HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2024 October 16.

Published in final edited form as:

Anal Chem. 2024 October 08; 96(40): 15970–15979. doi:10.1021/acs.analchem.4c03256.

Streamlining Phenotype Classification and Highlighting Feature Candidates: A Screening Method for Non-Targeted Ion Mobility Spectrometry-Mass Spectrometry (IMS-MS) Data

Jessie R. Chappel,
Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina 27606, United States

Kaylie I. Kirkwood-Donelson,
Immunity, Inflammation, and Disease Laboratory, National Institute of Environmental Health Sciences, Durham, North Carolina 27709, United States

James N. Dodds,
Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, United States

Jonathon Fleming,
Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina 27606, United States

David M. Reif,
Predictive Toxicology Branch, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Durham, North Carolina 27709, United States

Erin S. Baker
Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, United States

Abstract

Nontargeted analysis (NTA) is increasingly utilized for its ability to identify key molecular features beyond known targets in complex samples. NTA is particularly advantageous in exploratory studies aimed at identifying phenotype-associated features or molecules able to classify various sample types. However, implementing NTA involves extensive data analyses and labor-intensive annotations. To address these limitations, we developed a rapid data screening capability compatible with NTA data collected on a liquid chromatography, ion mobility

Corresponding Authors David M. Reif – Predictive Toxicology Branch, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Durham, North Carolina 27709, United States; david.reif@nih.gov; **Erin S. Baker** – Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, United States; erinmsb@unc.edu.

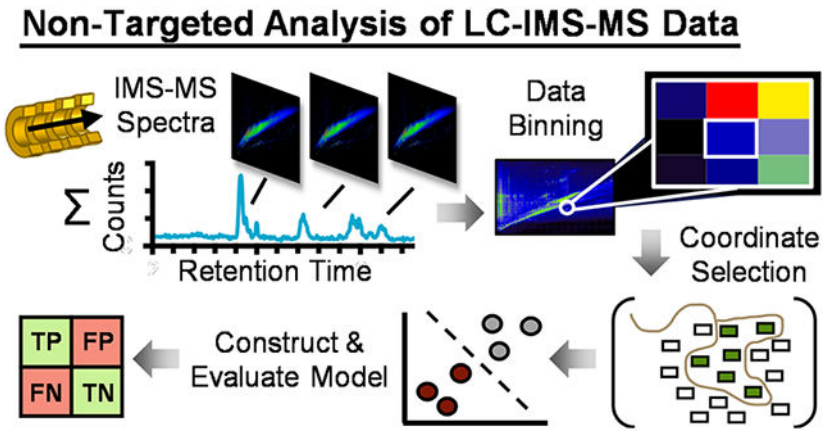
Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.4c03256.
Data set descriptions; workflow discussion; example IMS-MS heatmaps; tabulated coordinates and descriptors for lipids and PFAS; analysis of passive sampler data with reduced *m/z* bin size; binning process for IMS-MS frames; suspected PFAS correlations to known analytes (PDF)

The authors declare no competing financial interest.

spectrometry, and mass spectrometry (LC-IMS-MS) platform that allows for sample classification while highlighting potential features of interest. Specifically, this method aggregates the thousands of IMS-MS spectra collected across the LC space for each sample and collapses the LC dimension, resulting in a single summed IMS-MS spectrum for screening. The summed IMS-MS spectra are then analyzed with a bootstrapped Lasso technique to identify key regions or coordinates for phenotype classification via support vector machines. Molecular annotations are then performed by examining the features present in the selected coordinates, highlighting potential molecular candidates. To demonstrate this summed IMS-MS screening approach, we applied it to clinical plasma lipidomic NTA data and exposomic NTA data from water sites with varying contaminant levels. Distinguishing coordinates were observed in both studies, enabling the evaluation of phenotypic molecular annotations and resulting in screening models capable of classifying samples with up to a 25% increase in accuracy compared to models using annotated data.

Graphical Abstract



INTRODUCTION

The use of mass spectrometry to measure and understand molecular changes in diverse sample types has shown significant implications in diagnostics,[1] phenotype evaluations,[2-5] environmental assessments,[6-10] and many other fields.[11,12] For these assessments, two broad methodologies, either targeted or nontargeted analyses, are utilized for characterizing molecular changes depending on whether quantitative or comprehensive results are desired.[13] Specifically, targeted analyses are designed to detect and quantify a predefined set of known molecules, while nontargeted or untargeted analyses (NTA), aim to provide a comprehensive assessment of all detectable molecules within each sample of interest. When applicable, targeted approaches are typically preferred due to their enhanced selectivity, sensitivity, speed, and cost-efficiency compared to NTA.[14] However, targeted studies require *a priori* knowledge of the specific molecule(s) of interest, which is a limitation when studying poorly understood or novel sample types. In such cases, NTA are desirable due to their ability to provide a more comprehensive view of the molecular landscape and identify novel molecular connections with outcomes of interest.[15]

While NTA workflows claim to offer an impartial snapshot of the detectable molecules within a sample, the commonly utilized instrumental and data analysis processes can be time-consuming and challenging. One such challenge arises from the diverse physicochemical properties that may be present across molecules in a sample, encompassing variations in size, polarity, charge, and chemical structure.[16] To address these complexities, advanced analytical techniques such as high-resolution mass spectrometry (MS) instrumentation coupled with front-end separations such as gas or liquid chromatography (GC and LC) and ion mobility mass spectrometry (IMS) are often used to aid in molecular characterization.[17-19] Following laboratory analysis, the thousands of multidimensional molecular features are annotated via a database comparison to known compounds. This process involves matching the observed molecular descriptors, such as mass to charge ratios (*m/z*), isotopic distributions, retention times, fragmentation data, and other chemical fingerprints to entries in the reference databases.[20,21] Navigating this process can be tedious, as numerous databases exist, each differing in the breadth of molecules covered and type of molecular descriptors available.[22] Additionally, novel molecules may not exist in any database, resulting in unidentified features. For example, in metabolomic NTA, as few as 2–10% of detected features are reliably annotated.[23] Moreover, even when database matches are found, molecules cannot be accurately quantified or identified with the highest confidence without the use of a reference standard, which only exist for a limited number of chemicals.[24,25] Once identified, statistical analyses are necessary to determine which molecules or features are representative of conditions of interest, adding another layer of complexity to the process.[26]

To date, a majority of NTA experiments are conducted using either GC-MS or LC-MS, and as such provide two dimensions of molecular separation for feature finding and molecular annotation based on polarity, boiling point, and mass. However, IMS adds a unique dimension by separating analytes based on their gas-phase molecular size or collision cross section (CCS, typically denoted in units of square angstroms, $Å^2$).[27,28] Currently, there are several commercially available IMS platforms that can be interfaced between LC and MS separations, however drift tube IMS (DTIMS) is particularly advantageous for NTA due to its application of a uniform electric field within the drift tube which provides direct correlation between an ion's CCS and measured drift time. DTIMS scans are also incredibly rapid (conducted on a millisecond time scale), and easily nested into existing LC-MS workflows.[28] LC-IMS-MS methods therefore provide complementary separation mechanisms of polarity, size, and mass to enable multidimensional profiling of features in each sample.[29,30] Although several publications have highlighted the utility of IMS-MS separations conducted in the absence of chromatography,[17,31] utilizing LC in conjunction with IMS-MS affords several analytical advantages including significant reductions in ion suppression, minimization of matrix effects, and increased separation capacity for isobaric and isomeric species.[32] While the improved annotation confidence for chemical unknowns obtained through multidimensional LC-IMS-MS data has been demonstrated in previous works,[7,29] the additional dimensionality of the data structure presents unique challenges to traditional feature finding methods in NTA workflows. LC-IMS-MS data requires additional computational resources for feature finding and as not all MS vendors utilize IMS, most commercially available software does not possess the capability to process these files.

