications to the current implementation of the tool are on-going, with automated iteration across available "unknowns" for optimization of the hybrid GA parameter settings planned. The overall framework was rapidly prototyped using MathWork's Matlab software language and is being translated to the general purpose, platform independent language, Python, to support wide dissemination of the tool. The MeDDL tool dramatically reduced manpower costs in our research by providing scaffolding for the rapid development and verification of new algorithms without the need to create a large amount of supporting software. MeDDL also offered the potential for staff scientists to visualize data in new ways, providing novel insight into the experimental results and facilitating metabolomic biomarker discovery. It should be noted that a number of tools have recently been proposed in the literature which show great advancements in metabolomic and LC-MS analysis capability [95-99, 121]. However, the MeDDL tool, through its emphasis on visualization, provides unique opportunities by combining the following: easy use of both GC-MS and LC-MS data; use of both manufacturer specific data files as well as netCDF;

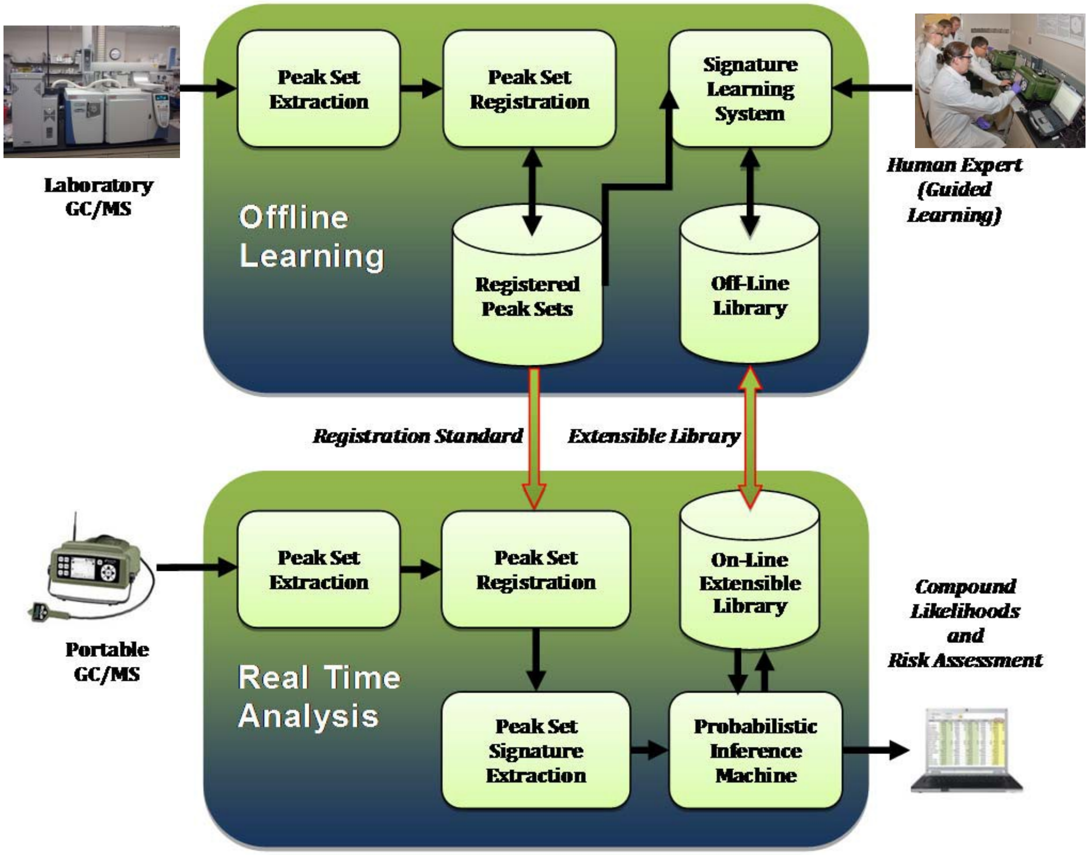


Figure 28: Two-Phase System for Compound Identification and Risk Analysis

preprocessing (peak registration and alignment in both time and mass); powerful visualization tools; and built in data analysis functionality.

A new effort to build on the success of the MeDDL tool has been initiated by our laboratory. Portable GC/MS systems are currently deployed by the DoD in a variety of situations where the identification of the presence of various VOCs is required for protection of forces. These systems rely on limited, build in VOC libraries and NIST reference spectra for identification of individual component compounds and are typically operated by personnel with minimal technical background. Often times, materials of interest/concern are actually composed of complex mixtures, whose relative component ratios are the determining factor required for identification. This research aims to enhance the capability of the fielded, portable GC/MS systems by providing rapid, on-site volatile signature recognition. The prototype software system is envisioned as being composed of two cooperative sub-systems (Figure 28). The offline learning subsystem will be used in a laboratory setting to assemble an extensible library of labeled sensor features representing known volatile organic compounds. The labeled library of VOC's will seed the second software subsystem; the online VOC analysis system, capable of supporting real time analysis of unknown VOC's captured using a field deployable GC/MS system. Both subsystems will exploit algorithms and techniques developed as part of the MeDDL system described above as well as a probabilistic pattern recognition scheme previously utilized for processing synthetic-aperture radar data.

8.2 Sampling Future Direction: Exhaled Breath Collection and Biomarker Discovery

Screening for VOC's in exhaled breath was an important tool during previous aircraft investigations support by our laboratory. However, there were significant limitations in the assessment of the data as there were noteworthy levels of VOC's observed in some incidents but it was not clear if those compounds were generated from physiological/ metabolic functions within an individual or from an environmental exposure that may have occurred during the flight. Additionally, the summa canister based method previously utilized [122] for sample collection, while very useful for trapping and analysis of VOC's with boiling points near that of naphthalene and below, is not capable of releasing higher molecular weight compounds which compose the majority of fuels/fluids utilized on aircraft. On-going work by our group seeks to address the later of these concerns through both the characterization and evaluation of emerging and novel methodologies for breath collection (see Figure 29) as well as identification of breath markers of aviation relevant physiological states such as hypoxia, fatigue, and stress.

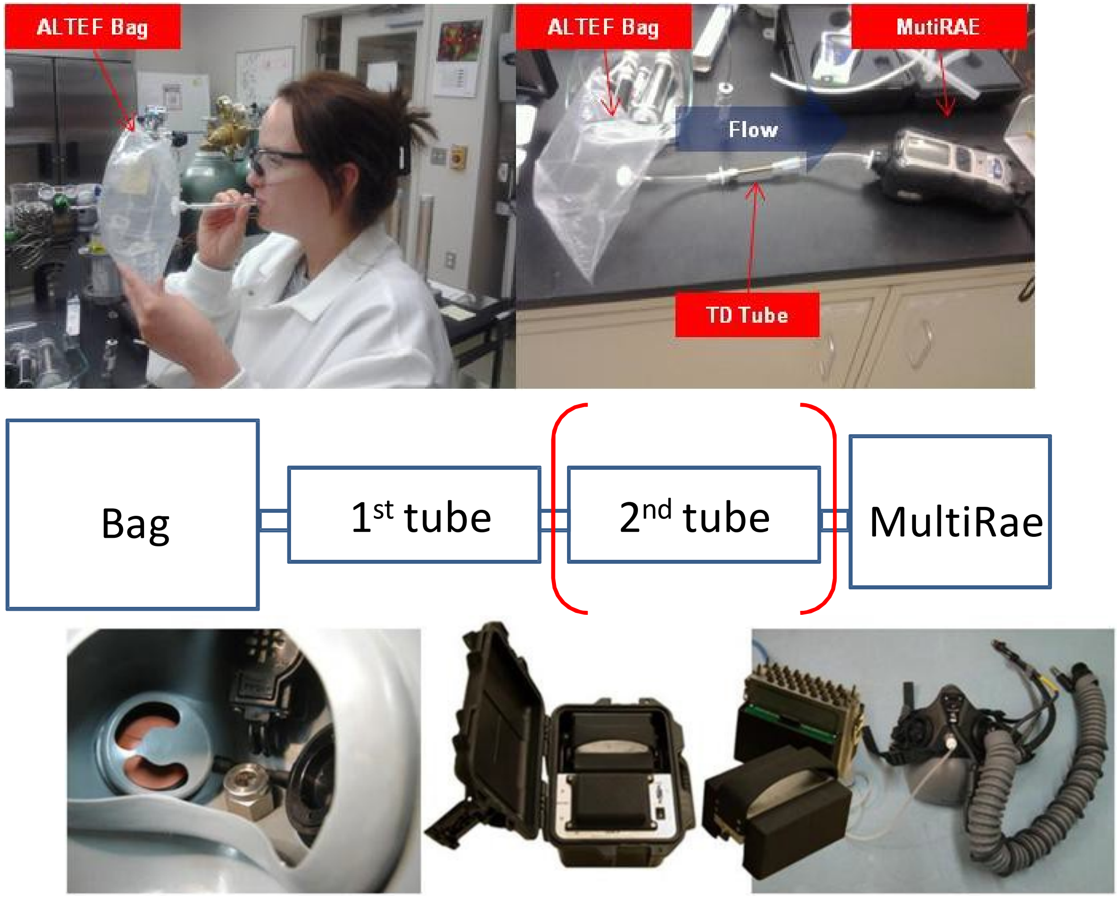


Figure 29: Gas Bag and TD Tube Time Series-Based Breath Collection Methods
Gas bag and TD tube time series based breath collection methods currently utilized by our laboratory and collaborators for on-going breath biomarker discovery projects.

These on-going experiments by our laboratory and collaborators seek to expand the understanding and ability to detect the complex VOC profiles that are likely to be present physiologically, occupationally, and in the environment so that we are better able to monitor and protect personnel in the future. The preceding, developed informatics tools and processing pipeline have created the foundation for these efforts and have provided our group a wonderful opportunity to successfully complete much of the work described.

9.0 REFERENCES

1. Napoli, C., et al., *Microarray analysis: A Novel Research Tool for Cardiovascular Scientists and Physicians.* Heart, 2003. **89**(6): p. 597-604.
2. Liu, B., S. Li, and J. Hu, *Technological Advances in High-Throughput Screening.* American Journal of Pharmacogenomics, 2004. **4**(4): p. 263-276.
3. Sok, J.C., et al., *Tissue-specific Gene Expression of Head and Neck Squamous Cell Carcinoma in Vivo by Complementary DNA Microarray Analysis.* Archives of Otolaryngology—Head & Neck Surgery, 2003. **129**(7): p. 760.
4. Pusztai, L., et al., *Gene Expression Profiles Obtained from fine-Needle Aspirations of Breast Cancer Reliably Identify Routine Prognostic Markers and Reveal Large-Scale Molecular Differences between Estrogen-Negative and Estrogen-Positive Tumors.* Clinical Cancer Research, 2003. **9**(7): p. 2406-2415.
5. Hirabayashi, Y., et al., *Mechanism of benzene-induced hematotoxicity and Leukemogenicity: Current Review with Implication of Microarray Analyses.* Toxicologic Pathology, 2004. **32**(2 suppl): p. 12-16.
6. Smith, M.T., et al., *Use of 'Omic'technologies to Study Humans Exposed to Benzene.* Chemico-Biological Interactions, 2005. **153**: p. 123-127.
7. Wouters, L., et al., *Graphical Exploration of Gene Expression Data: A Comparative Study of Three Multivariate Methods.* Biometrics, 2003. **59**(4): p. 1131-1139.
8. Salter, A.H. and K.C. Nilsson, *Informatics and Multivariate Analysis of Toxicogenomics Data.* Current opinion in drug discovery & development, 2003. **6**(1): p. 117.
9. Jirapech-Umpai, T. and S. Aitken, *Feature Selection and Classification for Microarray Data Analysis: Evolutionary Methods for Identifying Predictive Genes.* BMC Bioinformatics, 2005. **6**(1): p. 148.
10. Zhao, Y., M.-C. Li, and R. Simon, *An Adaptive Method for cDNA Microarray Normalization.* BMC Bioinformatics, 2005. **6**(1): p. 28.
11. Verducci, J.S., et al., *Microarray Analysis of Gene Expression: Considerations in Data Mining and Statistical Treatment.* Physiological genomics, 2006. **25**(3): p. 355-363.
12. Li, C. and W.H. Wong, *Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection.* Proceedings of the National Academy of Sciences, 2001. **98**(1): p. 31-36.
13. Quackenbush, J., *Microarray Data Normalization and Transformation.* Nature Genetics, 2002. **32**: p. 496-501.
14. Peterson, L.E., *Partitioning Large-Sample Microarray-Based Gene Expression Profiles using Principal Components Analysis.* Computer Methods and Programs in Biomedicine, 2003. **70**(2): p. 107-119.
15. Boutros, P.C. and A.B. Okey, *Unsupervised Pattern Recognition: An Introduction to the Whys and Wherefores of Clustering Microarray Data.* Briefings in Bioinformatics, 2005. **6**(4): p. 331-343.
16. Pan, W., *A comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments.* Bioinformatics, 2002. **18**(4): p. 546-554.

17. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance Analysis of Microarrays Applied to the Ionizing Radiation Response.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 5116-5121.

18. Tibshirani, R., et al., *Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression.* Proceedings of the National Academy of Sciences, 2002. **99**(10): p. 6567-6572.

19. Méndez, M.A., et al., *Discriminant Analysis to Evaluate Clustering of Gene Expression Data.* FEBS letters, 2002. **522**(1): p. 24-28.

20. Yong, M., et al., *Parameters Selection in Gene Selection using Gaussian Kernel Support Vector Machines by Genetic Algorithm.* Journal of Zhejiang University Science B, 2005. **6**(10): p. 961-973.

21. Zhang, X., et al., *Recursive SVM Feature Selection and Sample Classification for Mass-Spectrometry and Microarray Data.* BMC bioinformatics, 2006. **7**(1): p. 197.

22. Fu, W.J., R.J. Carroll, and S. Wang, *Estimating Misclassification Error with Small Samples via Bootstrap Cross-Validation.* Bioinformatics, 2005. **21**(9): p. 1979-1986.

23. Molinaro, A.M., R. Simon, and R.M. Pfeiffer, *Prediction Error Estimation: A Comparison of Resampling Methods.* Bioinformatics, 2005. **21**(15): p. 3301-3307.

24. Hua, J., et al., *Noise-Injected Neural Networks Show Promise for Use on small-Sample Expression Data.* BMC Bioinformatics, 2006. **7**(1): p. 274.

25. Huang, T.M. and V. Kecman, *Gene Extraction for Cancer Diagnosis by Support Vector Machines—An Improvement.* Artificial Intelligence in Medicine, 2005. **35**(1): p. 185-194.

26. Berrar, D., I. Bradbury, and W. Dubitzky, *Avoiding Model Selection Bias in Small-Sample Genomic Datasets.* Bioinformatics, 2006. **22**(10): p. 1245-1250.

27. Huggett, J., et al., *Real-Time RT-PCR Normalisation; Strategies and Considerations.* Genes and Immunity, 2005. **6**(4): p. 279-284.

28. Thongboonkerd, V., et al., *Proteomic Identification and Immunolocalization of Increased Renal Calbindin-D28k Expression in OVE26 Diabetic Mice.* Rev Diabet Stud, 2005. **2**(1): p. 19-26.

29. Xiao, Z., et al., *Proteomic Patterns: their Potential for Disease Diagnosis.* Molecular and Cellular Endocrinology, 2005. **230**(1-2): p. 95-106.

30. Ransohoff, D., *Lessons from Controversy: Ovarian Cancer Screening and Serum Proteomics.* Journal of the National Cancer Institute, 2005. **97**(4): p. 315.

31. Zhang, L., et al., *Contribution of Human Alpha-Defensin 1, 2, and 3 to the Anti-HIV-1 Activity of CD8 antiviral Factor.* Science, 2002. **298**(5595): p. 995.

32. Wadsworth, J., et al., *Serum Protein Profiles to Identify Head and Neck Cancer.* Clinical Cancer Research, 2004. **10**(5): p. 1625.

33. Petricoin, E., et al., *Serum Proteomic Patterns for Detection of Prostate Cancer.* Journal of the National Cancer Institute, 2002. **94**(20): p. 1576.

34. Xiao, X., et al., *Development of Proteomic Patterns for Detecting Lung Cancer.* Disease Markers, 2003. **19**(1): p. 33-39.

35. Petricoin III, E., et al., *Use of Proteomic Patterns in Serum to Identify Ovarian Cancer.* The Lancet, 2002. **359**(9306): p. 572-577.

36. BA EZ, L., et al., *Diagnostic Potential of Serum Proteomic Patterns in Prostate Cancer.* The Journal of Urology, 2003. **170**(2): p. 442-446.

37. Koopmann, J., et al., *Serum Diagnosis of Pancreatic Adenocarcinoma using Surface-Enhanced Laser Desorption and Ionization Mass Spectrometry.* Clinical Cancer Research, 2004. **10**(3): p. 860.

38. Uchida, T., et al., *Application of a Novel Protein Biochip Technology for Detection and Identification of Rheumatoid Arthritis Biomarkers in Synovial Fluid.* Journal of Proteome Research, 2002. **1**(6): p. 495-499.