Significant progress has been made recently with the publication of data analysis tools such as MZmine3 and LC-IMS-MS Feature Finder as a reflection of communal interest toward incorporation of IMS separations into traditional LC-MS workflows, though further progress will still be beneficial as the separations become more common.[33,34]

Thus, NTA currently faces a paradoxical situation: while the use of advanced analytical methods increases knowledge about present compounds, processing this multidimensional data imposes a computational bottleneck limiting the rate at which NTA data sets can be analyzed. To balance these opposing factors, we propose a rapid screening method that streamlines sample classification in LC-IMS-MS workflows while highlighting potential feature candidates of interest. Our approach involves collapsing the chromatographic dimension by summing the IMS-MS spectra across the LC space to yield a singular summed 2-dimensional IMS-MS spectrum. Summed IMS-MS spectra are then analyzed in R using a bootstrapped least absolute shrinkage and selection operator (Lasso) approach to locate phenotype distinguishing *m/z* and drift time coordinates and construct classification models using support vector machines for test data assessment. Model verification is then performed to determine the predictive value of the selected *m/z* and drift time coordinates, and individual features within these coordinates are examined to see if annotations to chemical structures can be made.

To demonstrate the utility of this summed IMS-MS screening method, we examined environmental and clinical data that has been previously evaluated using existing NTA workflows.[35] Both NTA LC-IMS-MS data from passive aquatic samplers deployed upstream and downstream from a fluorochemical manufacturer in North Carolina's Cape Fear River (Figure 1A) and plasma lipidomic data collected on the day of delivery from either control or preeclamptic individuals (Figure 1B) were assessed to see the application range of the summed IMS-MS screening method. In both cases, the screening method was able to differentiate samples, as will be further detailed in the Results section. The summed IMS-MS screening approach was also compared to the previously used NTA pipeline where individual features were first annotated and associated abundances were used for sample delineation. In this comparison, the summed IMS-MS screening approach not only resulted in improved predictions, but also highlighted unique features not identified in initial NTA assessments.

METHODS

Comprehensive descriptions of the data sets used can be found in the Supporting Information.

Overview of Screening Method.

For both the passive sampler and pregnancy study, data were collected using an LC-IMS-MS workflow and processed as illustrated in Figure 2. The DTIMS platform utilized for data collection (the Agilent 1290 UPLC and Agilent 6560 IM-QTOF MS (Santa Clara, CA)) has been utilized for both lipid and PFAS analysis previously.[7,36] Raw data files (.d files) were then imported into Agilent's IM-MS Browser (v. 10.0), where summed IMS-MS spectra were constructed for the feature space by summing signal abundances for each *m/z* and drift

time value across the LC space. The single IMS-MS data frame was exported from IM-MS Browser as a.CSV file consisting of drift time and *m/z* bins as a two-dimensional (2D) matrix with corresponding peak areas as the observed variable. These data frames were then exported into R (v4.2.1) for further analysis. Initial steps in R included binning data to create coordinate windows, or summing together signals with similar *m/z* and drift time values, normalizing these abundances, and filtering out spurious coordinates. Selection of unique *m/z* and drift time coordinates that distinguished phenotypes was performed on the training data. These coordinates were then used to build a sample classification model, whose final performance was assessed using the testing data.

LC-Summation and Binning Process.

Each .d file acquired using LC-IMS-MS was opened in IM-MS Browser and the LC dimension was integrated over the entire scan period (16.5 min for PFAS analysis and 38.5 min for lipid profiling) and the summed IMS-MS spectra for each datafile was exported as a .CSV file for subsequent analysis. Though the entirety of the chromatographic profile was summed for these data sets, selection of a predefined window of elution is also possible to prohibit integration of the void volume or column wash and focus on a specific biomolecular class with known elution window as desired. The time required for summation of each PFAS datafile was ~3 min, whereas the larger lipid datafiles took ~10 min to process. Agilent's IM Browser currently does not utilize multiple processor threads, though implementation of this capability could increase the temporal throughput by more than an order of magnitude dependent on the computational capabilities of the user platform. To ensure that individual samples had consistent labels, *m/z* and drift time bins were created. The first step of this process was to limit the *m/z* and drift time values to relevant ranges for our molecules of interest. For the passive sampler data, the *m/z* range was limited to values ≥100 and ≤1000, while the *m/z* range for the pregnancy data was limited to values ≥200 and ≤1200. To determine the appropriate minimum drift time cutoff, the "Find Features" function from IM-MS Browser was used to identify spectral features. From these lists, it was determined that the minimum drift time corresponding to a real feature was approximately 13 ms (ms) for both data sets, and thus this value was used as the cutoff. Once the ranges were chosen, the *m/z* values were binned by summing together *m/z* values that were within 2 Da, and drift times were binned by summing values within 0.5 ms. Implementation of this binning approach affords us the ability to break continuous data into discrete coordinates that retain most relevant feature information while being amenable to use in popular machine learning and statistical approaches. Although it should be noted that the choice of binning parameters can impact downstream results, as too wide of bins may result in multiple features being combined which may confound statistical significance, while too narrow of bins may result in a feature being split across multiple bins. Determining if this has occurred and to what extent can be done following feature selection by examining the coordinates in the raw data, and can be alleviated by modifying the bin sizes, which can be achieved readily by the end-user in the code provided. To demonstrate this, we also demonstrate predictive results with a 1 Da and 0.5 ms drift time bins with the passive sampler data, which are shown in the Supporting Information. However, in cases where comprehensively identifying all significant features is important, this approach may not be suitable.

Split Data.

To assess our screening method, samples were randomly split into training and testing sets using a 3:1 ratio using the function “createDataPartition” from the caret package.[37] With this approach, training samples were used to determine the appropriate normalization and filtering parameters, pick out coordinates of interest, and construct predictive models, while the testing data were withheld for final model evaluation.

Normalize and Filter.

To determine the appropriate normalization and filtering steps, data distributions were first assessed, and transformations were selected based on the training data, which were then subsequently applied to the testing data. Following binning, unique *m/z* and drift time value combinations were formed to represent coordinates on the original IMS-MS spectra. For the passive sampler data, these values were first normalized by dividing by the mass of the sampler, and then coordinate values for a passive sampler not deployed were subtracted (blank) to minimize signal associated with the sampler itself. Negative values resulting from this subtraction were set to zero. Normality of the abundances across coordinates in the training data for both data sets were then assessed by looking at density plots, which revealed a consistent, strong right skew. To alleviate this skew, abundances were $\log_2$ transformed. Once coordinates were normalized, they were next filtered based on abundance. As coordinates with small abundances are more likely to contain noise rather than actual features, removing low abundance coordinates prior to model construction was thought to increase the likelihood of pinpointing relevant features. For the passive sampler data, this was achieved by identifying coordinates whose median abundance was in the lowest 25% of the training data, and then removing these coordinates from both data splits. Given the IMS-MS region for halogenated molecules tends to be less noisy than the biological region, a stricter filter was used for the lipidomic data, and thus coordinates whose median abundance was in the lowest 75% of the training data were removed to avoid any noise and low level features.[23] After filtering, sample outliers were assessed using principal components analysis (PCA) and hierarchical clustering with a Euclidean distance metric. Doing so revealed one sample in the pregnancy data had outlying data distributions in both PCA and clustering, and thus was excluded from downstream analysis.

Select Coordinates.

Before building predictive models, distinguishing coordinates between phenotypes were identified using Lasso logistic regression. While other feature selection methods could also be implemented here, we opted for Lasso as it results in a sparse model focused on the most relevant coordinates. This analysis was conducted in R using the glmnet package, with phenotypes as dependent variables and coordinates as independent variables.[38] The optimal regularization strength (lambda) was determined via 5-fold cross-validation. Stability of selected coordinates was enhanced by combining Lasso with bootstrapping.[39] This involved generating 1000 bootstrap samples, each equal in size and created by sampling with replacement from the training data, conducting Lasso on each training set, and retaining features selected in at least 200 of the 1000 trials.

Construct Model.

After coordinate selection, classification models were developed to predict the underlying phenotype associated with each sample. This was done using the selected coordinates from the training data using a support vector machine (SVM) with a linear kernel and $C = 1$ and implemented using the package "caret" in R.[37] To assess performance, 5-fold cross-validation was performed on the training data, utilizing metrics such as accuracy, sensitivity, specificity, and the kappa statistic.

Evaluate Model and Annotate Features.

To evaluate the final model, classification was performed on the withheld testing data that was not used during coordinate selection or model construction. Predicted phenotype labels were then compared to the true sample labels, and model accuracy, sensitivity, specificity, and the kappa statistic were calculated for each comparison. To determine the relevancy of input coordinates, importance scores were calculated using the function "varImp" from the caret package, which considers the coefficients assigned to each coordinate in the SVM model.[37] Selected coordinates were then mapped on to summed IMS-MS spectra using the package ggplot2, with open circles highlighting the regions of interest. Abundances of these coordinates for different phenotypes were also visualized using boxplots. Feature annotations were performed using the software Skyline following the procedures outlined by Kirkwood-Donelson et al.[6,30] Briefly, for each coordinate defined as a significant *m/z* and drift time bin highlighted by the developed model, an extracted ion chromatogram (EIC) was obtained to characterize the signal's retention time and precursor mass. While it is possible for co-binning of signals to generate multiple "peaks" in the EIC, generally the bulk abundance of the signal could be traced to individual MS signals, and this occurrence could be reduced by using smaller *m/z* and drift time bin widths as appropriate on a per-instrument basis. Signals of interest were compared against pre-existing lipid and PFAS libraries wherein accurate mass, retention time, CCS and fragmentation spectra when available were utilized to aid in molecular annotation on a per coordinate basis.

Annotated Model Construction.

Given that the data sets used in this study have already been explored through a conventional NTA approach,[6,35] there was interest in evaluating how the outlined screening models, would fare against models that depend solely on annotated features and their abundances. To assess this, a modified version of the model building process described above was repeated using the molecular annotations provided by Odenkirk et al. for the lipidomic data and Dodds et al. for the passive sampler data.[35,40] For the lipidomic data, total ion chromatogram (TIC) normalization and $\log_2$ transformation was performed, while the peak areas in the passive sampler data were divided by the mass of resin collected. Since the annotated data was already preprocessed, the classification pipeline started by splitting samples into training and testing sets, which consisted of the same samples used for the screening models to ensure consistency. Important molecules were then selected using the described bootstrapped Lasso, and finally leveraged into a SVM classification model for comparison.

RESULTS & DISCUSSION

The summed IMS-MS screening method in this work was developed to facilitate rapid sample classification and expedite feature annotation in LC-IMS-MS data. While traditional approaches characterize differences between sample groups by first performing feature finding followed by statistical analysis, our approach reverses this order by first finding statistical differences in two-dimensional (2D) sample summed IMS-MS spectra, verifying these differences through the assessment of classification models, and then assessing resulting coordinates to see if they retain relevant features. The constructed classification models can be readily applied to new samples and application types, allowing for immediate assessments when sample groupings are prioritized over understanding specific molecular compositions. We believe these advancements will considerably improve the rate at which NTA data analysis can be performed. Additional discussion of the details of this method can be found in the Supporting Information.

Model Evaluation.

To evaluate our proposed workflow, we applied our summed IMS-MS screening to two distinct data sets illustrating both environmental and clinical applications. The first study used exposomic NTA data gathered from passive aquatic samplers deployed upstream and downstream of a fluorochemical manufacturer along North Carolina's Cape Fear River (Figure 1A). Given the prior documentation of PFAS contaminants in this river due to industrial activities,[41-43] we anticipated that samples taken from the different locations would have variable molecular profiles. The second study involved lipidomic NTA data from blood plasma samples collected from either control or preeclamptic pregnant individuals on their day of delivery (Figure 1B). Previous work by Odenkirk et al. characterized differences in lipidomic profiles between these two phenotypes, leading us to hypothesize that differences would be present in the summed IMS-MS spectra.[35] These two data sets not only represent distinct fields of application—environmental science and clinical medicine—but also encompass two primary molecule types: PFAS and lipids. Therefore, our results demonstrate that this approach is not specific to certain domains, but instead can be applied to various sample types for which there are discernible molecular differences.

To determine the efficacy of our classification approach on the described passive sampler and pregnancy data sets, we initially conducted a 5-fold cross-validation on the training data. For each fold, standard performance metrics such as accuracy, sensitivity, and specificity were calculated, and their averages were used to assess the overall performance. Due to the class imbalance present in both data sets (Figure 1), we also calculated Cohen's Kappa statistic, which accounts for the possibility of random chance in the classification process, providing a more accurate reflection of the model's predictive accuracy.[44] After assessing the training data, we applied our model to the withheld testing data to further validate its accuracy. When doing so, the model for the passive sampler data had 100% accuracy in both data splits, indicating that it could perfectly distinguish between upstream and downstream samples (Table 1A, **left**). This accuracy raised concerns about possible overfitting. To investigate this, we examined the abundances of the influential coordinates to see if model output could be traced back to the training data as shown in the boxplots in

Figure 3A. In these plots, it is evident that the selected coordinates show higher abundances in downstream training samples, with many upstream samples exhibiting nondetects for these coordinates. The most pronounced difference between the upstream and downstream samples was observed in coordinate (636–638, 24–24.5) (*m/z* value, drift time). Here, the abundance distributions for the upstream and downstream samples are entirely distinct, thereby reinforcing the model's discriminatory power. An example of this difference is shown for select sample IMS-MS spectra in Figure S1A, which illustrates that the abundance of this coordinate downstream is much higher than the representative coordinate in the upstream sample, where it is not detected.