39. Lewczuk, P., et al., *Amyloid [Beta] Peptides in Cerebrospinal Fluid as Profiled with Surface Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: Evidence of Novel Biomarkers in Alzheimer's Disease.* Biological Psychiatry, 2004. **55**(5): p. 524-530.

40. O'Farrell, P., *High Resolution Two-Dimensional Electrophoresis of Proteins.* Journal of Biological Chemistry, 1975. **250**(10): p. 4007.

41. O'Farrell, P., H. Goodman, and P. O'Farrell, *High Resolution Two-Dimensional Electrophoresis of Basic as well as Acidic Proteins.* Cell, 1977. **12**(4): p. 1133-1142.

42. Bjellqvist, B., et al., *Isoelectric Focusing in Immobilized pH Gradients: Principle, Methodology and Some Applications.* Journal of Biochemical and biophysical Methods, 1982. **6**(4): p. 317-339.

43. Görg, A., et al., *The Current State of Two-Dimensional Electrophoresis with Immobilized pH Gradients.* Electrophoresis, 2000. **21**(6): p. 1037-1053.

44. Görg, A., et al., *Two Dimensional Electrophoresis of Proteins in an Immobilized pH 4–12 Gradient.* Electrophoresis, 1998. **19**(8 9): p. 1516-1519.

45. Simonian, M. and E. Betgovargez, *Proteome Analysis of Human Plasma with the ProteomeLab PF 2D System.* Beckman Coulter, Inc. Application Information Bulletin A-1963A, 2003.

46. Billecke, C., et al., *Analysis of Glioma Cell Platinum Response by Metacomparison of Two-Dimensional Chromatographic Proteome Profiles.* Molecular & Cellular Proteomics, 2006. **5**(1): p. 35.

47. Linke, T., A. Ross, and E. Harrison, *Proteomic Analysis of Rat Plasma by Two-Dimensional Liquid Chromatography and Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry.* Journal of Chromatography A, 2006. **1123**(2): p. 160-169.

48. Wall, D., S. Parus, and D. Lubman, *Three-Dimensional Protein Map According to pI, Hydrophobicity and Molecular Mass.* Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 2002. **774**(1): p. 53-58.

49. Wang, Y., et al., *Classification of Cancer Cell Lines using an Automated Two-Dimensional Liquid Mapping Method with Hierarchical Clustering Techniques.* Molecular & Cellular Proteomics, 2006. **5**(1): p. 43.

50. *Ettan DIGE System User Manual 18-1173-17 Edition AA.* Amersham Biosciences Corp. Piscataway N.J. USA.

51. Wilm, M. and M. Mann, *Analytical Properties of the Nanoelectrospray Ion Source.* Anal. Chem, 1996. **68**(1): p. 1-8.

52. Matthias Wilm, A., et al., *Femtomole Sequencing of Proteins from Polyacrylamide Gels by nano-Electrospray Mass Spectrometry.* Nature, 1996. **379**: p. 1.

53. Servais, A., J. Crommen, and M. Fillet, *Capillary Electrophoresis-Mass Spectrometry, an Attractive Tool for Drug Bioanalysis and Biomarker Discovery.* Electrophoresis, 2006. **27**(13): p. 2616-2629.

54. Link, A., et al., *Direct Analysis of Protein Complexes using Mass Spectrometry.* Nature Biotechnology, 1999. **17**(7): p. 676-682.

55. Eng, J., A. McCormack, and J. Yates III, *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.* Journal of the American Society for Mass Spectrometry, 1994. **5**(11): p. 976-989.

56. Washburn, M., D. Wolters, and J. Yates, *Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology.* Nature Biotechnology, 2001. **19**(3): p. 242-247.

57. Yates, J., et al., *Automated Protein Identification using Microcolumn Liquid Chromatography-Tandem Mass Spectrometry.* Methods in Molecular Biology-Clifton then Totowa-, 1999. **112**: p. 553-570.

58. Yates III, J., J. Eng, and A. McCormack, *Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases.* Analytical Chemistry, 1995. **67**(18): p. 3202-3210.

59. Peng, J., et al., *Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC- MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome.* Journal of Proteome Research, 2003. **2**(1): p. 43-50.

60. Nicholson, J.K., J.C. Lindon, and E. Holmes, *'Metabonomics': Understanding the Metabolic Responses of Living Systems to Pathophysiological Stimuli via Multivariate Statistical Analysis of Biological NMR Spectroscopic Data.* Xenobiotica, 1999. **29**(11): p. 1181-9.

61. Pham-Tuan, H., et al., *Method Development in High-Performance Liquid Chromatography for High-Throughput Profiling and Metabonomic Studies of Biofluid Samples.* Journal of Chromatography B, 2003. **789**(2): p. 283-301.

62. Lindon, J., E. Holmes, and J. Nicholson, *Metabonomics Techniques and Applications to Pharmaceutical Research & Development.* Pharmaceutical Research, 2006. **23**(6): p. 1075-1088.

63. Robertson, D., *Metabonomics in Toxicology: A Review.* Toxicological Sciences, 2005. **85**(2): p. 809-822.

64. Nicholson, J., et al., *Metabonomics: A Platform for Studying Drug Toxicity and Gene Function.* Nature Reviews Drug Discovery, 2002. **1**(2): p. 153-161.

65. Reo, N., *NMR-Based Metabolomics.* Drug and Chemical Toxicology, 2002. **25**(4): p. 375-382.

66. Dunn, W. and D. Ellis, *Metabolomics: Current Analytical Platforms and Methodologies.* Trends in Analytical Chemistry, 2005. **24**(4): p. 285-294.

67. Wang, Y., et al., *Metabonomic Investigations in Mice Infected with Schistosoma Mansoni: An Approach for Biomarker Identification.* Proceedings of the National Academy of Sciences, 2004. **101**(34): p. 12676-12681.

68. Zlatkis, A., R.S. Brazell, and C.F. Poole, *The Role of Organic Volatile Profiles in Clinical Diagnosis.* Clin Chem, 1981. **27**(6): p. 789-97.

69. Mace, J.W., et al., *The child with an Unusual Odor. A Clinical Resume.* Clin Pediatr (Phila), 1976. **15**(1): p. 57-62.

70. *Body Odor and Metabolic Defects.* Nutr Rev, 1968. **26**(4): p. 107-11.

71. Jellum, E., O. Stokke, and L. Eldjarn, *Application of Gas Chromatography, Mass Spectrometry, and Computer Methods in Clinical Biochemistry.* Anal Chem, 1973. **46**(7): p. 1099-106.

72. Burke, D.G., et al., *Profiles of Urinary Volatiles from Metabolic Disorders Characterized by Unusual Odors.* Clin Chem, 1983. **29**(10): p. 1834-8.

73. Zlatkis, A. and H.M. Liebich, *Profile of volatile Metabolites in Human Urine.* Clin Chem, 1971. **17**(7): p. 592-4.

74. Zlatkis, A., et al., *Profile of volatile Metabolites in Urine by Gas Chromatography-Mass Spectrometry.* Anal Chem, 1973. **45**(4): p. 763-7.

75. Novotny, M., et al., *Chemical Studies of the Primer Mouse Pheromones.* Chemical Signals in Vertebrates and Aquatic Invertebrates. D. Muller-Schwarze, RM Silverstein,(eds.). Plenum Press, New York, 1980: p. 377-390.

76. Harvey, S., B. Jemiolo, and M. Novotny, *Pattern of Volatile Compounds in Dominant and Subordinate Male Mouse Urine.* Journal of Chemical Ecology, 1989. **15**(7): p. 2061-2072.

77. Zlatkis, A., et al., *Profiles of Organic Volatiles in Biological Fluids as an Aid to the Diagnosis Of Disease.* Analyst, 1981. **106**(1260): p. 352-60.

78. Winberry, W.T., N.T. Murphy, and R. Riggan, *Compendium of Methods for the Determination of Toxic Organic Compounds in Ambient Air.* 1992.

79. Eller, P.M. and M.E. Cassinelli, *NIOSH Manual of Analytical Methods*. 1994: DIANE Publishing.

80. Seethapathy, S., T. Górecki, and X. Li, *Passive Sampling in Environmental Analysis.* Journal of Chromatography A, 2008. **1184**(1): p. 234-253.

81. Harper, M. and L.V. Guild, *Experience in the Use of the NIOSH Diffusive Sampler Evaluation Protocol.* American Industrial Hygiene Association Journal, 1996. **57**(12): p. 1115-1123.

82. Belardi, R. and J. Pawliszyn, *The Application of Chemically Modified Fused Silica Fibres in the Extraction of Organics from Water Matrix Samples and their Transfer to Capillary Columns.* Water Pollut Res J Can, 1989. **24**: p. 179-82.

83. Arthur, C. and J. Pawliszyn, *Solid Phase Microextraction with Thermal Desorption using Fused Silica Optical Fibers.* Analytical Chemistry (Washington, DC), 1990. **62**(19): p. 2145-2148.

84. Wercinski, S., *Solid Phase Microextraction: A Practical Guide*. 1999: CRC.

85. Pawliszyn, J., *Solid Phase Microextraction: Theory And Practice*. 1997: Vch Verlagsgesellschaft Mbh.

86. Mani, V. and J. Pawliszyn, *In Applications of Solid Phase Microextraction.* Cambridge: Royal Society Chemistry, 1999.

87. Supelco, *Solid Phase Micrextraction of Volatile Compounds.*

88. Pawliszyn, J., *Solid Phase Microextraction*, in *Headspace Analysis of Foods and Flavors*. 2001, Springer. p. 73-87.

89. Grigsby, C.C., et al. *Differential Profiling of Volatile Organic Compound Biomarker Signatures Utilizing a Logical Statistical Filter-Set and Novel Hybrid Evolutionary Classifiers*. in *Proc. of SPIE Vol*. 2012.

90. Zhang, Z. and J. Pawliszyn, *Headspace Solid-Phase Microextraction.* Analytical Chemistry, 1993. **65**(14): p. 1843-1852.

91. Van Berkel, J., et al., *A Profile of Volatile Organic Compounds in Breath Discriminates COPD Patients from Controls.* Respiratory Medicine, 2010. **104**(4): p. 557-563.

92. Van Berkel, J., et al., *Development of Accurate Classification Method Based on the Analysis of Volatile Organic Compounds from Human Exhaled Air.* Journal of Chromatography B, 2008. **861**(1): p. 101-107.

93. Silverstein, R.M., *Spectrometric Identification of Organic Compounds*. 2005 ed. 2005, Hoboken: Wiley. 502.

94. Baran, R., et al., *MathDAMP: A Package for Differential Analysis of Metabolite Profiles.* BMC Bioinformatics, 2006. **7**(1): p. 530.

95. Broeckling, C.D., et al., *MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics.* Anal Chem, 2006. **78**(13): p. 4334-41.

96. Bunk, B., et al., *MetaQuant: A Tool for the Automatic Quantification of GC/MS-Based Metabolome Data.* Bioinformatics, 2006. **22**(23): p. 2962-5.

97. Katajamaa, M. and M. Oresic, *Processing Methods for Differential Analysis of LC/MS Profile Data.* BMC Bioinformatics, 2005. **6**: p. 179.

98. Luedemann, A., et al., *TagFinder for the Quantitative Analysis of Gas Chromatography--Mass Spectrometry (GC-MS)-Based Metabolite Profiling Experiments.* Bioinformatics, 2008. **24**(5): p. 732-7.

99. Smith, C.A., et al., *XCMS: Processing Mass Spectrometry Data for Metabolite Profiling using Nonlinear Peak Alignment, Matching, and Identification.* Anal Chem, 2006. **78**(3): p. 779-87.

100. Grigsby, C., et al., *Metabolite Differentiation and Discovery Lab (MeDDL): A New Tool for Biomarker Discovery and Mass Spectral Visualization.* Analytical Chemistry, 2010. **82**(11): p. 4386-4395.

101. Van Vleet, T. and R. Schnellmann. *Toxic Nephropathy: Environmental Chemicals*. 2003. Elsevier.

102. Soto, A., et al., *D-Serine Exposure Resulted in Gene Expression Changes Indicative of Activation of Fibrogenic Pathways and Down-Regulation of Energy Metabolism and Oxidative Stress Response.* Toxicology, 2008. **243**(1-2): p. 177-92.

103. Carone, F. and C. Ganote, *D-Serine Nephrotoxicity. The Nature of Proteinuria, Glucosuria, and Aminoaciduria in Acute Tubular Necrosis.* Archives of Pathology, 1975. **99**(12): p. 658.

104. Ganote, C., D. Peterson, and F. Carone, *The Nature of D-serine-Induced Nephrotoxicity.* The American Journal of Pathology, 1974. **77**(2): p. 269.

105. Pilone, M.S., *D-Amino Acid Oxidase: New Findings.* Cell Mol Life Sci, 2000. **57**(12): p. 1732-47.

106. Serra, J., *Image Analysis and Mathematical Morphology. 11: Theoretical Advances*. 1988, London: Academic Press.

107. Buschmann, F., *Pattern-Oriented Software Architecture : A System of Patterns*. 1996, Chichester ; New York: Wiley. xvi, 457 p.

108. Rosenfeld, A., R. Hummel, and S. Zucker, *Scene labeling by Relaxation Operations.* IEEE Transactions on Systems, Man and Cybernetics, 1976. **6**(6): p. 420-433.

109. Wu, Q., *A Correlation-Relaxation-Labeling Framework for Computing Opticalflow-Template Matching from a New Perspective.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995. **17**(9): p. 843-853.

110. Richmond, B., et al., *Temporal Encoding of Two-Dimensional Patterns by single Units in Primate Inferior Temporal Cortex. I. Response Characteristics.* Journal of Neurophysiology, 1987. **57**(1): p. 132.

111. Peters, L.L., et al., *The Mouse as a Model for Human Biology: a Resource Guide for Complex Trait Analysis.* Nature Reviews Genetics, 2007. **8**(1): p. 58-69.

112. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*. 2nd ed. 2001, New York: Wiley. xx, 654 p.

113. Huang, C.L. and C.J. Wang, *A GA-Based Feature Selection and Parameters Optimization for Support Vector Machines.* Expert Systems with applications, 2006. **31**(2): p. 231-240.

114. Lu, J., T. Zhao, and Y. Zhang, *Feature Selection Based-on Genetic Algorithm for Image Annotation.* Knowledge-Based Systems, 2008. **21**(8): p. 887-891.

115. Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989: Addison-wesley.

116. Russell, S.J. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2010: Prentice Hall.

117. Kwak, J., et al., *Differential Binding between Volatile Ligands and Major Urinary Proteins Due to Genetic Variation in Mice.* Physiology & Behavior, 2012. **107**(1): p. 112-120.

118. Kwak, J., et al., *Changes in Volatile Compounds of Mouse Urine as it Ages: Their interactions with Water and Urinary Proteins.* Physiology & Behavior, 2013.

119. Kwak, J., et al., *Changes in Volatile Compounds of Human Urine as it Ages: Their Interaction with Water.* Journal of Chromatography B, 2013.

120. McClenny, W.A. and M. Colón, *Measurement of Volatile Organic Compounds by the US Environmental Protection Agency Compendium Method TO-17: Evaluation of Performance Criteria.* Journal of Chromatography A, 1998. **813**(1): p. 101-111.

121. Baran, R., et al., *MathDAMP: A Package for Differential Analysis of Metabolite Profiles.* BMC Bioinformatics, 2006. **7**: p. 530.

122. Pleil, J.D. and A.B. Lindstrom, *Exhaled Human Breath Measurement Method for Assessing Exposure to Halogenated Volatile Organic Compounds.* Clinical Chemistry, 1997. **43**(5): p. 723-730.

10.0 ACRONYMS

2D-DIGE	2D Difference in-Gel Electrophoresis
2D-SDS PAGE	Sodium dodecyl sulfate – Polyacrylamide gel Electrophoresis
AFB	Air Force Base
ANOVA	Analysis of Variance
BFB	Bromofluorobenzene
CCV	Calibration Check Verification
CO	Carbon Monoxide
CO2	Carbon Dioxide
D-AAO	D-Aminoacid Oxidase
DNA	Deoxyribonucleic Acid
DoD	Department of Defense
EFS	Environmental, Food and Safety
EI	Electron Impact
EPA	Environmental Protection Agency
ESI	Electro Spray Ionization
FSOT	Fused Silica Open Tubular
GA	Genetic Algorithm
GC/MS	Gas Chromatography Mass Spectometry
IEF	Isoelectric Focusing
IP	Intraperitoneal
JBER	Joint Base Elmendorf Richardson
KNN	K-nearest neighbor
LC/MS	Liquid Chromatography - Mass Spectometry
LCS	Laboratory Control Sample
LOD	Limit of Detection
LOO	Leave-One-Out
m/z	Mass Over Charge
MALDI	Matrix Assisted Laser Desorption Ionization
MeDDL	Metabolite Differentiation and Discovery Lab

MS	Mass Spectrometry
MUPs	Major Urinary Proteins
MVC	Model-View-Controller
NaNs	Not a Number's
netCDF	Network Common Data Form
NIST	National Institute of Standards and Technologies
NMR	Nuclear Magnetic Resonance
O2	Oxygen
OBOGS	On-Board Oxygen Generating System
PCA	Principal Component Analysis
PDMS	Polydimethylsiloxane
PMT	Post Translational Modifications
PTFE	Polytetrafluoroethylene
QToF	Quadrupole Time-of-Flight
ROC	Receiver Operating Characteristic
RP	Reversed Phase
RP-HPLC	Reverse-Phase High Pressure Liquid Chromatography
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SCX	Stron Cation Exchange
SELDI	Surface Extraction Laser Desorption/Ionization
SPE	Solid Phase Extraction
SPME	Solid phase Microextraction
TD	Thermal Desorption
TIC	Total Ion Current
TOF	Time of Flight
UPLC	Ultra Pure Liquid Chromatography
USAF	United States Air Force
VOC	Volatile Organic Compound

APPENDIX A - Software User Guide

Please note: a complete, online version of the MeDDL user guide is available at: http://meddl.cs.wright.edu

PEAK DATA

The Peak Data tab contains the main data table and will also serve as the location of any external data files that you load. For more information about loading external data, see the entry on External Data.