Using the summed IMS-MS screening method for the pregnancy data, both the positive and negative mode ionization data were evaluated. However, differences were only observed in the positive mode data. This observation is consistent with the data set's previous characterization by Odenkirk et al., wherein most lipids signals were observed in positive ion polarity vs negative mode (221 analytes vs 99). As with the passive sampler data, high accuracy was achieved in the training and testing data (Table 1A, **right**). In both the training and testing data, the specificity was lower than the sensitivity, suggesting that the model is more adept at identifying true positives (PRE samples) than true negatives (control samples). Distributions of the selected coordinates are shown in Figure 3B, wherein the discrepancies between phenotypes are apparent. This can be seen clearest in the coordinates with the highest model weights, such as, (848–850, 36.5–37) and (1080–1082, 42.5–43), where the interquartile ranges do not overlap. Figure S1B visualizes these differences for select summed IMS-MS spectra. In contrast with the passive sampler coordinates, there are no nondetections in the pregnancy lipidomic data. This suggests that while certain coordinates are distinctly indicative of either upstream or downstream samples in the passive sampler data, the variations in the pregnancy data set are more subtle, which may contribute to why more coordinates were selected. This pattern matches what we would anticipate, as it is expected that specific PFAS contamination would be localized to downstream of a point source, while it is not expected that specific lipids would only be expressed in one phenotype.

In traditional NTA data pipelines, feature finding is performed, statistical methods are applied to determine which molecules differentiate phenotypes or groups of interest, and feature annotation is performed. Prior to development of our outlined screening method, the data sets presented in this study were analyzed following this standard approach.[35] In both omic studies, molecules that differentiated phenotypes were identified, however the predictive power of these differences was not assessed. Therefore, in this work we also utilize the previously published annotated data to construct classification models to compare performance to the summed IMS-MS screening model. Examining these results for the passive sampler data (Table 1B, **left**), we see that the classifications are comparable to the screening model. While the presented accuracy and associated metrics are slightly lower in the training data, there is equivalent performance on the testing data. This is once again expected, as several of the identified compounds were primarily detected downstream and remained absent upstream. Interestingly, the classification model for the pregnancy data using annotated data performed considerably worse (Table 1B, **right**), with a 25% reduction

in accuracy, 16.6% reduction in sensitivity, and a 40% reduction in specificity. While the exact source of this discrepancy is unclear, one potential reason for this difference is gaps that exist in the present lipid annotations or databases. Due to the highly isomeric nature of lipids, as well as their diverse physicochemical properties, it is highly unlikely that all lipids that are detected in a given sample are able to be identified.[45] Therefore, an additional advantage to classification with our summed IMS-MS screening approach is that it utilizes the complete feature space, and consequently may incorporate differential signals or features that are typically lost in the annotation process. Another potential reason for this performance gap may relate to the preprocessing of the data prior to model construction. While formation of the assessed coordinates followed the described binning and filtering procedure, normalized lipid data was used directly from Odenkirk et al. Differences in this initial handling may have impacted downstream model performance.

Coordinate Assessment.

When performing coordinate assessment of the summed IMS-MS screening approach, a total of 4 coordinates for the PFAS passive sampler data and 19 for the lipid pregnancy data were deemed important as highlighted in Figure 3. Each of the coordinates were profiled across the most strongly correlated samples. The coordinate binning process of each summed IMS-MS frame is illustrated in Figure S2, wherein the dashed gray and orange lines are indicative of the applied *m/z* and drift time bins for coordinate 4 of the passive sampler data. Based on the resolving power for the current IMS-IMS platform (Agilent 6560, characterized in detail in previous manuscripts),[46,47] each ion's drift time distribution is often ~0.5 ms wide at full width at half-maximum (fwhm). This drift time bin width could be adjusted for different IMS platforms possessing lower or higher resolving power as required by the end user. While narrower drift time bins generate more specific coordinates for data profiling, if the applied bin width is too narrow, signal abundance for each feature may be split across multiple bins. The ability to adjust the desired bin size is demonstrated in the Supporting Information, where results from using a *m/z* bin of 1 Da and drift time bin of 0.5 ms on the passive sampler data are shown (Figure S2). With this adjustment, equivalent predictive performance was achieved, with both unique and shared coordinates selected compared to the wider *m/z* bin.

Upon evaluation of the passive samplers up- and downstream of the chemical manufacturer using the outlined selective criteria, 4 coordinates were highlighted for the passive samplers as shown in Figures 3A and S1A. The feature descriptors did not match any entries in an in-house library of >100 PFAS with known *m/z*, retention time, and CCS values,[23] and their masses did not return matches using extensive public PFAS databases.[48,49] Figure S4 displays the feature CCS, retention time, and mass defect values plotted against their respective *m/z* values. The 57 PFAS detected in the same downstream passive sampler data set previously are also plotted to demonstrate the expected PFAS-specific trends.[6] PFAS have been shown to have distinctly low CCS values as well as low or negative mass defect due to the prevalence of fluorine.[6,23] The features pinpointed by the highlighted coordinates align well with the previously reported PFAS (Figure S4A-C), indicating that these features did arise from highly fluorinated chemicals. For example, coordinate (886–888, 28–28.5) corresponded to *m/z* 886.85, which had a larger mass defect than the other PFAS. However,

it is likely this feature is a gas-phase dimer as they are commonly observed in PFAS data. Specifically, several were reported by Kirkwood-Donelson et al., including two isobars of *m/z* 886.85, which were dimers of PFO5DoA (Figure S4) and Nafion byproduct 1.[6] The corresponding monomer is also plotted in Figure S4 and aligns well with the other PFAS. In the previous study from Kirkwood-Donelson et al., 8 of the detected PFAS were previously unreported chemical structures, and 11 were assigned putative molecular formulas but had unknown chemical structures. Thus, given the presence of PFAS signatures but absence of suspect screening matches, these features likely also correspond to unknown PFAS produced by the same fluorochemical manufacturer. Further evaluation of these features to identify their chemical structures is therefore warranted as they could be novel PFAS which have not been previously observed or annotated.

In the pregnancy study, comparison of the control and PRE samples generated 19 significant coordinates as shown in the representative summed IMS-MS spectra (Figure 4). Non-targeted lipidomics data for the samples was collected using an “alternating frames” method, wherein precursor and fragmentation data is collected by alternating MS^1, MS2, MS^1,…[35,50] In this workflow, half of the IMS-MS frames represent fragmentation spectra and half precursor data. Hence it is not surprising that 7 of the 19 highlighted coordinates arose from fragments produced during collision-induced dissociation (CID). Of these fragment coordinates, several were identified as the neutral loss of a single fatty acyl (FA) chain from specific complex lipids. These include coordinate (646–648, 40–40.5) corresponding to TG 20:4_20:4_X, coordinate (478–480, 20.5–21) corresponding to PC 16:0_X, and coordinate (308–310, 39.5–40) corresponding to PC O-18:0_X, where X is any FA. The ability to evaluate fragments shared by multiple structurally related lipids is a key advantage of this approach, especially given the importance of FA components for enzyme specificity and lipid function.[51-53] Here, it can be concluded that several or all TGs with at least two 20:4 FAs, PCs with at least one 16:0 FA, and plasmalogen PCs with an O-18:0 FA are significantly altered in PRE samples relative to control. Figure 4B displays the EIC of the *m/z* value corresponding to M-HG-X for the PC O-18:0/X $[M + Na]^+$ in representative PRE and control samples. In this example, there are several appreciable signals for this fragment corresponding to multiple unique plasmalogen PCs containing O-18:0 FAs in the PRE sample, whereas this signal is among the noise in the control sample. Moreover, previous work using a traditional lipidomics approach on the same PRE samples found that PC 16:0/16:0, PC 16:0 18:1, PC 16:0_20:4, and PC 16:0_22:6 were all significantly upregulated in PRE samples compared to control.[35] Thus, the coordinate (308–310, 39.5–40) likely encompasses all these changes as the shared fragment representing the neutral loss of FA 16:0, 18:1, 20:4, 22:6, and potentially others. These conclusions would be difficult to draw using traditional lipidomics approaches, where complete FA assignment followed by comprehensive evaluation of those FA components within each class can be challenging. While the structural-based connectivity and omic phenotype evaluations (SCOPE) software toolbox includes the ability to perform such FA evaluations, confident annotations with appropriate lipid speciation are required, which is not the case using this approach.[54]

Multiple coordinates from the pregnancy data found to arise from precursor signals corresponded to adjacent heavier ^{13}C isotope peaks within a single isotopic envelope,

i.e., the M + 1, M + 2, M + 3 or M + 4 peaks, rather than the precursor ion signal itself. One such example is coordinate (528–530, 30.5–31) which corresponds to the M + 4 isotopic peak for protonated LPC 18:0 (*m/z* 524.374). This represents another unique aspect of this approach, as it may be able to evaluate and correct for the oversaturation of abundant lipid species. A well-known challenge in lipidomics is the large dynamic range of common lipid samples such as plasma, which can lead to oversaturation when using common time-of-flight instruments with an analog-to-digital (ADC) converter subject to detector saturation. Some researchers perform multiple measurements of the same sample at low and high concentrations to avoid oversaturation of highly abundant lipids such as phosphati-dylcholines (PCs) without eliminating lower abundance lipid classes. Here, the precursor ion signal for protonated LPC 18:0 is visually oversaturated, which may lead to the conclusion that it is not statistically significant between groups. Thus, the lower abundance M + 4 isotopic peak may represent abundance differences more accurately than the previous oversaturated isotopic peaks. This has been demonstrated previously by Bilbao and co-workers, where the M + 2 isotope peak of PC(16:0/16:0) was used to correct the abundance of this oversaturated lipid posthoc.[55] Other tentatively identified precursors include multiple oxidized lipids, such as the acylcarnitines AC 16:1;O (414–416, 26–26.5) and AC 20:1;O3 (502–504, 29–29.5) among others (Table S1). The assessment of oxidized lipids is also extremely important as they are relatively understudied in traditional lipidomics analyses, primarily due to a lack of analytical standards and absence in commonly searched databases, although they are recognized as markers of oxidative stress.[56]

CONCLUSIONS

The objective of this study was to develop a method to reduce the complexity of NTA data collected with a multidimensional LC-IMS-MS platform and screen for phenotypic differences. While the method parameters used in this study were able to differentiate the phenotypes investigated, future modifications may prove necessary to delineate outcomes of interest. Some changes may include additional adjustments of bin sizes for other IMS-MS platforms to account for variations in resolving power and selectivity, or replacing the SVM with other classification models, such as a random forest. Following the classification modeling, there were also several challenges associated with definitively linking coordinates with fully annotated molecular structures in the PFAS data. However, the molecular descriptors for the features within the coordinates of interest support their identification as highly fluorinated compounds. Therefore, these coordinates may represent unknown PFAS which have not been annotated to date. In the lipidomic data, several molecular annotations were made across the 19 selected coordinates. Among these was the identification of neutral losses of a single fatty acyl (FA) chain from specific complex lipids, which showcased the ability of this method to identify fragments shared by multiple structurally related lipids, which may be missed in existing pipelines. Other annotations included precursor signals corresponding to adjacent ^{13}C isotope peaks, which may allow for the detection of lipid differences between phenotypes that are often missed due to saturation of high concentration species. Thus, the showcased rapid data analyses, potential localization of novel PFAS, ability to find fragments of interest and assess saturated lipids, makes the summed IMS-MS screening approach of great value to the NTA scientific community.

The molecules examined in this work, lipids and PFAS, were ideal for the outlined approach due to their strong mass versus mobility relationships, resulting in readily distinguishable summed IMS-MS spectra. Other molecule types, such as carbohydrates or peptides, may contain more overlap in this space, making samples harder to discern. However, we believe the outlined pipeline functions as a proof of concept and can readily be applied to assess other data dimensions. For example, 2D LC-MS or GC-MS spectra could be used as input for this workflow in cases where the use of IMS is not appropriate or feasible. Future work will include determining the utility of our approach on different data dimensions, as well as incorporating molecular descriptors not included in the raw instrument output, such as mass defect.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was funded by grants from the National Institute of Environmental Health Sciences (P42 ES027704 and P42 ES031009), the National Institute of General Medical Sciences (R01 GM141277), the National Institute of Health National Cancer Institute (R33 CA229068), and a cooperative agreement with the Environmental Protection Agency (STAR RD 84003201). This research was supported [in part] by the Intramural Research Program of the NIH (ZIC ES103363). The views expressed in this manuscript do not reflect those of the funding agencies.

Data Availability Statement

Source data and code required for statistical analysis and figure generation have been deposited in GitHub and are available at https://github.com/BakerLabMS/IMS_Screening.

REFERENCES

(1). Banerjee S. ACS Omega 2020, 5 (5), 2041–2048. [PubMed: 32064364]

(2). Guo L; Milburn MV; Ryals JA; Lonergan SC; Mitchell MW; Wulff JE; Alexander DC; Evans AM; Bridgewater B; Miller L; et al. Proc. Natl. Acad. Sci. U. S. A 2015, 112 (35), E4901–E4910. [PubMed: 26283345]

(3). Wu X; Wang Z; Luo L; Shu D; Wang K Front. Med. Technol 2023, 4, No. 1065506. [PubMed: 36688143]

(4). Molloy MP; Hill C; O'Rourke MB; Chandra J; Steffen P; McKay MJ; Pascovici D; Herbert BR J. Proteome Res 2022, 21 (4), 1196–1203. [PubMed: 35166117]

(5). Chappel JR; King ME; Fleming J; Eberlin LS; Reif DM; Baker ES Anal. Chem 2023, 95 (34), 12913–12922. [PubMed: 37579019]

(6). Kirkwood-Donelson KI; Dodds JN; Schnetzer A; Hall N; Baker ES Sci. Adv 2023, 9 (43), No. eadj7048. [PubMed: 37878714]

(7). Kirkwood KI; Fleming J; Nguyen H; Reif DM; Baker ES; Belcher SM Environ. Sci. Technol 2022, 56 (6), 3441–3451. [PubMed: 35175744]

(8). Dodds JN; Alexander NLM; Kirkwood KI; Foster MR; Hopkins ZR; Knappe DRU; Baker ES Anal. Chem 2021, 93, 641–656. [PubMed: 33136371]