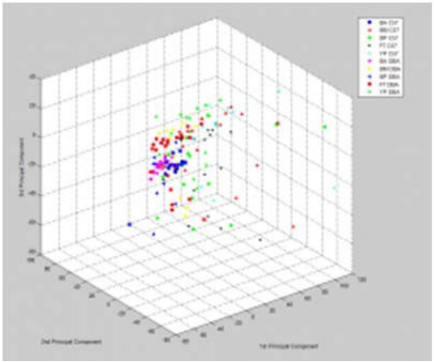
MAIN BUFFER

This tab contains the main table that displays a summary of all of the registered peaks in a given data set. This table allows you to preform a variety of actions on any number of selected peaks. At the bottom of the table you will see a row of buttons that you will use to further analyze the data set. These include: Line Plot, Box Plot, PCA, Correlate, Data Table, and Copy. To find a peak in the Heat Map, simply right-click on it and a green circle will surround that peak. You also select multiple peaks and right-click on the selection to find multiple peaks in the Heat Map. Right-clicking once will only highlight peaks while you hold the right mouse button down. To lock highlighting around a peak, right-click twice.

Line and Box Plot

These plots will display a respective plot for each of the peaks selected. Be advised that you do not want to select too many peaks at one time because each plot will be displayed in a single figure. These buttons have two settings associated with them: Plot Display Group and Plot Color Group. These settings are defined in the Plot Filter tab. The Plot Display Group is the attribute that you wish to plot and the Plot Color Group is the attribute that you wish to use to split up the plots by. For example, if your data set has the attributes Strain (containing three possiblities: AKR, B6, and BALB_B) and MUP Protein (containing two possibilities: Intact and Denatured), then you can view a plot of the Strain versus MUP Protein by setting your Plot Display Group as Strain and Plot Color Group as MUP Protein. For a box plot this means that you will have two separate plots because you have two possible tags under attribute MUP Protein (Intact and Denatured) with each containing three boxes along the x-axis (one for each of the tags under the attribute Strain). For a Line Plot this means that you will have one plot that has each Strain along the x-axis with each MUP Protein as a different marker.

The Line Plot is unique in that it provides access to the raw spectra data for a particular peak. You may double click on any of the plotted points in a line plot. This will open a summary window. Click the Explore button to view a peak in the raw spectra. The window can be dismissed using the Done button.



PCA

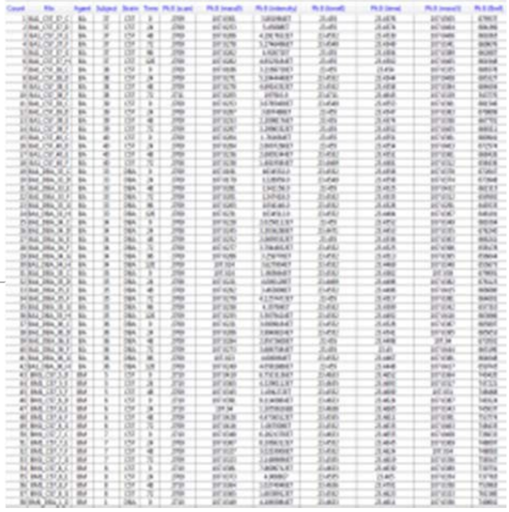
The PCA button will generate a 3D plot of the prinple components in the data set as well as generate a variance pareto graph. The PCA uses the last two attributes in a given data set to generate labels. For example, if your experiment.ini file contains the following attributes: [File Subject Strain MUP_Protein], then you will have a PCA labeled with all possible combinations of the tags for the last two labels. Again, if the attribute Strain has three tags: AKR, B6, and BALB_b, and the attribute MUP_Protein contains two tags: Intact and Denatured, the PCA will have the labels: Denatured-AKR, Denatured-B6, Denatured-BALB_b, Intact-AKR, Intact-B6, and Intact-BALB_b. Note that you do not have to have any peak(s) selected to preform a PCA; the PCA is based off of the entire data set.

Correlate

The Correlate button will take a single selected peak and update the Rank column to contain a score for the correlation of the peaks in respect to the peak selected. The peak that was used to correlate from will have a score of 1.0. The correlate function will get the all of the intensities for a selected peak and pass them into the Matlab corr function. This will return a vector that is the same size as the main feature vector. This vector will contain Pearson product-moment correlation coefficients for each of the comparisons against the selected feature, which will be from -1.0 to 1.0. See the corr function for more information.

Data Table

The Data Table button will display a table for the selected peaks that shows detailed information about the peak(s) across all of the samples/files. This detailed information includes data such as mass, intensity, and time.



GROUP FILTER

The group filter tab allows users to create partitions and groups. For an explanation on partitions and groups, see the Data Management page. After users have created at least one group, they can begin to add groups to it. The All partition is a default partition that contains all of the files.

CREATING A PARTITION

To create a partition, simply type a label in the partition text box and click the 'Save Partition' button. After you create a partition it becomes the active partition on the group filter tab. This means that any groups you save will be saved within it. To change the active partition, change the selected partition in the partition combo box.

CREATING A GROUP

To create a group, you will need to have created a partition (see above). Once you have a partition, toggle the check boxes under each tag of your data that you wish to include in the group. Note that if only one field is selected under a given tag, MeDDL automatically inserts that field into the groups label. Once the check boxes are set for the group, click the 'Save Group' button. Remember that this group will be added to the active partition (whichever partition is selected in the partition combo box). You cannot save a group in the All partition. The groups within a partition must be mutually exclusive.

OUTLIER FILTER

Summary

Removes peaks that are outside of the range of the n number of standard deviations from the mean. This filter inhibits the use of the PCA functionality. Users specify the number of standard deviations to accept. Data outside of this range will not be shown throughout MeDDL.

Pseudocode

```
mu = mean(intensities)
stand = std(intensities)


for i = 1 : number of files
  for j = 1 : number of features
    if(intensities(i, j) < (mu(j) - (factor * stand(j))) || outlierInten(i,j) > (mu(j) + (factor
* stand(j))))
      intensities(i, j) = NaN
    end
  end
end
```

Factor is the user-defined parameter representing the number of standard deviations from the mean to accept.

Theory

The means and standard deviations are calculated across all of the files for each peak. Each value is checked to ensure that it is within the user-defined number of standard deviations from the mean.The figure on the left is was generated using the original data and the figure on the right was generated after the outliers were removed. A setting of 1 standard deviation was used in this case.

TOTAL ION CURRENT (TIC) NORMALIZATION

Summary

Total Ion Current Normalization involves summing the intensities contained in each file of the spectra. The user selects which file to use as the seed file. The seed file will be used to build the normalization ratio.

Pseudocode

```
function computeTICs
  TICs = []


  for i = 1 : number of features
    TICs = [TICs; sum(intensities(i, :))]
  end
  return TICs
```

```
end


TICs = computeTICs
normailizationRatio =  seedTic / TICs


for i = 1 : number of TICs
  intensity(i, :) = normailizationRatio(i) * intensity(i, :);
end
```

Theory

As you can see in the above pseudocode, a TIC score is calculated for each file in the data by summing the intensities in said file. This array of TIC scores is returned to a different function that applies the normalization to the data. The seed TIC is simply the TIC score from the file that the user selects in a drop down box. The TIC vector becomes the right division of the seed TIC and TIC vector. That is, each score in the TIC vector becomes the seed TIC divided by the original TIC score. Note that this makes the normalization ratio for the seed file to be 1; the intensities in the seed file will remain unchanged. The normalization ratio is applied to all of the intensities within each file.

OLYMPIC AVERAGE NORMALIZATION

Summary

The upper and lower percentiles are calculated for all of the features across each file. These percentiles are then used to normalize the intensity values in each file. Intensity values that become negative as a result of the normalization are replaced with NaNs.

Enter the upper percentage and lower percentage in the respective text boxes. A setting of 90 and 10 (as shown at the top of this page) will calculate the 10th and 90th percentiles. Thus, the data will be normalized using the upper and lower 10% of the data.

Pseudocode

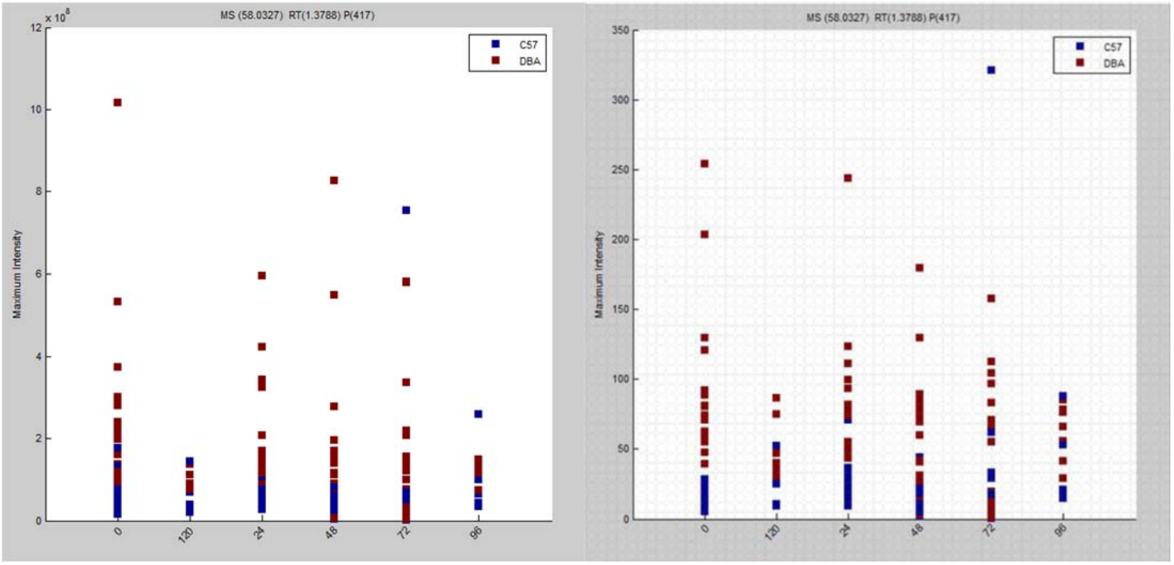
```
lower_percentile = prctile(intensity, lower, 2)
upper percentile = prctile(intensity, upper, 2)


for i = 1 : number of files
  normalizedIntensity(i,:) = (intensity(i,:) - lower_percentile(i)) / (upper_percentile(i) - lower_percentile(i));
end
```

It is understood that the variables lower_percentile and upper_percentile can be calculated in one line. For example, the following command would return an array containing the upper and lower percentiles: prctile(intensity, [lower, upper], 2) Here we break them up for simplicity.

Theory

The upper and lower percentiles are calculated across all of the files using the user defined values. This gives us two vectors containing percentiles, each with a size of n by 1, where n is the number of features. These percentile vectors are used to normalize the intensity values for across each file. The difference of the original intensities in a file and the lower percentile value is divided by the difference of the upper and lower percentiles. This technique will scale down the intensities by several orders of magnitude.

Consider the following Line Plots for the same peak. The left plot shows the original data and the right plot shows the data that has been normalized using the Olympic Average Normalization technique. The upper bound of the y-axis for the unnormalized data is 12 x 10^8 and the upper bound of the y-axis for the normalized data is 350.



DATA FILTER

The Data Filter tab contains all of the filtering methods that allow you to down select your data so that you can focus on significant data. These filters create objects called peak sets. A peak set can be viewed in the Visualization Tab or classified in the Machine Learning Tab.

P-VALUE FILTER

Summary

The P-Value filter checks for statistical significance using ANOVA, calculating a p-value for each peak in study. After the p-values are calculated for each peak, a new feature vector is produced using the user defined p-value threshold (confidence bounds). This feature vector is saved as a peak set.

Theory

There are two options for the P-Value filter:

- N-Way: performs an n-way ANOVA using the Matlab function anovan which accepts all of the groups in the active partition. The features are down selected to be only from files that are included in the active partition. The down selected features are sent to anovan. Thus, features from each file are compared to features from the remaining files.

- Pairwise: performs a pairwise ANOVA using the Matlab function anovan iteratively generating all possible combinations of the groups in the active partition. The indices of the comparisons relate to the indices of all nonzero elements in a strictly upper triangular matrix, i.e. 1 to 2, 1 to 3, 2 to 3, etc. Groups are not compared to themselves.

TIME BINNING FILTER

Summary

The Time Binning Filter allows you to reduce GC-MS data that contains fragments caused by electron impact.

Pseudocode

```
while sum(maxIntensities > 0)
  maxIndex = max(maxIntensities);

  intensityLogical(maxIndex) = 2;
  maxIntensities (maxIndex) = 0;

  timeCenter = timeAvg(maxIndex);
  upper = timeCenter + (deltaT / 2);
  lower = timeCenter - (deltaT / 2);

  index = find( intensityLogical == 1 AND timeAvg >= lower AND timeAvg <
upper);
  intensityLogical(index) = 0;
  maxIntensities(index) = 0;
end
```

Theory

The Time Binning Filter drills down on the data set starting with the most intense peak of the data set. After the maximum peak is found, the peaks with times within $\pm{}^{t}/_{2}$ from the time t_0 are removed from the data, where t is the user defined time parameter and t_0 is the time of the most intense peak in the data set. This process is repeated until all peaks are processed.

The intensities in the 'maxIntensity' array are set to 0 as the data is processed to indicate that they are of no further interest. The array 'intensityLogical' is a pseudo-boolean array. 0's indicate features of no interest, 1's indicate potential interest, and 2's indicate an intensity index that is the maximum within the time window.

FOLD CHANGE FILTER

Summary

The Fold Change Filter takes the average of each feature across all of the files that are included in a partition. This produces a group average vector for each group. The average vectors are divided by each other using right division, left division, a combination of both to produce a fold value vector (containing a fold value for each feature). Features with fold values that meet or exceed the user-defined threshold are kept.

Pseudocode

```
 left  = B / A;
right = A / B;

if typeChange == absolute
    result = left >= foldChange OR right >= foldChange;
    value = max(left, right);
else if typeChange == positive
    result = left >= foldChange;
else
    result = right >= foldChange;
```

Theory

The components of the resulting fold vector are compared to the threshold that the user defines and only those features that meet or exceed the threshold pass the filter. If there are more than two groups in the active partition, then pairwise comparisons of all of the groups are generated. As in the ANOVA filter, the indices of the comparisons relate to the indices of all nonzero elements in a strictly upper triangular matrix. The disjunction of all of the feature vectors from each of the pairwise comparisons is generated: features must meet the fold criteria in at least one fold comparison.

The Fold Change Filter has the options to accept only positive or only negative folds that fulfill the threshold, where a positive fold is defined as the right division of A and B and a negative fold is defined as the left division of A and B. Note: we define left division as B(i, j) / A(i, j) and right division as A(i, j) / B(i, j). See MathWorks for more information. The default setting is absolute, which accepts a feature if either the left division or right division is greater than the set threshold. In the case of the absolute setting, the resulting fold value vector will contain the maximum fold change from either the left division or right division of the average matrices.

- Absolute: Calculate the fold using both the left and right division of the pair of matrices. The fold value will be the maximum for that feature from either matrix.
- Positive: Calculate the fold using the right division of the average vectors.
- Negative: Calculate the fold using the left division of the average vectors.

GROUP INTENSITY FILTER

Summary

The Group Intensity filter peaks that have an average intensity greater than the set intensity.

The Strict setting specifies whether or not all of the groups in the partition must pass the filter. If Strict is used the conjunction of the feature vectors is taken to produce the final peak set. Otherwise, the disjunction of the feature vectors is taken to produce the peak set.

Pseudocode

```
for all samples in the group on the filter tab

  if strict is selected
    remaining peaks = all the peaks AND peaks with an average intensity
greater than the given absolute intensity

    otherwise
    remaining peaks = all the peaks ORpeaks with an average intensity greater
than the given absolute intensity
```

GROUP DISTRIBUTION

Summary

Group Distribution filters out peaks that do not fall within the given standard deviation in the group it is contained in. If the peak is more Standard Deviations away then the given value the peak is excluded from the final peak set.

The Strict setting specifies whether or not all of the groups in the partition must pass the filter. If Strict is used the conjunction of the feature vectors is taken to produce the final peak set. Otherwise, the disjunction of the feature vectors is taken to produce the peak set.