(9). Hollender J; Schymanski EL; Singer HP; Ferguson PL Environ. Sci. Technol 2017, 51 (20), 11505–11512. [PubMed: 28877430]

(10). Kirkwood-Donelson KI; Chappel J; Tobin E; Dodds JN; Reif DM; DeWitt JC; Baker ES Chemosphere 2024, 354, No. 141654. [PubMed: 38462188]

(11). Ahuja V; Singh A; Paul D; Dasgupta D; Urajová P; Ghosh S; Singh R; Sahoo G; Ewe D; Saurav K Chem. Res. Toxicol 2023, 36 (12), 1834–1863. [PubMed: 38059476]

(12). Dueñas ME; Peltier-Heap RE; Leveridge M; Annan RS; Büttner FH; Trost M EMBO Mol. Med 2023, 15 (1), No. e14850. [PubMed: 36515561]

(13). Manz KE; Feerick A; Braun JM; Feng YL; Hall A; Koelmel J; Manzano C; Newton SR; Pennell KD; Place BJ; et al. J. Expo Sci. Environ. Epidemiol 2023, 33 (4), 524–536. [PubMed: 37380877]

(14). Cajka T; Fiehn O Anal. Chem 2016, 88 (1), 524–545. [PubMed: 26637011]

(15). Chen L; Zhong F; Zhu J Metabolites 2020, 10 (9), 348. [PubMed: 32867165]

(16). Dettmer K; Aronov PA; Hammock BD Mass Spectrom. Rev 2007, 26 (1), 51–78. [PubMed: 16921475]

(17). Zheng X; Wojcik R; Zhang X; Ibrahim YM; Burnum-Johnson KE; Orton DJ; Monroe ME; Moore RJ; Smith RD; Baker ES Annu. Rev. Anal. Chem 2017, 10 (1), 71–92.

(18). Baker PRS; Armando AM; Campbell JL; Quehenberger O; Dennis EA J. Lipid Res 2014, 55 (11), 2432–2442. [PubMed: 25225680]

(19). Glish GL; Vachet RW Nat. Rev. Drug Discov 2003, 2 (2), 140–150. [PubMed: 12563305]

(20). Xiao JF; Zhou B; Ressom HW TrAC Trends Analyt. Chem 2012, 32, 1–14.

(21). Dunn WB; Erban A; Weber RJM; Creek DJ; Brown M; Breitling R; Hankemeier T; Goodacre R; Neumann S; Kopka J; et al. Metabolomics 2013, 9 (S1), 44–66.

(22). Zedda M; Zwiener C Anal. Bioanal. Chem 2012, 403 (9), 2493–2502. [PubMed: 22476785]

(23). Foster M; Rainey M; Watson C; Dodds JN; Kirkwood KI; Fernandez FM; Baker ES Environ. Sci. Technol 2022, 56 (12), 9133–9143. [PubMed: 35653285]

(24). McCord JP; Groff LC; Sobus JR Environ. Int 2022, 158, No. 107011. [PubMed: 35386928]

(25). Schymanski EL; Jeon J; Gulde R; Fenner K; Ruff M; Singer HP; Hollender J Environ. Sci. Technol 2014, 48 (4), 2097–2098. [PubMed: 24476540]

(26). Chappel JR; Kirkwood-Donelson KI; Reif DM; Baker ES Anal. Bioanal. Chem 2023, 416, 2189–2202. [PubMed: 37875675]

(27). Dodds JN; Baker ES J. Am. Soc. Mass Spectrom 2019, 30 (11), 2185–2195. [PubMed: 31493234]

(28). May JC; McLean JA Anal. Chem 2015, 87 (3), 1422–1436. [PubMed: 25526595]

(29). Dodds JN; Wang L; Patti GJ; Baker ES Anal. Chem 2022, 94 (5), 2527–2535. [PubMed: 35089687]

(30). Kirkwood KI; Pratt BS; Shulman N; Tamura K; MacCoss MJ; MacLean BX; Baker ES Nat. Protoc 2022, 17 (11), 2415–2430. [PubMed: 35831612]

(31). Zang X; Monge ME; Gaul DA; Fernández FM Anal. Chem 2018, 90 (22), 13767–13774. [PubMed: 30379062]

(32). Annesley TM Clin. Chem 2003, 49 (7), 1041–1044. [PubMed: 12816898]

(33). Schmid R; Heuckeroth S; Korf A; Smirnov A; Myers O; Dyrlund TS; Bushuiev R; Murray KJ; Hoffmann N; Lu M; et al. Nat. Biotechnol 2023, 41 (4), 447–449. [PubMed: 36859716]

(34). Crowell KL; Slysz GW; Baker ES; LaMarche BL; Monroe ME; Ibrahim YM; Payne SH; Anderson GA; Smith RD Bioinformatics 2013, 29 (21), 2804–2805. [PubMed: 24008421]

(35). Odenkirk MT; Stratton KG; Gritsenko MA; Bramer LM; Webb-Robertson B-JM; Bloodsworth KJ; Weitz KK; Lipton AK; Monroe ME; Ash JR; et al. Mol. Omics 2020, 16 (6), 521–532. [PubMed: 32966491]

(36). Kirkwood KI; Christopher MW; Burgess JL; Littau SR; Foster K; Richey K; Pratt BS; Shulman N; Tamura K; MacCoss MJ; et al. J. Proteome Res 2022, 21 (1), 232–242. [PubMed: 34874736]

(37). Kuhn M. Journal of Statistical Software 2008, 28 (5), 1–26. [PubMed: 27774042]

(38). Friedman J; Hastie T; Tibshirani R J. Stat. Softw 2010, 33 (1), 1–22. [PubMed: 20808728]

(39). Laurin C; Boomsma D; Lubke G Stat. Appl. Genet. Mol. Biol 2016, 15 (4), 305–320. [PubMed: 27248122]

(40). Dodds JN; Kirkwood-Donelson KI; Boatman AK; Knappe DRU; Hall NS; Schnetzer A; Baker ES Sci. Total Environ 2024, 947, No. 174574. [PubMed: 38981548]

(41). Hopkins ZR; Sun M; DeWitt JC; Knappe DRU J. AWWA 2018, 110 (7), 13–28.

(42). Sun M; Arevalo E; Strynar M; Lindstrom A; Richardson M; Kearns B; Pickett A; Smith C; Knappe DRU Environmental Science & Technology Letters 2016, 3 (12), 415–419.