Pseudocode

```
for all samples in the group
    upper limit allowed = group average plus the standard deviations
    lower limit allowed = group average minus standard deviations

   if strict is selected on the filter tab
     filtered peaks = filtered peaks AND intensities that are less than
     the upper limit allowed    AND intensities that are less than the
     lower limit allowed

   otherwise
     filtered peaks = filtered peaks OR intensities that are less than
     the upper limit   allowed OR intensities that are less than the
     lower limit allowed
```

GROUP SEPARATION FILTER

Summary

This filter compares all possible pair-wise combinations of the groups. The means of both groups are calculated, then the difference is found between the mean of every peak in each group. If the separation of that peak is larger than the given separation value then the peak passes the filter and depending on strict or loose may or may not be included in the final peak set.

The Strict setting specifies whether or not all of the groups in the partition must pass the filter. If Strict is used the conjunction of the feature vectors is taken to produce the final peak set. Otherwise, the disjunction of the feature vectors is taken to produce the peak set.

Pseudocode

```
for all possible combinations
     setA = means for set A
     setB = means for set B

     if strict is selected on the filter tab
          filteredPeaks = filteredPeaks AND peaks that pass |setA - setB| >
          separationFactor

     otherwise
          filteredPeaks = filteredPeaks OR peaks that pass |setA - setB| >
          separationFactor
```

RATIO FILTER

Summary

There are two options for this filter, top heavy and singular. Top heavy squares the numerator of the equation used to find the ratios and singular does not. The filter calculates ratios based on a formula and returns the given amount of peaks with the highest ratios. The ratio is calculated by first generating all pair-wise combinations of the groups. Then the Standard Deviation and average intensities are calculated for each of those groups. Then the ratio is calculated by adding all the comparison combinations ratios together, ratios are calculate by the absolute value of the average of set one minus the average of set 2, intensity that is, divided by the standard deviation of set 1 times the standard deviation of set 2. The statistical significance would be finding peaks that are tightly formed and separated with respects to intensity.

Top Heavy:

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\left(\mu_i - \mu_j\right)^2}{\sigma_i \sigma_j}$$

Otherwise:

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\left|\mu_i - \mu_j\right|}{\sigma_i \sigma_j}$$

Pseudocode

```
For all pairwise comparisons of the groups
  setAMu = means of set A
  setAStandardD = standard deviation of set A
  setBMu = means of set B
  setBStandardD = standard deviation of set B

    if top heavy is selected on the filter tab
    ratios = ratios + (setAMu - setBMu)^2 / (setAStandardD * setBStandardD)
    otherwise
    ratios = ratios + abs((setAMu - setBMu)) / (setAStandardD *
setBStandardD)

pick the n highest ratio peaks
```

MASS-TIME EXCLUSION WINDOW

Summary

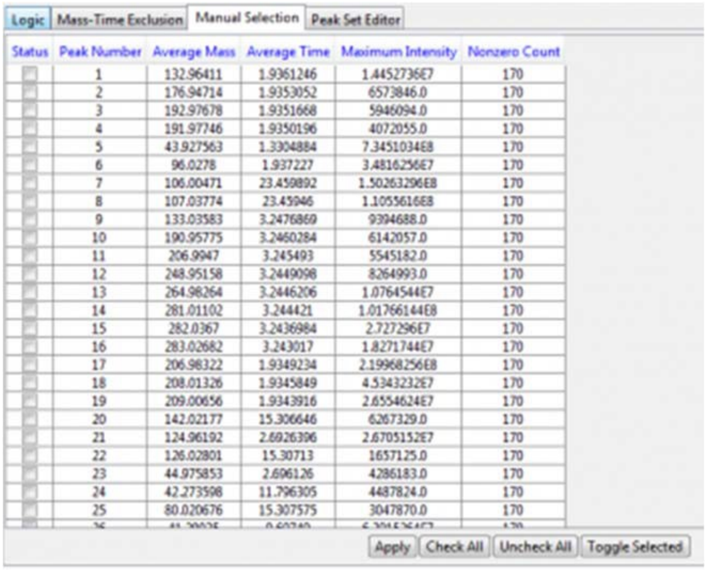
The mass-time exclusion filter allows you to specify an exclusion window of mass and time. This can be done in two ways: manually or with the heat map. To manually specify an exclusion window, enter the upper and lower bounds for both mass and time, and then click the Add button. This will add the window to the table on the right. You may now enter another window and add it to the table or click the Apply button to generate a peak set that excludes all of the windows listed in the table. To specify a mass-time exclusion window using the heat map, simply left click and drag a selection box around the peaks that you wish to exclude. The coordinates (upper and lower bounds) of the selected region will be displayed in the text fields for the filter. You may click the Add button to add the window to the table or modify the upper and lower bounds by typing them in manually or by drawing a new selection on the heat map.

MANUAL SELECTION

Summary

The Manual Selection Filter allows you to hand select peaks to be included in a feature vector. This can be used to include peaks of interest or to create a mask of peaks that are extraneous to your study. If the peaks are of interest, then you can view them as or logically OR them with another peak set. If they are extraneous, you can negate the feature vector and then AND the resulting peak set with any other peak set.

Use the Select All/Deselect All to toggle all of the check boxes on or off, respectively. Use the toggle button to



Status	Peak Number	Average Mass	Average Time	Maximum Intensity	Nonzero Count
□	1	132.96411	1.9361246	1.4452736E7	170
□	2	176.94714	1.9353052	6573846.0	170
□	3	192.97678	1.9351668	5946094.0	170
□	4	191.97746	1.9350196	4072055.0	170
□	5	43.927563	1.3304884	7.3451034E8	170
□	6	96.0278	1.937227	3.4816256E7	170
□	7	106.00471	23.459892	1.50263296E8	170
□	8	107.03774	23.45946	1.1055616E8	170
□	9	133.03583	3.2476869	9394688.0	170
□	10	190.95775	3.2460284	6142057.0	170
□	11	206.9947	3.245493	5545182.0	170
□	12	248.95158	3.2449098	8264993.0	170
□	13	264.98264	3.2446206	1.0764544E7	170
□	14	281.01102	3.244421	1.01766144E8	170
□	15	282.0367	3.2436984	2.727296E7	170
□	16	283.02682	3.243017	1.8271744E7	170
□	17	206.98322	1.9349234	2.19968256E8	170
□	18	208.01326	1.9345849	4.5343232E7	170
□	19	209.00656	1.9343916	2.6554624E7	170
□	20	142.02177	15.306646	6267329.0	170
□	21	124.96192	2.6926396	2.6705152E7	170
□	22	126.02801	15.30713	1657125.0	170
□	23	44.975853	2.696126	4286183.0	170
□	24	42.273598	11.796305	4487824.0	170
□	25	80.020676	15.307575	3047870.0	170
□	26	[illegible]	[illegible]	[illegible]	170

invert the state of all of the check boxes. Finally, the Apply button will generate a peak set from the selected peaks. All of the columns in the table support sorting, including the check box column. Left click on the column header to sort in the features in decreasing order. Click the header again to sort in nondecreasing order.

PEAK SET EDITOR

The Peak Set Editor allows you to manage peak sets. Use the delete button to delete a selection of peak sets. The Ctrl key allows you to select more than one peak set. The Delete All button will delete all of the peak sets.

MACHINE LEARNING

MeDDL offers users with several types of classification methods. These methods use Matlab and Waikato Environment for Knowledge Analysis classifiers. Each of these allow users to classify data that is internal or external to MeDDL. The internal data classification allows users to classify the peak sets that they have created using the tool. The external data classification is currently designed to process a comma separated value file (.CSV). All classification methods support classifying intensities or ratios of intensities.

MATLAB TREE CLASSIFIER

- Greedy, sequential feature selection algorithm given by the sequentialfs Matlab function.
- Uses a forward selection method (need to test the results of a reverse selection method).
- If data has been filtered using the fold filter, the data will be sorted so that the fold values are in descending order.
- User specifies the number of features to select and the number of folds for cross validation.
- Allows the user to visualize the results by displaying a tree for the desired peak(s) or a bar graph of the peak(s). The tree is often a simple decision stump, but can have more depth to it. The bar graphs list the files along the x-axis and show the intensities for each file along the y-axis.

Pseudocode:

```
Initialize features vector to zeros while number of features is less than the requested number

Call the sequentialfs function

**The sequentialfs function uses the functions ClassificationTree.fit and predict to produce a classification score.

Add the newly selected features to the features vector

end
```

As you can see, the sequentialfs method is very important in the classification method. It picks what features are important by adding them and seeing if classification performance is improved upon. Classification performance is generated by using the ClassificationTree.fit method. This method produces a tree. The tree is then used to classify the data using the predict method. The classifier's labels are compared to the actual labels to produce a score.

MATLAB SVM CLASSIFIER

- Same greedy, sequential forward selection algorithm as the tree classifier (sequentialfs).
- Uses the svmtrain Matlab function.
- Allows users to display a bar graph of the selected peak(s) as described above.

The code for the SVM classifier is fairly similar to that of the Tree classifier. The primary difference is that the sequentialfs method uses the methods svmtrain and svmclassifiy to produce a score for the feature. Based on this score, sequentialfs decides whether the feature is kept or removed.

WEKA SVM CLASSIFIER

- Support vector machine using sequential minimal optimization (SMO) technique.
- Uses training and test sets. If no test set is given, the training will be cross validated.
- The number of folds for the cross validation is specified by the user.
- If a test data set is not supplied, a dialog box is displayed with information about the classifier's performance, including a confusion matrix. If a test set is supplied, then a summary of the accuracy percentages for each of the samples in the test set is displayed.

Pseudocode: If a test set is supplied the program preforms the following:

```
Load the two files and convert their structures to an .ARFF format
 Extract the data and set the class indices
 Instantiate the WEKA classifier
 Build the classifier based on the training set
 For each instance in the test set
   classify the nth instance (this returns a label)
   compare the label from the above line and the real label
 end
 Calculate the accuracy and display a summary dialog
```

If a test set is not supplied the program performs the following:

```
Instantiate an instance of Evaluation with the training set
Call the crossValidateModel function of the instance of Evaluation with the
following arguments:
    the type of classifier (SMO in this case)
    the training set
    the number of folds for the cross validation
    the options for the classifier
    a pseudo-random seed
Display the results in a dialog window
```

WEKA SVM ATTRIBUTE SELECTOR

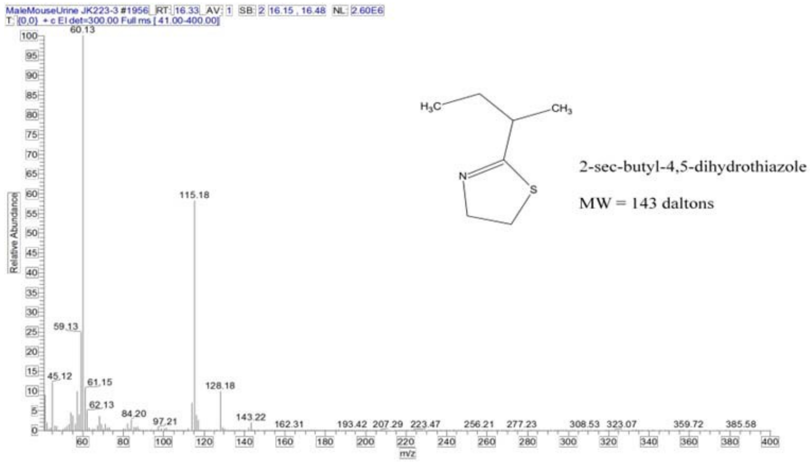
- Support vector machine using sequential minimal optimization (SMO) technique.
- The number folds for cross validation and attributes to select are specified by the user.
- Accepts one file to use as a training data set.
- Displays a dialog box with information about the selected attributes.

```
Load the file and convert its structure to a .ARFF format
Extract the data and set the class index
Instantiate an instance of AttributeSelection, SVMAttributeEval, and Ranker
Set the evaluator of the AttributeSelection as the SVMAttributeEval and the
search as the Ranker
Run the attribute selection
```

APPENDIX B - Peak Alignment Validation

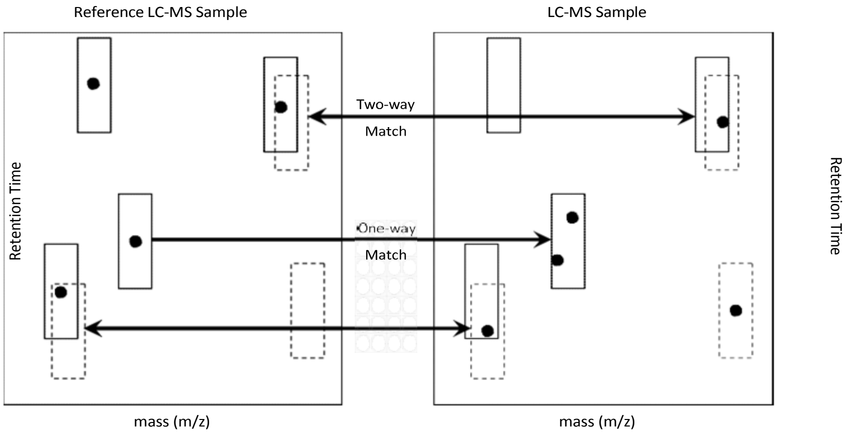
Peak Alignment v2.0 (MeDDL Modification)

- Below are typical landmark volatile peaks observed in male murine urine:



Average Mass	Average Time	Compound ID
(57)	7.92-8.51	3-heptanone
(60), 115, 128	14.72-15.32	2-*sec*-butyl-4,5-dihydrothiazole [SBT]
(71),41,93,121	16.97-17.72	linalool

- Top figure is EI spectra for SBT
- Below is diagram defining a unique 2-way match used in MeDDL for alignment
- Registration windows (time and m/z) are defined in peak.ini by user prior to registration



Example Study Details:

Date range of raw files: 1/20/2011 – 10/4/2011

N=151

Exposure group date of analysis:

• FT1	1/20/2011
• BP1	1/31/2011
• BM2	2/16/2011
• BP2	2/23/2011
• YP2	8/6/2011
• CTR1	8/24/2011
• BA2	9/11/201
• BA3	10/1/2011

Experimental Time Shift:

To accurately register this study, taking into account the 9 month length of analysis, it is necessary to determine time and mass variation of a subset of the above landmark peaks in murine urine. Samples from the C (48hr) timepoint will be used as these are present in each exposure group tested.

Samples used for peak landmark comparison:

FT1_DBA_17_C

BP1_DBA_25_C

BM2_DBA_41_C

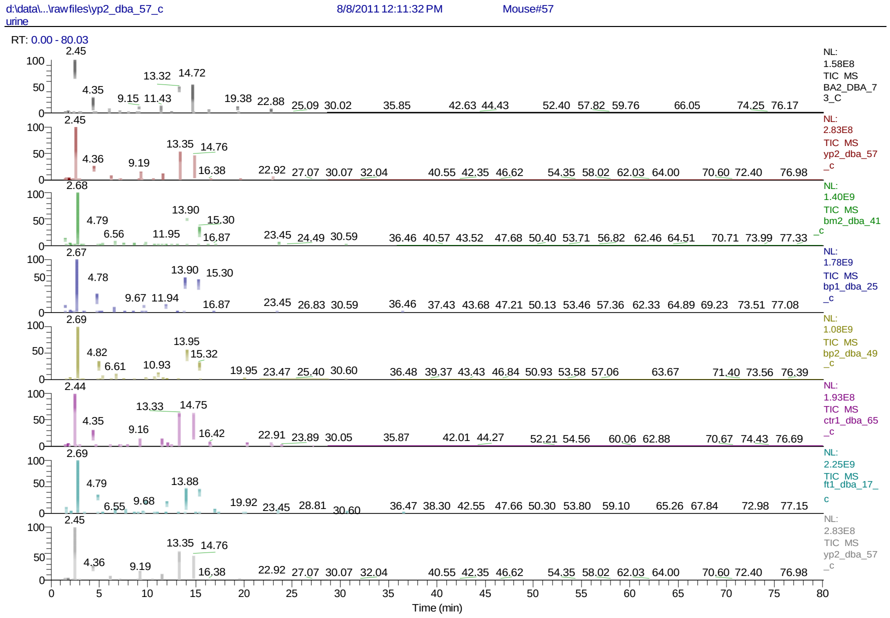
BP2_DBA_49_C

YP2_DBA_57_C

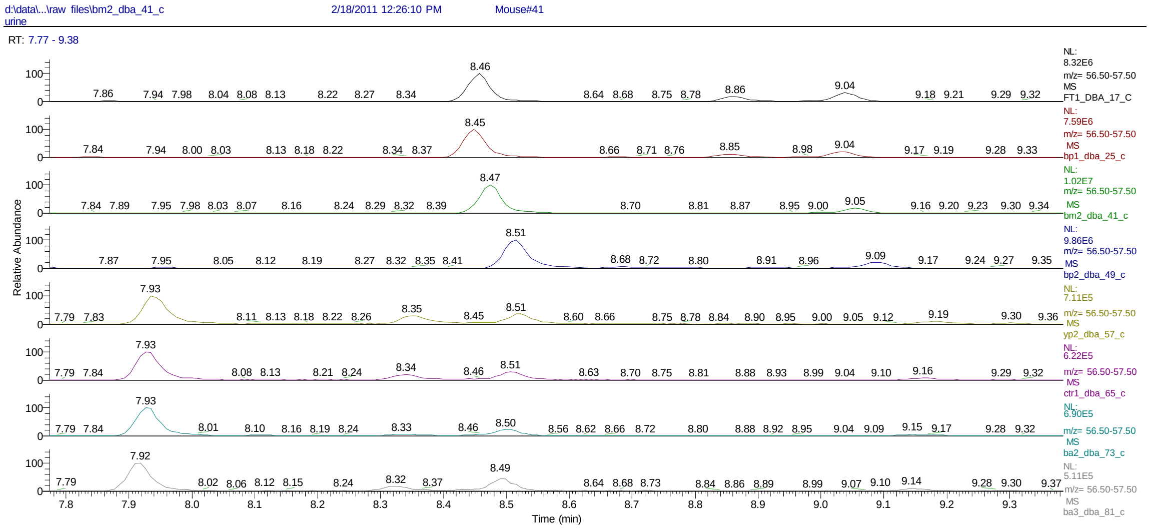
CTR1_DBA_65_C

BA2_DBA_73_C

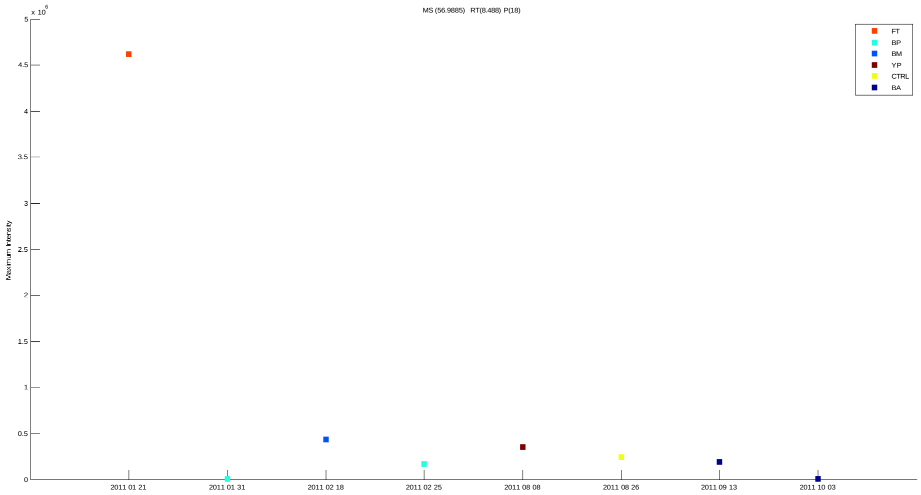
BA3_DBA_81_C



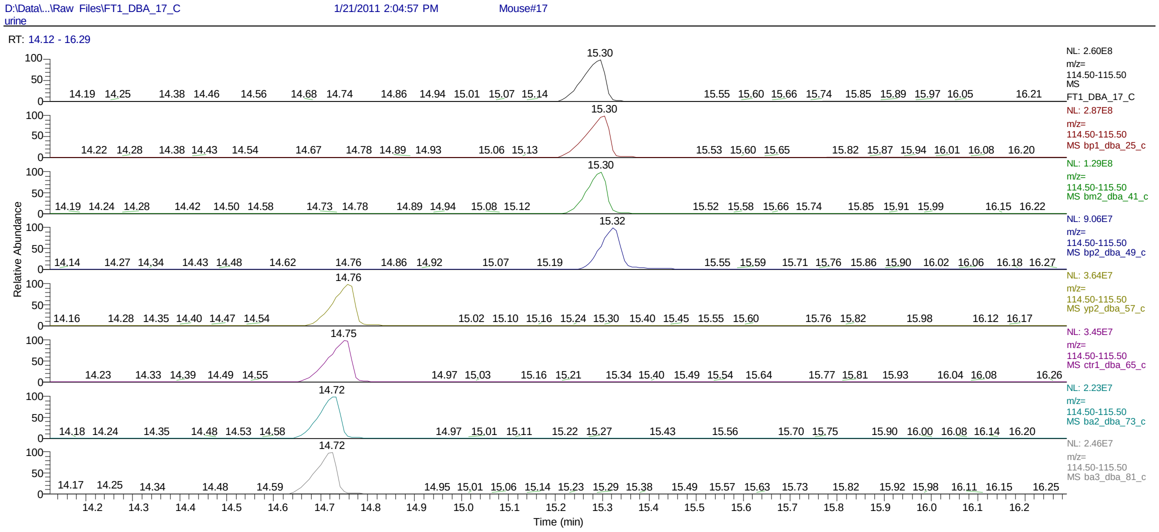
3-Heptanone



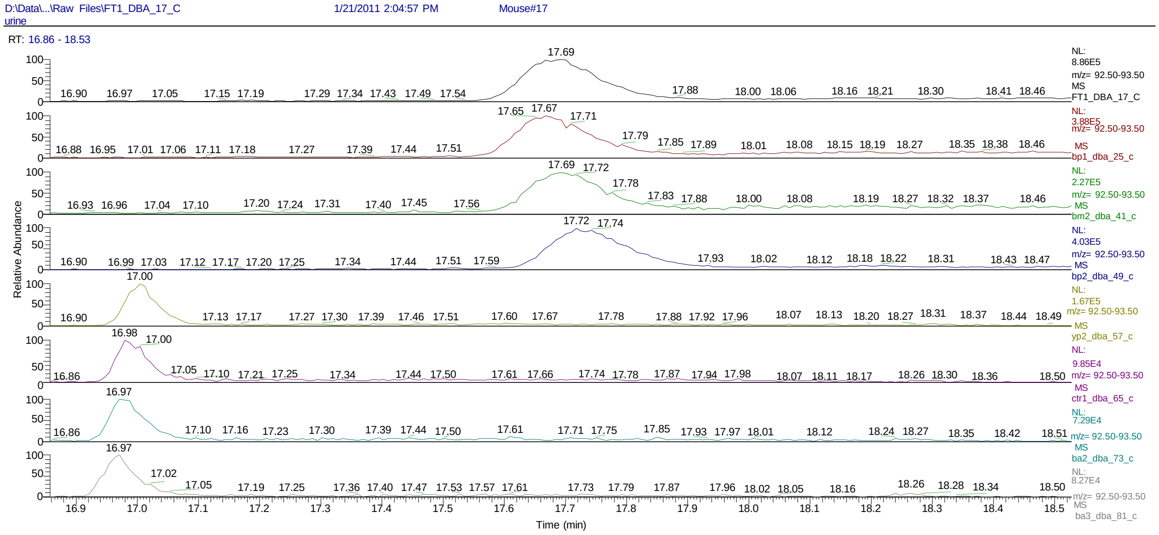
Current registration settings/algorithm not accurately capturing and registering peak (see below scatterplot).



SBT



Linalool

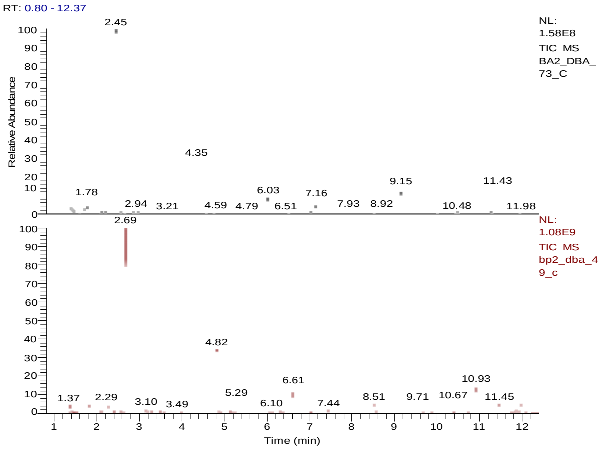


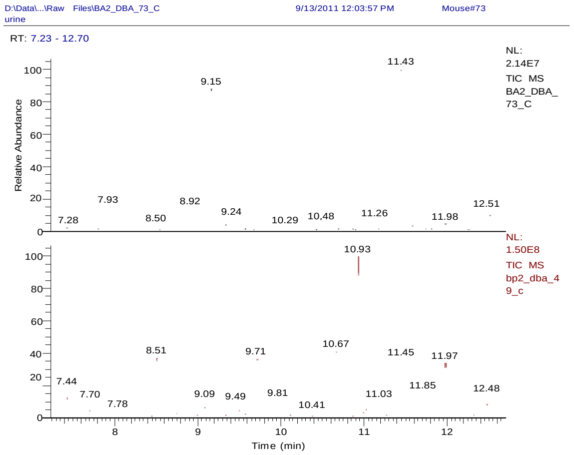
Distribution A. Approved for public release; distribution unlimited.
88ABW-2015-1753; Cleared 07 April 2015

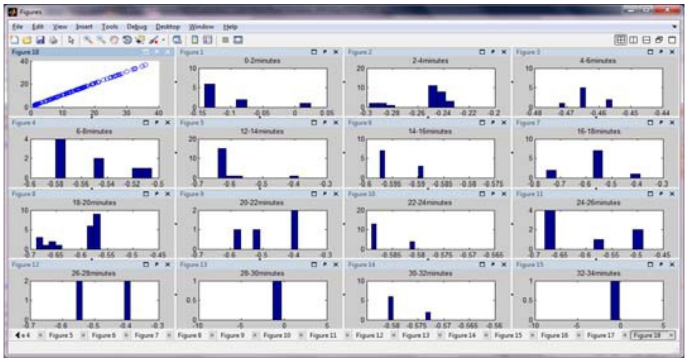
Delta-T Histogram approach: Calculation of chromatographic region/binned (delta T – Chromatographic time shift) histograms

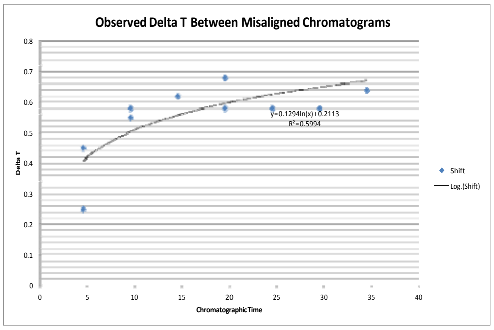
- Generate discrete histograms of the time deltas for misaligned peaks (comparing reference image to new candidate images)
- Only unique, 2-way peak matches are utilized by MeDDL for generating alignment via polynomial fit.
- **Purpose of new function is to down-select identified unique two-way matches for generation of accurate curve for polynomial fit capable of aligning significant time shifts observed.**
- Utilized both static and dynamic deltaT bin sizes (in number of peaks and time windows)
- Initially considered discarding first few minutes due to compounds eluting irrespective of column properties (bimodal deltaT distribution noted).
- DeltaT histograms demonstrated logarithmic increase in time shift over chromatographic time (NOT linear)
- Resulting alignment correction (polyfit) using deltaT distributions limited to third order polynomial as this showed best results

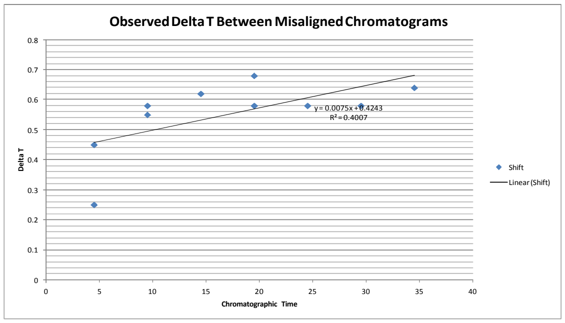
Below are two examples of the misalignment from different time regions:





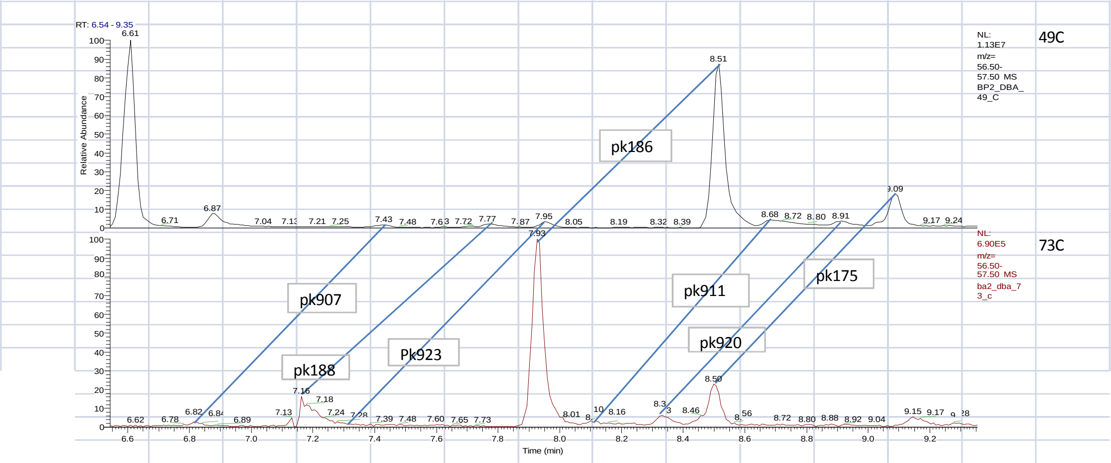

Figures
File Edit View Insert Tools Debug Desktop Window Help
0-2minutes
2-4minutes
4-6minutes
6-8minutes
12-14minutes
14-16minutes
16-18minutes
18-20minutes
20-22minutes
22-24minutes
24-26minutes
26-28minutes
28-30minutes
30-32minutes
32-34minutes
Figure 5 Figure 6 Figure 7 Figure 8 Figure 9 Figure 10 Figure 11 Figure 12 Figure 13 Figure 14 Figure 15 Figure 16 Figure 17 Figure 18





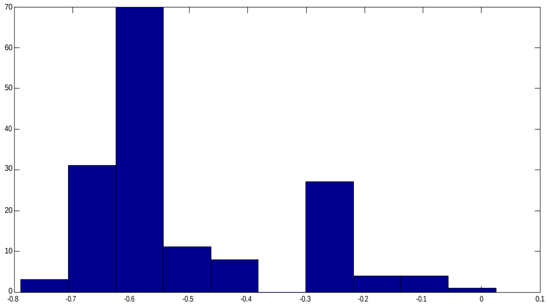
Generated DeltaT histograms, static 2 minute time window:

2 Sample Histogram Alignment Example:

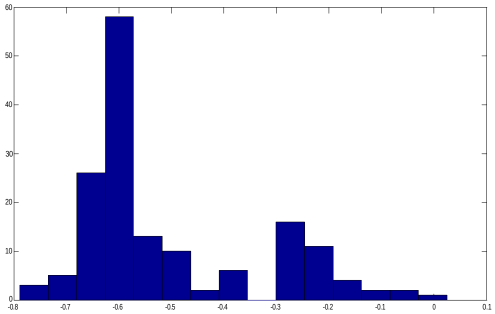


Optimization of number of Bins Per Histogram would be Required. 10 appears to provide adequate resolution for this example.