(43). Pétré MA; Salk KR; Stapleton HM; Ferguson PL; Tait G; Obenour DR; Knappe DRU; Genereux DP Science of The Total Environment 2022, 831, No. 154763. [PubMed: 35339537]

(44). McHugh ML Biochem Med. (Zagreb) 2012, 22 (3), 276–282. [PubMed: 23092060]

(45). Olshansky G; Giles C; Salim A; Meikle PJ Prog. Lipid Res 2022, 87, No. 101177. [PubMed: 35780914]

(46). Stow SM; Causon TJ; Zheng X; Kurulugama RT; Mairinger T; May JC; Rennie EE; Baker ES; Smith RD; Mclean JA; et al. Anal. Chem 2017, 89 (17), 9048–9055. [PubMed: 28763190]

(47). May JC; Dodds JN; Kurulugama RT; Stafford GC; Fjeldsted JC; McLean JA Analyst 2015, 140 (20), 6824–6833. [PubMed: 26191544]

(48). Mohammed Taha H; Aalizadeh R; Alygizakis N; Antignac JP; Arp HPH; Bade R; Baker N; Belova L; Bijlsma L; Bolton EE; et al. Environ. Sci. Eur 2022, 34 (1), 104. [PubMed: 36284750]

(49). McEachran AD; Sobus JR; Williams AJ Anal. Bioanal. Chem 2017, 409 (7), 1729–1735. [PubMed: 27987027]

(50). Tötsch K; Fjeldsted JC; Stow SM; Schmitz OJ; Meckelmann SW J. Am. Soc. Mass Spectrom 2021, 32 (10), 2592–2603. [PubMed: 34515480]

(51). Begum H; Li B; Shui G; Cazenave-Gassiot A; Soong R; Ong RT; Little P; Teo YY; Wenk MR Sci. Rep 2016, 6, No. 19139. [PubMed: 26743939]

(52). Chua EC; Shui G; Lee IT; Lau P; Tan LC; Yeo SC; Lam BD; Bulchand S; Summers SA; Puvanendran K; et al. Proc. Natl. Acad. Sci. U. S. A 2013, 110 (35), 14468–14473. [PubMed: 23946426]

(53). Li J; Condello S; Thomes-Pepin J; Ma X; Xia Y; Hurley TD; Matei D; Cheng JX Cell Stem Cell 2017, 20 (3), 303–314.e5. [PubMed: 28041894]

(54). Odenkirk MT; Zin PPK; Ash JR; Reif DM; Fourches D; Baker ES Analyst 2020, 145 (22), 7197–7209. [PubMed: 33094747]

(55). Bilbao A; Gibbons BC; Slysz GW; Crowell KL; Monroe ME; Ibrahim YM; Smith RD; Payne SH; Baker ESInt. J. Mass Spectrom 2018, 427, 91–99. [PubMed: 29706793]

(56). Wölk M; Prabutzki P; Fedorova M Acc. Chem. Res 2023, 56 (7), 835–845. [PubMed: 36943749]

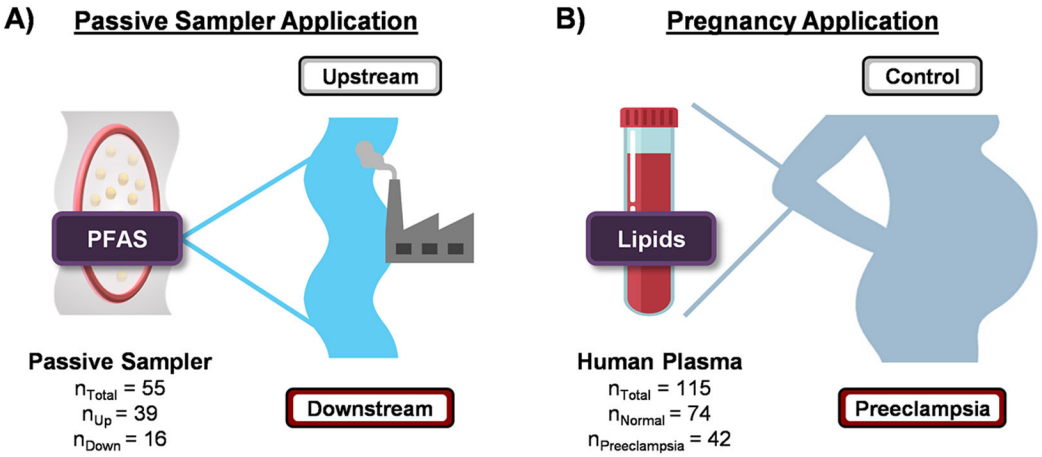


Figure 1.

Overview of data sets evaluated. (A) Passive aquatic samplers were deployed for ~2 weeks in the Cape Fear River of North Carolina either upstream or downstream from a fluorochemical manufacturer to monitor differences in water contamination. (B) Blood samples were taken from control or preeclamptic individuals near the time of birth to assess differences in lipid profiles.

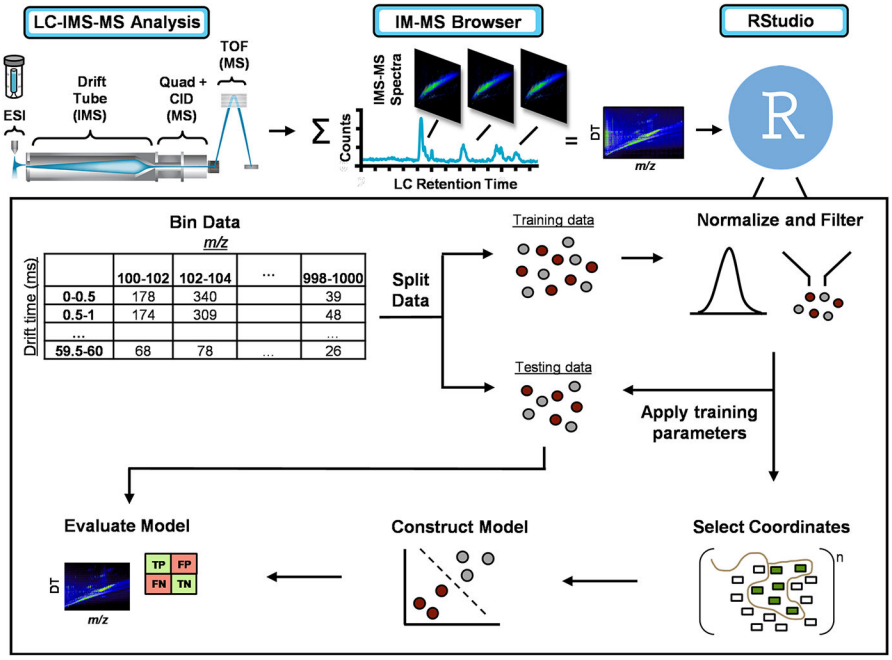


Figure 2.

Overview of IMS-MS screening workflow. Data are first collected on an LC-IMS-MS platform. The LC dimension is then collapsed by summing the individual IMS-MS spectra across all retention times, resulting in a single IMS-MS spectrum for each sample. The summed spectra are imported into R where the *m/z* values and drift times are binned into consistent intervals. Data are subsequently split into training and testing sets at the sample level, and normalization and filtering are performed based on the distribution of the training data before being applied to the testing data. Following all data cleaning steps, relevant coordinates are selected from the training data using a bootstrapped Lasso, which are then used to create classification models that are evaluated on the testing data.

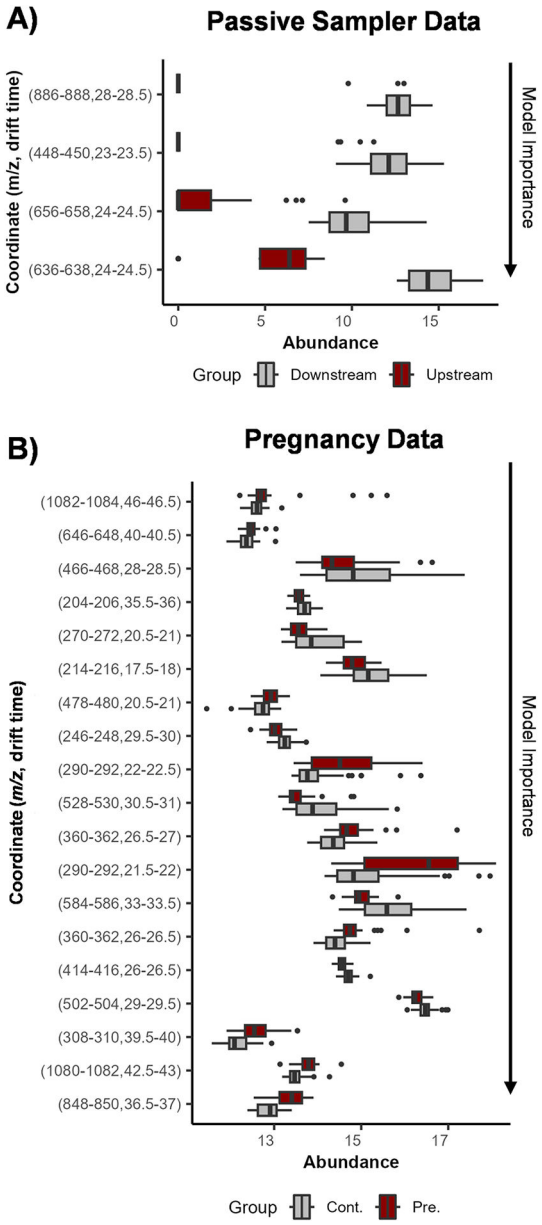


Figure 3.

Abundance distributions for selected training sample coordinates shown as (*m/z* range (Da) and drift time range (ms)) for the two (A) passive sampler and (B) pregnancy groups. Individual box and whiskers are ordered by their influence on their respective classification model, with box and whiskers closer to the bottom carrying more weight.

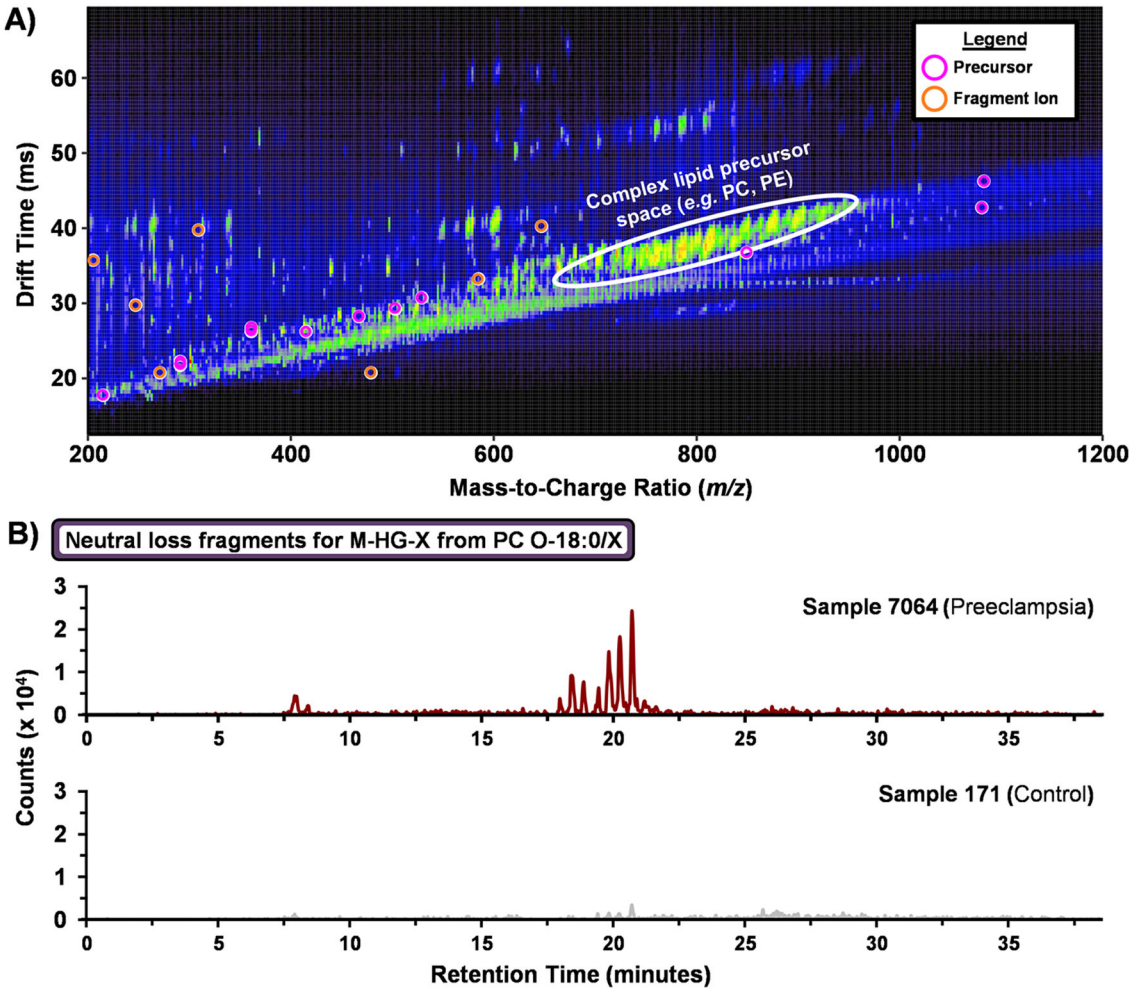


Figure 4.

(A) Summed IMS-MS spectrum highlighting coordinate regions indicated as important in the developed model. Highlighted coordinates circled in pink are indicative of precursor signals, while those in orange correspond to fragment ions generated by collision-induced dissociation. (B) An extracted ion chromatogram of *m/z* 309.28 from coordinate (308–310 *m/z*, 39.5–40 ms) corresponds to the neutral loss fragment of M-HG-X for PC O-18:0/X $[M + Na]^+$ where X is any FA for a representative preeclampsia versus control sample.

Table 1.

Results of Classification Models[a]

A) IMS-MS Screening Model Results

Passive Sampler Model			Pregnancy Model		
	Training	Testing		Training	Testing
# Samples	42	13	# Samples	87	28
Accuracy (%)	100	100	Accuracy (%)	90.8	92.9
Sensitivity (%)	100	100	Sensitivity (%)	92.7	94.4
Specificity (%)	100	100	Specificity (%)	87.1	90.0
Kappa	1	1	Kappa	0.80	0.84

B) Annotated Model Results

Passive Sampler Model			Pregnancy Model		
	Training	Testing		Training	Testing
# Samples	42	13	# Samples	87	28
Accuracy (%)	93.1	100	Accuracy (%)	72.4	67.9
Sensitivity (%)	86.7	100	Sensitivity (%)	85.5	77.8
Specificity (%)	96.7	100	Specificity (%)	49.5	50.0
Kappa	82.2	1	Kappa	0.37	0.28

[a](A) shows results using our IMS-MS screening method for passive sampler (left) and pregnancy data (right), while (B) shows results using annotated feature data. Metrics for the training data were calculated using 5-fold cross-validation.