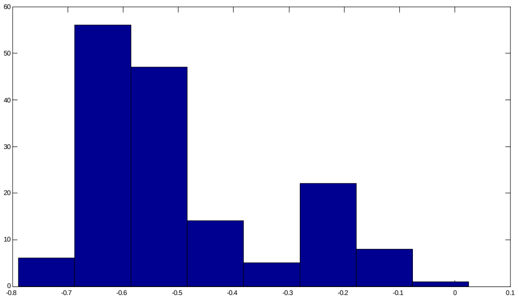
10 bins / 0.58 / matches with observed shift



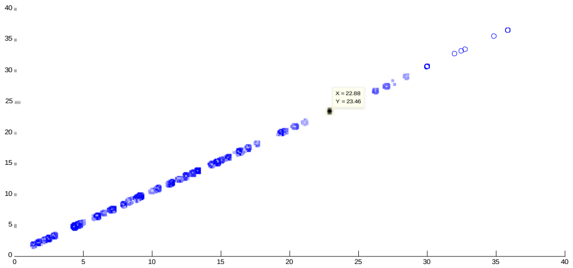
15 bins / 0.6



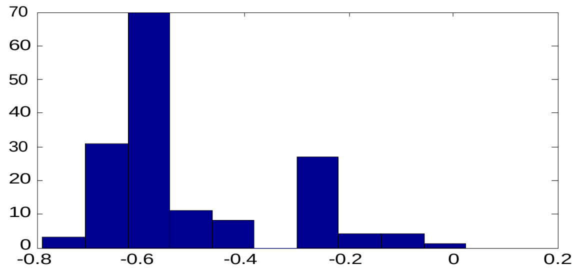
8 bins 0.64



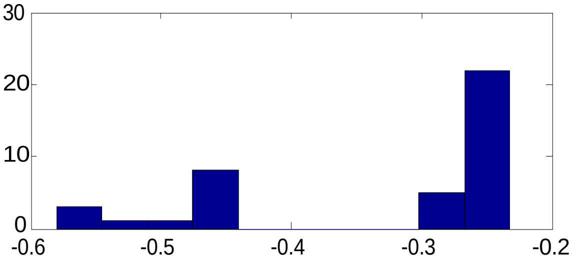
Correlation of unique 2 way matches (i.e. 22.8 / 23.46)

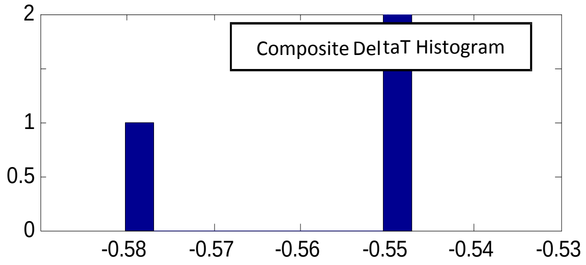


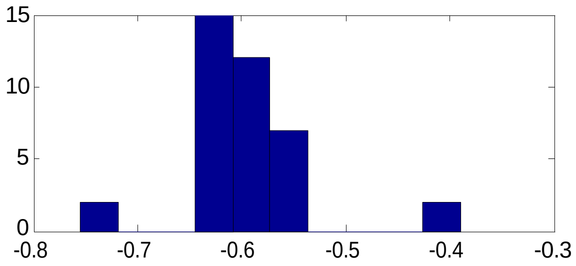
Static 5 Minute Window:



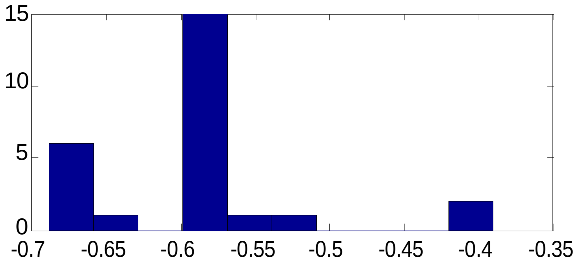
2-7minutes



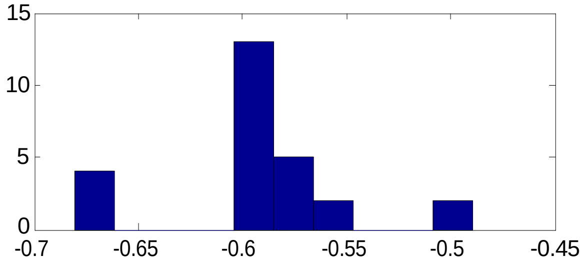




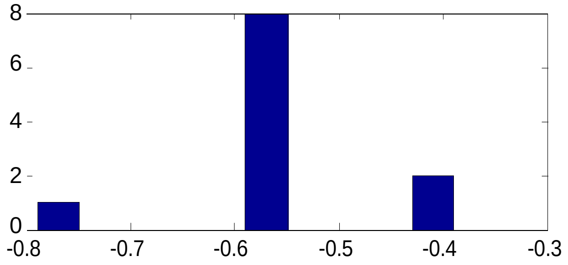
17-22minutes



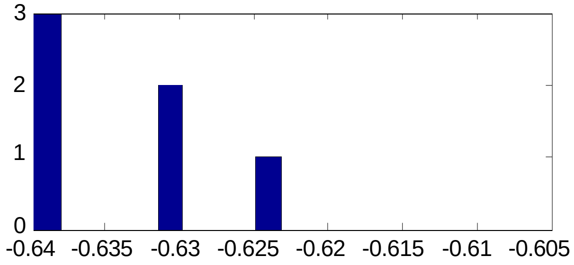
22-27minutes



27-32minutes



32-37minutes

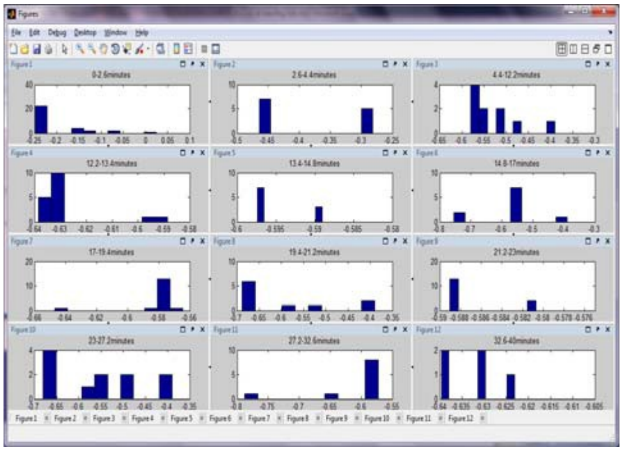


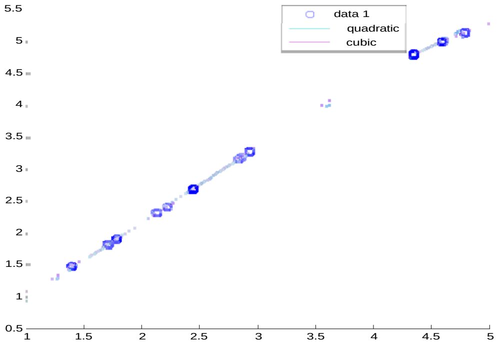
- To attempt to increase accuracy in polyfit and reduce bias in generated deltaT bins due to the peak density, bin sizes were next evaluated dynamically.
- The bin time window increased using defined steps until a bin minimum peak count is achieved (both parameters user specified in the peak.ini).

[HistStep PeaksPerHist]

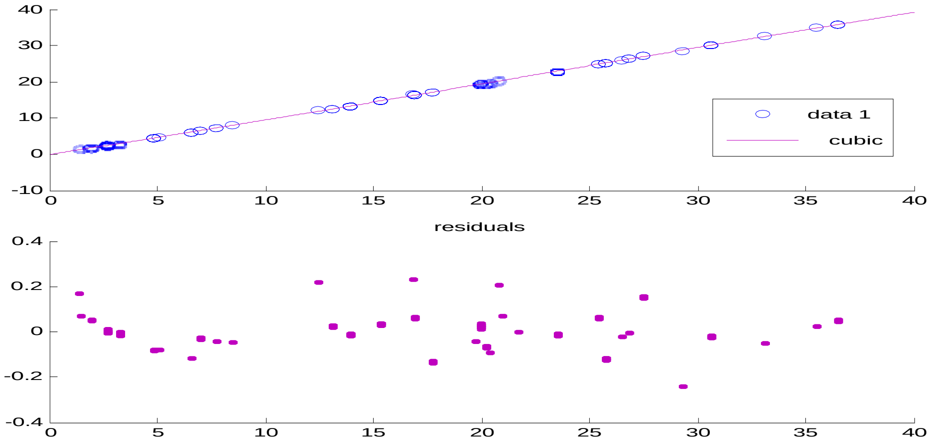
0.01 10

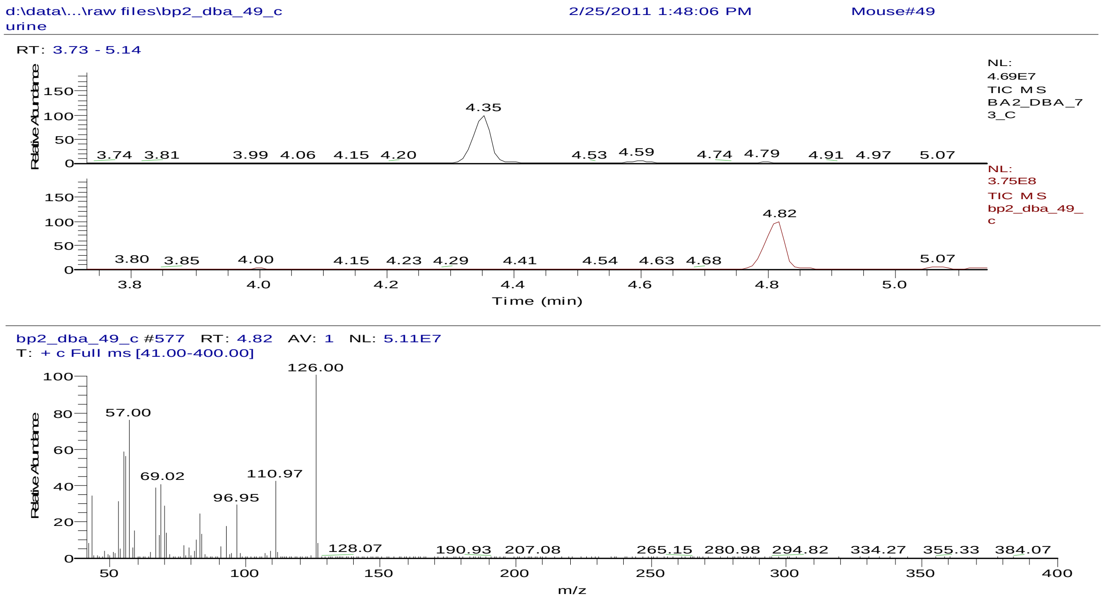
- Below figures show example of dynamic bins and resulting corrective polynomials/residuals.
- Note that in the 0-2.6 min window, a single cluster of 20 peaks exceeds the 10 peak threshold for bin creation.
- This is due to the "hard ionization" EI method utilized causing a large number of fragment peaks generated for a single compound (example below)
- Thus a single bin step can incorporate this molecular "bundle", and induce the bias/bimodal distributions noted in the deltaT histograms.





- Therefore, for GC/MS by EI, we MUST utilize a variant of our time binning filter algorithm in the down-selection of unique two way matching peaks to be used for polynomial fitting as the EI fragmentation patterns too heavily weight individual molecular peaks.



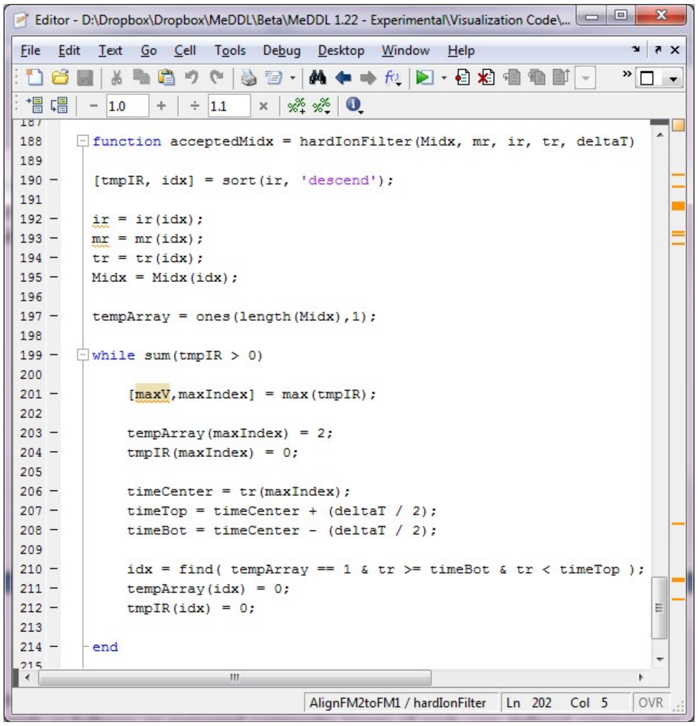

d:\data\...\raw files\bp2_dba_49_c 2/25/2011 1:48:06 PM Mouse#49
urine

RT: 3.73 - 5.14

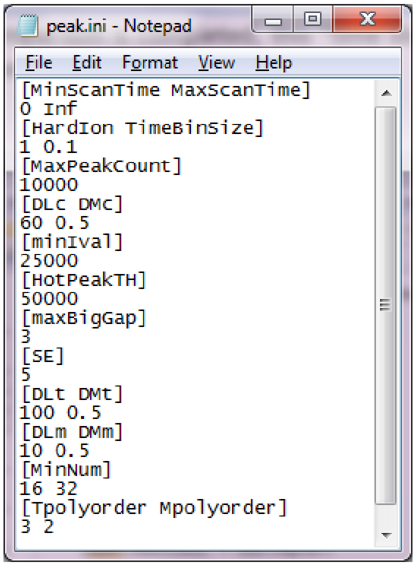
bp2_dba_49_c #577 RT: 4.82 AV: 1 NL: 5.11E7
T: + c Full ms [41.00-400.00]

Time Binned Down Selection of Unique Two Way Matches:

- Utilizes a modification of the time-bin algorithm originally created as a filter for EI spectra differential profiling described as follows: a supplementary time-binned, fold change algorithm which was labeled as “hard ionization”. In the “hard ionization” method, the analyst specified both a time window and peak intensity threshold for comparison. The comparison then proceeded as follows: an averaged composite image of each user-defined comparative group was generated; the most intense peak from all comparative groups was evaluated across all aligned images using criteria specified; once the comparison was completed, this “time slice” based upon the peak apex ± ½ of the specified time window is removed from further analysis and the next most intense set of peaks are compared.
- Application of time binning in the matched peak set allows for the logical and chromatographically distributed down-selection of the most predominant (intense) unique two-way matching peaks and avoids the weighting issues observed in deltaT binning (artifact of EI).
- This algorithm may be turned on/off and the time window parameter specified by the user prior to registration in the peak.ini (below).


Editor - D:\Dropbox\Dropbox\MeDDL\Beta\MeDDL 1.22 - Experimental\Visualization Code\...
File Edit Text Go Cell Tools Debug Desktop Window Help
1.0
1.1
AlignFM2toFM1 / hardIonFilter
Ln 202 Col 5
OVR

```
function acceptedMidx = hardIonFilter(Midx, mr, ir, tr, deltaT)

[tmpIR, idx] = sort(ir, 'descend');

ir = ir(idx);
mr = mr(idx);
tr = tr(idx);
Midx = Midx(idx);

tempArray = ones(length(Midx),1);

while sum(tmpIR > 0)

    [maxV,maxIndex] = max(tmpIR);

    tempArray(maxIndex) = 2;
    tmpIR(maxIndex) = 0;

    timeCenter = tr(maxIndex);
    timeTop = timeCenter + (deltaT / 2);
    timeBot = timeCenter - (deltaT / 2);

    idx = find( tempArray == 1 & tr >= timeBot & tr < timeTop );
    tempArray(idx) = 0;
    tmpIR(idx) = 0;

end
```



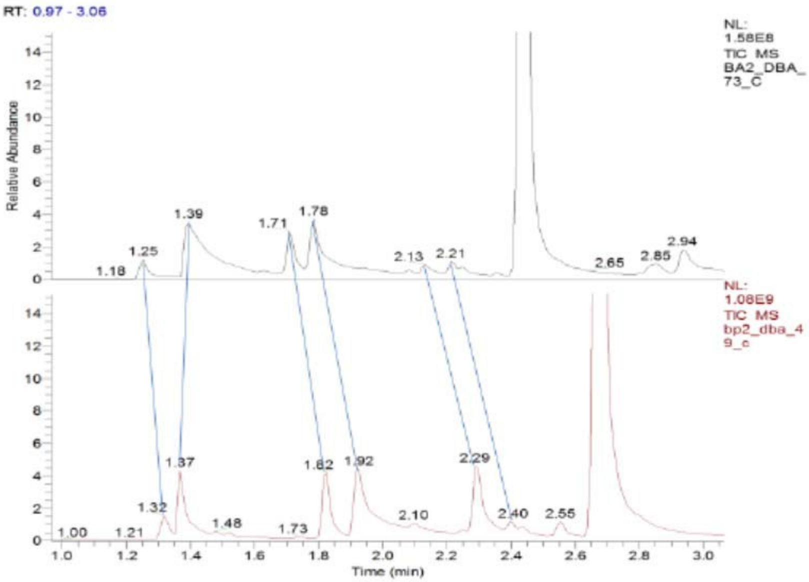
```
[MinScanTime MaxScanTime]
0 Inf
[HardIon TimeBinSize]
1 0.1
[MaxPeakCount]
10000
[DLc DMc]
60 0.5
[minIval]
25000
[HotPeakTH]
50000
[maxBigGap]
3
[SE]
5
[DLt DMt]
100 0.5
[DLm DMm]
10 0.5
[MinNum]
16 32
[Tpolyorder Mpolyorder]
3 2
```

Time Binning 2 Sample Validation:

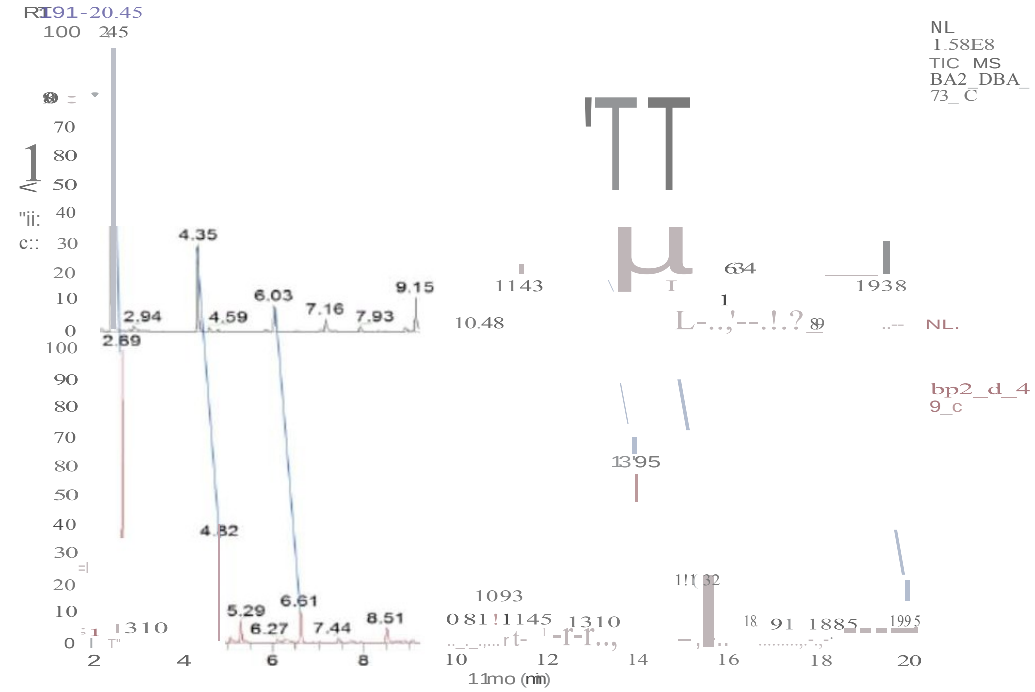
- 2 Sample Validation: Approximate 0.6min shift easily observed.
- Note that previous images show linear fitting inadequate for proper alignment.
- Binning studies show that this shift increased rapidly in the first 4 minutes.
- Due to observed behavior, alignment of the first few chromatographic minutes likely most problematic
- No significant peaks observed after 37 minutes

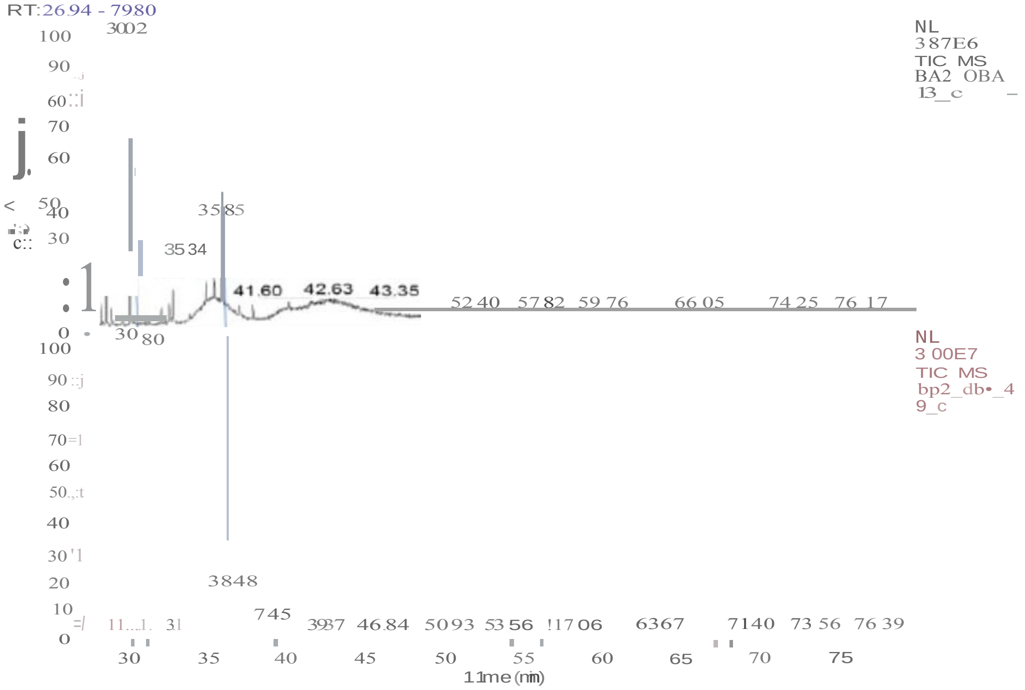
Example Registered Peak Table used for Verification:

											Intensity	Unaligned Time	Matched Time		
Pk74	1	BP2_DBA_49_C	20110225	BP	DBA	49	48	A_UNSTR	155	43.9281	1.63E+07	**1.3185**	**1.3105**	43.9268	23593
	2	BA2_DBA_73_C	20110913	BA	DBA	73	48	A_UNSTR	147	43.9502	1828209	**1.2524**	**1.4042**	43.9489	8690
Pk7	1	BP2_DBA_49_C	20110225	BP	DBA	49	48	A_UNSTR	161	58.0244	2.29E+07	**1.3685**	**1.3604**	58.0234	24580
	2	BA2_DBA_73_C	20110913	BA	DBA	73	48	A_UNSTR	164	58.0181	2732492	**1.3932**	**1.5575**	58.0148	9858
Pk57	1	BP2_DBA_49_C	20110225	BP	DBA	49	48	A_UNSTR	216	42.9806	2.63E+07	**1.8245**	**1.8159**	42.9793	33562
	2	BA2_DBA_73_C	20110913	BA	DBA	73	48	A_UNSTR	202	42.9901	2289576	**1.7083**	**1.8995**	42.9891	12836
Pk559	1	BP2_DBA_49_C	20110225	BP	DBA	49	48	A_UNSTR	228	206.9607	2.26E+07	**1.924**	**1.9153**	206.9314	35821
	2	BA2_DBA_73_C	20110913	BA	DBA	73	48	A_UNSTR	211	206.9519	2583981	**1.7828**	**1.9802**	207.0146	13847
Pk287	1	BP2_DBA_49_C	20110225	BP	DBA	49	48	A_UNSTR	272	71.9965	9676288	**2.2885**	**2.2796**	71.9952	44684
	2	BA2_DBA_73_C	20110913	BA	DBA	73	48	A_UNSTR	253	72.0009	215830	**2.1313**	**2.357**	71.9969	19575
Pk893	1	BP2_DBA_49_C	20110225	BP	DBA	49	48	A_UNSTR	286	56.0603	291657	**2.4047**	**2.3956**	56.0593	47465
	2	BA2_DBA_73_C	20110913	BA	DBA	73	48	A_UNSTR	263	56.0837	52237	**2.2141**	**2.4463**	56.0806	-20800



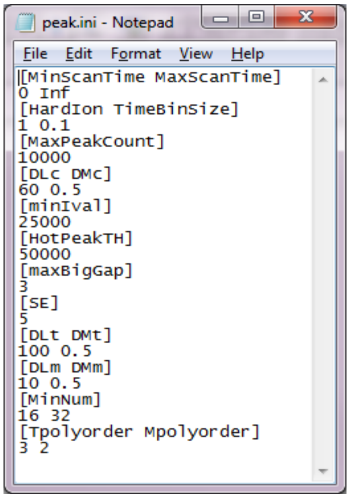
Distribution A. Approved for public release; distribution unlimited.
88ABW-2015-1753; Cleared 07 April 2015


RT: 1.91 - 20.45
NL
1.58E8
TIC MS
BA2_DBA_
73_C
2.45
4.35
6.03
7.16
7.93
9.15
2.94
4.59
2.69
10.48
11.43
19.38
NL
bp2_d_4
9_c
13.95
4.82
5.29
6.27
6.61
7.44
8.51
10.93
11.45
13.10
18.85
19.95
Time (min)

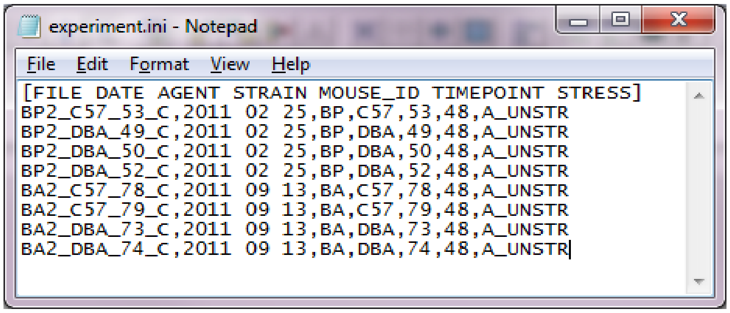

RT: 26.94 - 79.80
NL
3.87E6
TIC MS
BA2_OBA
13_c
30.02
35.85
35.34
41.60
42.63
43.35
52.40
57.82
59.76
66.05
74.25
76.17
30.80
NL
3.00E7
TIC MS
bp2_db_4
9_c
38.48
39.37
46.84
50.93
53.56
63.67
71.40
73.56
76.39
Time (min)

Time Binning 8 Sample Validation:

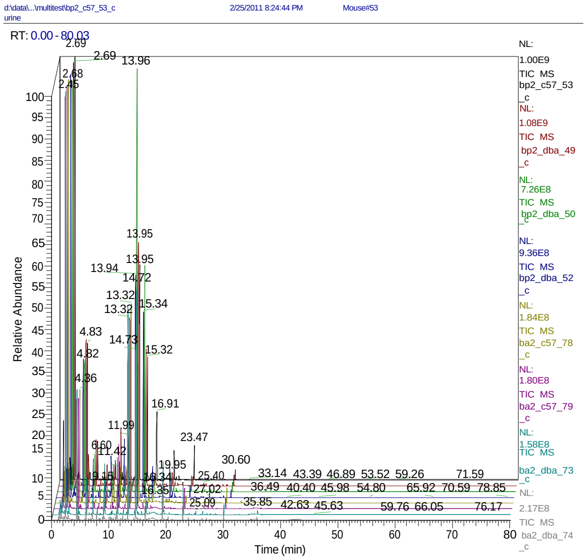
- Subset of sample validated: Limited to 8 by Thermo Xcalibur in number of spectral overlays possible
- Shifts similar to 2 sample validation

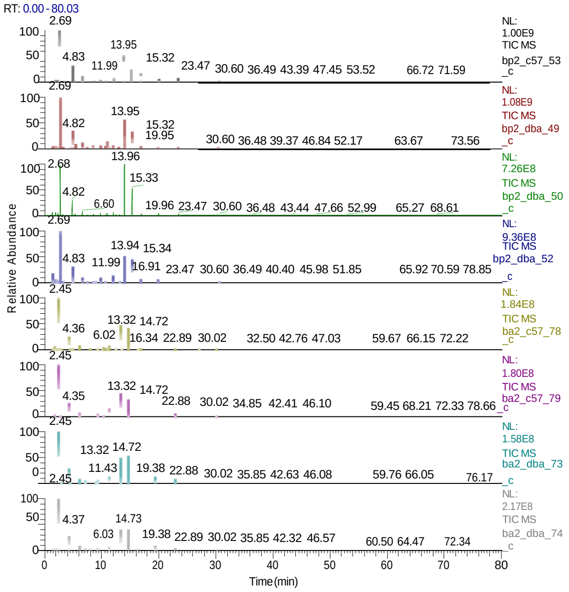


```
[MinScanTime MaxScanTime]
0 Inf
[HardIon TimeBinSize]
1 0.1
[MaxPeakCount]
10000
[DLc DMc]
60 0.5
[minIval]
25000
[HotPeakTH]
50000
[maxBigGap]
3
[SE]
5
[DLt DMt]
100 0.5
[DLm DMm]
10 0.5
[MinNum]
16 32
[Tpolyorder Mpolyorder]
3 2
```

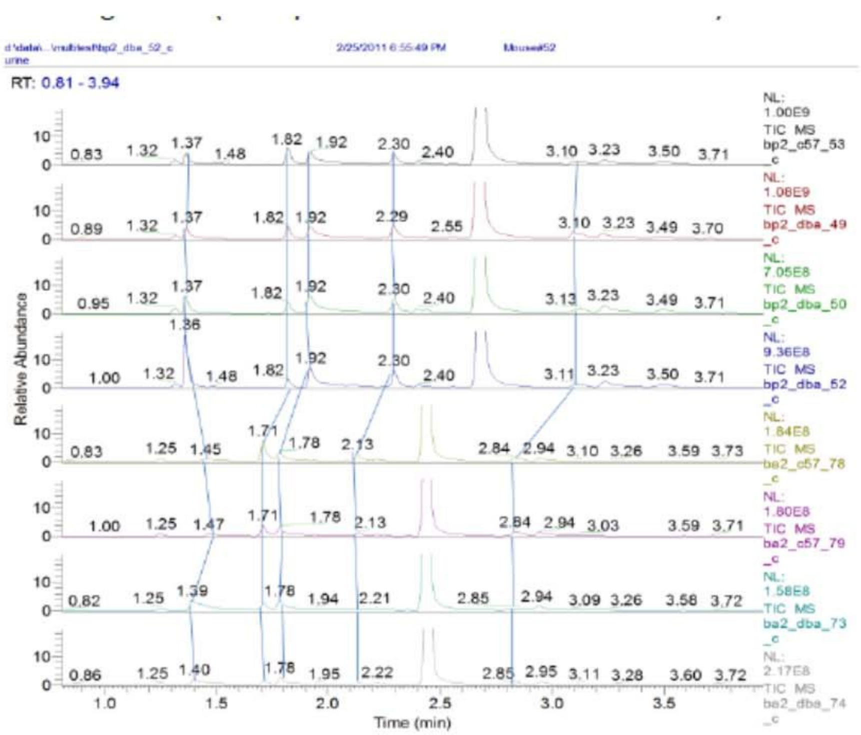


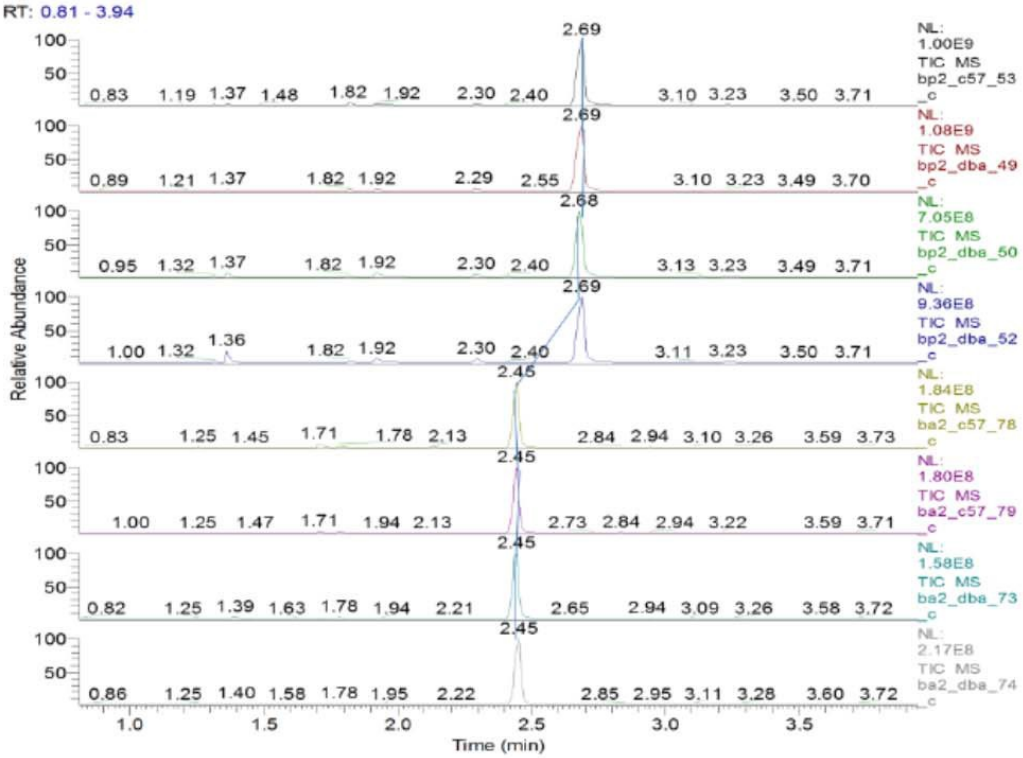
```
[FILE DATE AGENT STRAIN MOUSE_ID TIMEPOINT STRESS]
BP2_C57_53_C,2011 02 25,BP,C57,53,48,A_UNSTR
BP2_DBA_49_C,2011 02 25,BP,DBA,49,48,A_UNSTR
BP2_DBA_50_C,2011 02 25,BP,DBA,50,48,A_UNSTR
BP2_DBA_52_C,2011 02 25,BP,DBA,52,48,A_UNSTR
BA2_C57_78_C,2011 09 13,BA,C57,78,48,A_UNSTR
BA2_C57_79_C,2011 09 13,BA,C57,79,48,A_UNSTR
BA2_DBA_73_C,2011 09 13,BA,DBA,73,48,A_UNSTR
BA2_DBA_74_C,2011 09 13,BA,DBA,74,48,A_UNSTR
```



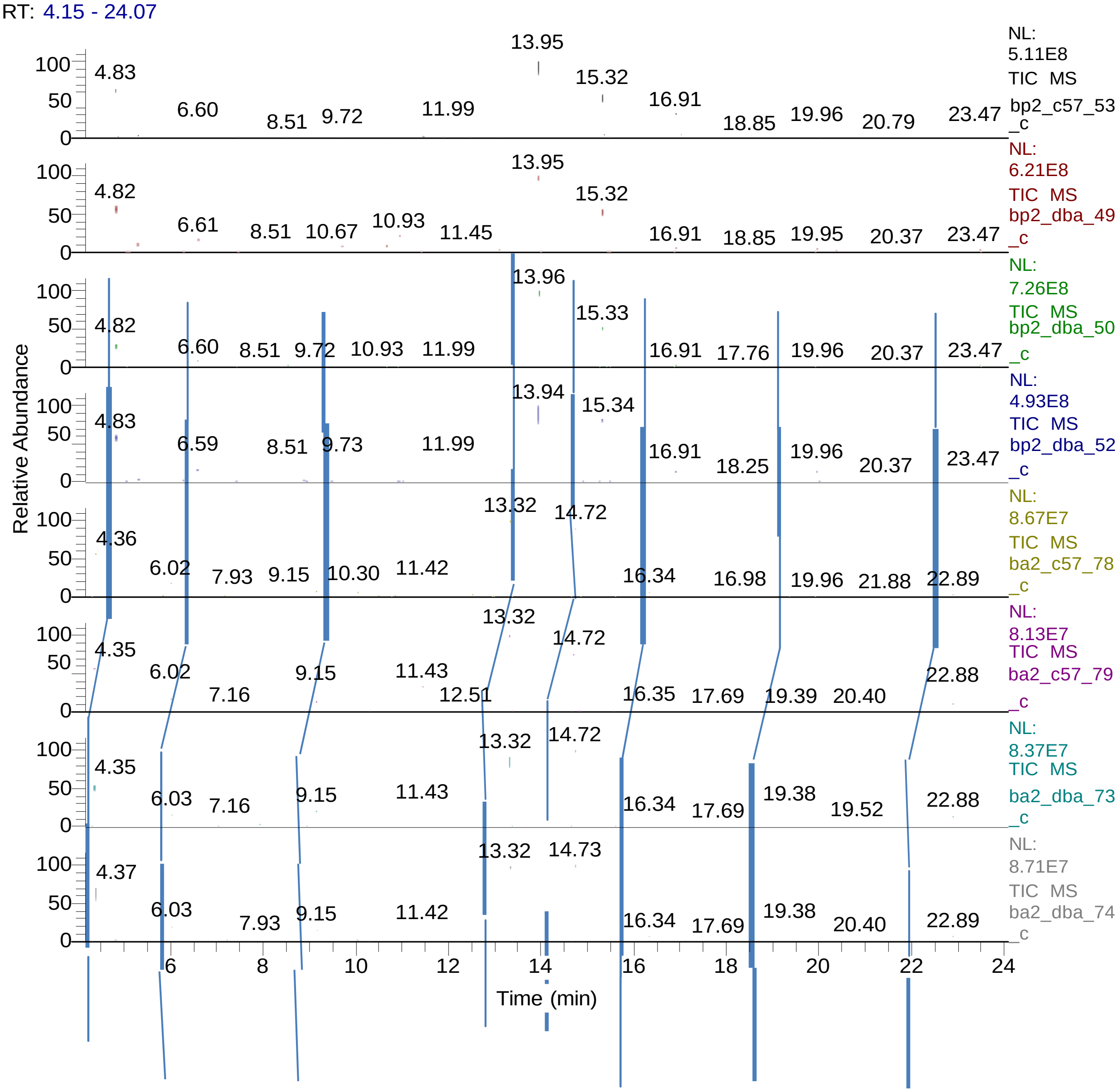


- The 0-4 minute time range arguably best illustrates the power of this method for EI GC/MS alignment (note peaks at 1.37 and 1.82 minutes).





- Please note response of pk437 and pk1892 (4.82 min).



Distribution A. Approved for public release; distribution unlimited.
88ABW-2015-1753; Cleared 07 April 2015

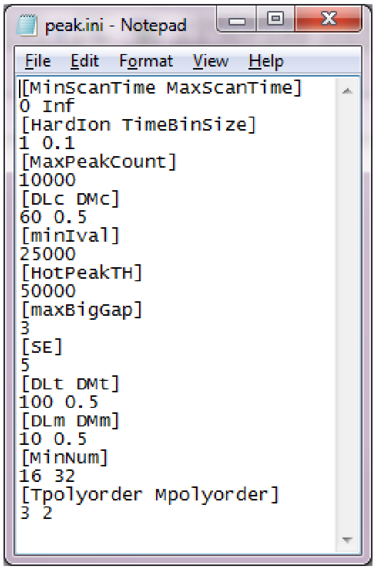
- Registration of the leading shoulder can be eliminated through the time binning filter with a 0.1 min window (verified)
- Not artifact, as new registration algorithm only down-selects 2 way matches used for alignment and does not affect peak detection.

Pk437	1	BP2_C57_53_C	2011 02 25	BP	C57	53	48	A_UNSTR	577	125.9954	4.73E+07	**4.8172**	**4.7937**	125.9941	115595
	2	BP2_DBA_49_C	2011 02 25	BP	DBA	49	48	A_UNSTR	577	125.9956	5.11E+07	**4.8172**	**4.7985**	125.9939	110748
	3	BP2_DBA_50_C	2011 02 25	BP	DBA	50	48	A_UNSTR	577	125.9974	3.13E+07	**4.8172**	**4.7972**	125.9935	114649
	4	BP2_DBA_52_C	2011 02 25	BP	DBA	52	48	A_UNSTR	578	126.0016	4.38E+07	**4.8253**	**4.802**	126.0033	116725
	5	BA2_C57_78_C	2011 09 13	BA	C57	78	48	A_UNSTR	521	125.9819	6808069	**4.3529**	**4.7263**	125.9872	65118
	6	BA2_C57_79_C	2011 09 13	BA	C57	79	48	A_UNSTR	521	126.0349	6433989	**4.3529**	**4.7304**	126.0305	64539
	7	BA2_DBA_73_C	2011 09 13	BA	DBA	73	48	A_UNSTR	521	125.983	6293168	**4.3529**	**4.7294**	125.9796	57477
	8	BA2_DBA_74_C	2011 09 13	BA	DBA	74	48	A_UNSTR	523	125.9842	8536832	**4.3695**	**4.7415**	125.9808	-65652
Pk1892	1	BP2_C57_53_C	2011 02 25	BP	C57	53	48	A_UNSTR	577	125.9954	4.73E+07	**4.8172**	**4.7937**	125.9941	-115595
	2	BP2_DBA_49_C	2011 02 25	BP	DBA	49	48	A_UNSTR	554	126.0493	103342	**4.6266**	**4.6088**	126.0476	105737
	3	BP2_DBA_50_C	2011 02 25	BP	DBA	50	48	A_UNSTR	580	125.0734	95976	**4.8422**	**4.822**	125.0694	115301
	4	BP2_DBA_52_C	2011 02 25	BP	DBA	52	48	A_UNSTR	577	126.0429	4.23E+07	**4.8172**	**4.794**	126.0446	-116493
	5	BA2_C57_78_C	2011 09 13	BA	C57	78	48	A_UNSTR	521	125.9819	6808069	**4.3529**	**4.7263**	125.9872	-65118
	6	BA2_C57_79_C	2011 09 13	BA	C57	79	48	A_UNSTR	521	126.0349	6433989	**4.3529**	**4.7304**	126.0305	-64539
	7	BA2_DBA_73_C	2011 09 13	BA	DBA	73	48	A_UNSTR	521	125.983	6293168	**4.3529**	**4.7294**	125.9796	-57477
	8	BA2_DBA_74_C	2011 09 13	BA	DBA	74	48	A_UNSTR	523	125.9842	8536832	**4.3695**	**4.7415**	125.9808	-65652

Peak Shoulder Registration:

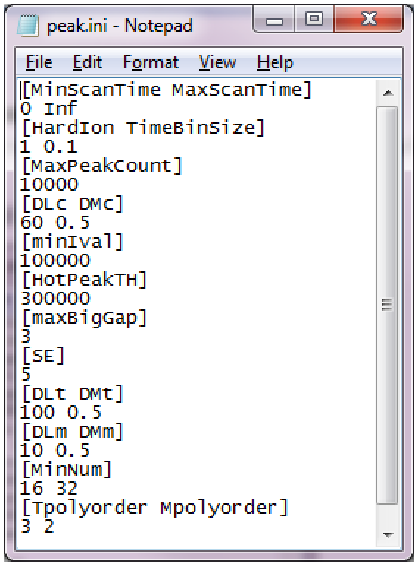
Increasing baseline to 300K absolute eliminates registration of shoulder peaks.

Old registration parameters



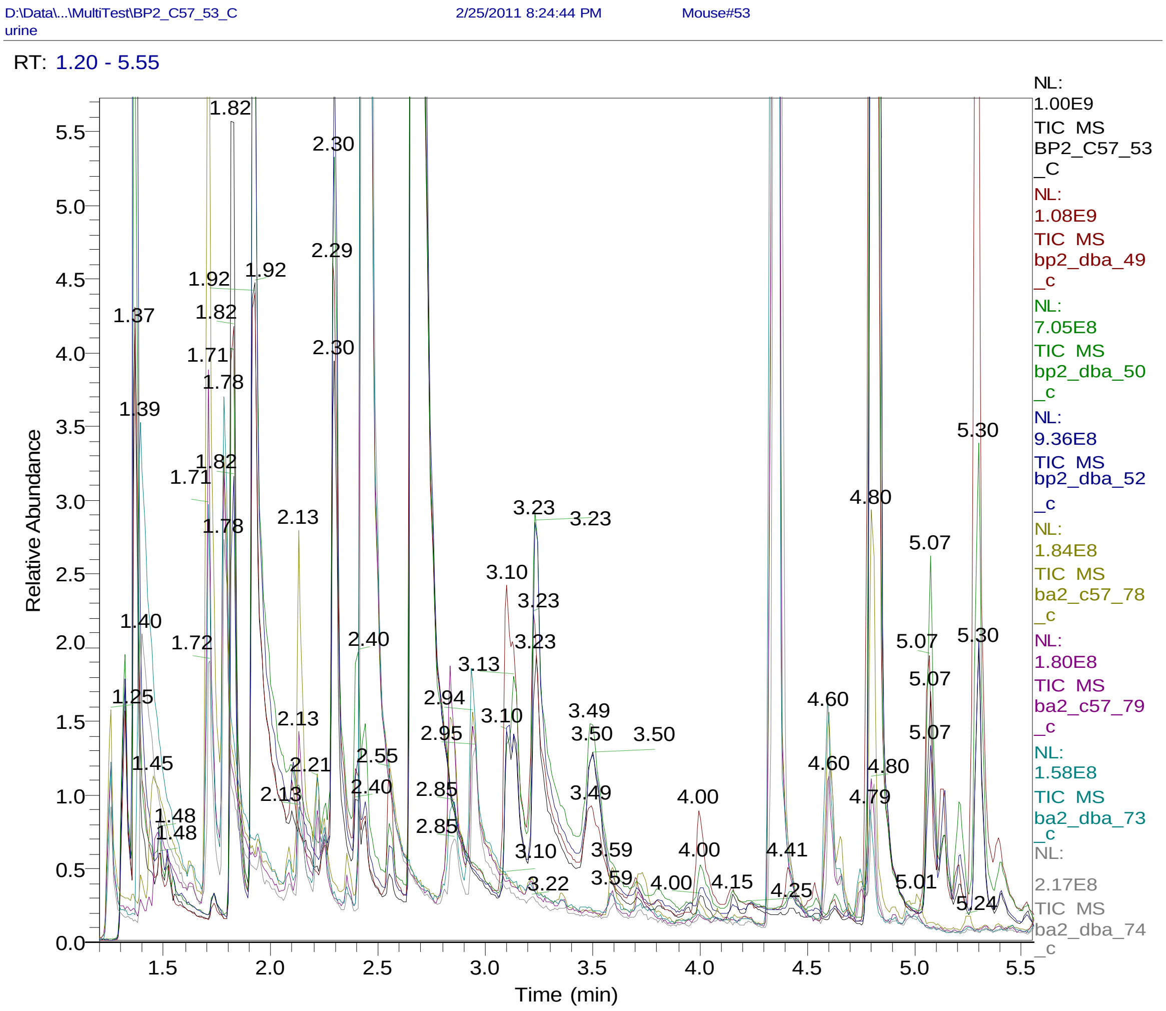
```
[MinScanTime MaxScanTime]
0 Inf
[HardIon TimeBinSize]
1 0.1
[MaxPeakCount]
10000
[DLc DMc]
60 0.5
[minIval]
25000
[HotPeakTH]
50000
[maxBigGap]
3
[SE]
5
[DLt DMt]
100 0.5
[DLm DMm]
10 0.5
[MinNum]
16 32
[Tpolyorder Mpolyorder]
3 2
```

New registration parameters



```
[MinScanTime MaxScanTime]
0 Inf
[HardIon TimeBinSize]
1 0.1
[MaxPeakCount]
10000
[DLc DMc]
60 0.5
[minIval]
100000
[HotPeakTH]
300000
[maxBigGap]
3
[SE]
5
[DLt DMt]
100 0.5
[DLm DMm]
10 0.5
[MinNum]
16 32
[Tpolyorder Mpolyorder]
3 2
```

Resulted in 2422 peaks detected	Resulted in 918 peaks detected
[minIval]	[minIval]
25000	100000
[HotPeakTH]	[HotPeakTH]
50000	300000



Additional manual examination of right combined spectra demonstrate that use of above peak.ini parameters resulted in no missed significant peaks and no duplicate registration of leading/tailing edges.

Conclusion: No overlap in peak registration (shoulders) for same mass seen when baseline increased to 300K absolute.

APPENDIX C - Peak Normalization

Normalization Test (8 Sample)

Purpose: To evaluate the effect of normalization on stress study data (8 sample example)

Example Study Details:

Date range of raw files: 1/20/2011 – 10/4/2011

N=151

Exposure group date of analysis:

FT1	1/20/2011	Group A
BP1	1/31/2011	
BM2	2/16/2011	
BP2	2/23/2011	
YP2	8/6/2011	Group B
CTR1	8/24/2011	
BA2	9/11/201	
BA3	10/1/2011	

Decreased Sensitivity:

To accurately analyze this study, normalization of some type is required. Following the 6 month down-time of Feb – Aug, a significant decrease in sensitivity was noted. Samples from the C (48hr) timepoint will be used for validation as these are present in each exposure group tested.

Samples used for peak landmark comparison:

FT1_DBA_17_C

BP1_DBA_25_C

BM2_DBA_41_C

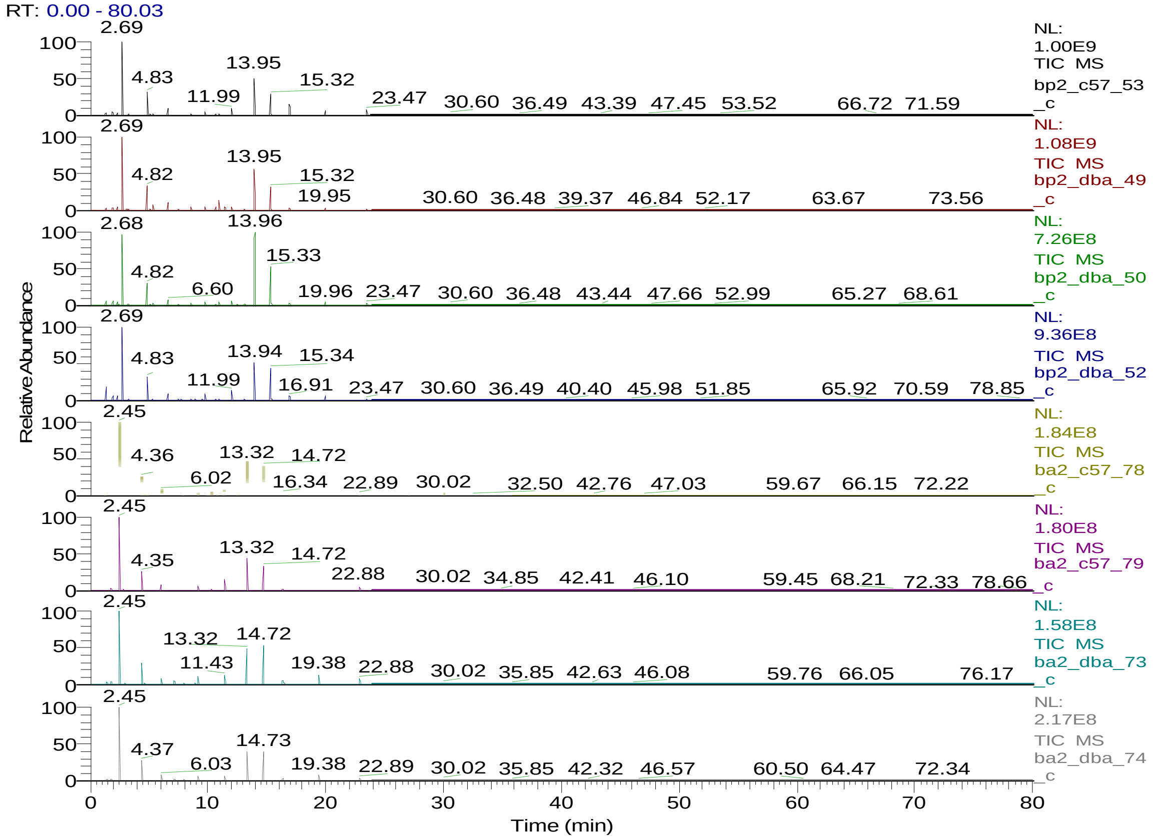
BP2_DBA_49_C

YP2_DBA_57_C

CTR1_DBA_65_C

BA2_DBA_73_C

BA3_DBA_81_C



Summary of Major Peaks:

Numbers for UnNorm, TIC Norm, and Olympic represent the relative fold change between analysis time groups (A/B) referenced above.

R.T.	Peak	UnNorm	TIC Norm	Olympic
2.7	126	5.51	1.09	1.15
4.78	152	6.18	1.03	1.02
6.54	153	6.3	1.04	1.02
13.93	141	6.56	1.11	1.06
15.32	60	4.57	1.29	1.35
19.95	131	5.31	1.1	1.13

Olympic used 80/20%