Supporting Information:

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

Jessie R. Chappel[1], Kaylie I. Kirkwood-Donelson[2], James N. Dodds[3], Jonathon Fleming[1], David M. Reif[4*], Erin S. Baker[3*]

[1] Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, NC 27606, USA

[2]Immunity, Inflammation, and Disease Laboratory, National Institute of Environmental Health Sciences, Durham, NC 27709, USA

[3]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

[4]Predictive Toxicology Branch, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Durham, NC 27709, USA

*** Correspondence:**
Co-corresponding Authors
david.reif@nih.gov; erinmsb@unc.edu

Table of Contents

Dataset descriptions:

Passive Sampler Data Collection:

Passive samplers (n=55) were deployed during the summer and fall months of 2016 and 2017 in the Cape Fear River of North Carolina as previously described.[1] This waterway was known to possess widespread PFAS contamination during this period as a result of local fluorochemical manufacturing and corresponding outfall.[2-4] Passive samplers were deployed both upstream and downstream of the chemical manufacturer for a period of ~2 weeks and were subsequently retrieved and extracted for PFAS profiling using LC-IMS-MS.[1, 5] All analyses were performed in negative ion mode over the 50-1700 *m/z* range and MS data was acquired solely in MS^1 mode for sensitive PFAS monitoring (precursor scan only).

Pregnancy Data Collection:

Previously, plasma samples from pregnant individuals diagnosed with preeclampsia (PRE) were collected on the day of delivery and compared to those from healthy controls for evaluation of potential lipidomic biomarkers of condition.[6] All study samples were obtained through the informed consent of the donors and de-identified prior to their receipt by the authors. Each sample was extracted for lipids and assessed using LC-IMS-MS. Data was acquired in both positive and negative ion mode over the same 50-1700 *m/z* range used for characterization of the passive samplers. For lipid characterization, both MS1 and MS/MS for monitoring both precursors and product ions was performed. Data analysis conducted in the previous study noted several proteins and lipid species dysregulated in PRE using targeted data processing of the NTA and hence served as an applicable model to evaluate our developed methods. For this analysis, a subset of the lipidomic portion of this data (n=115) was reprocessed by collapsing the LC dimension as

described and reassessed using the summed IMS-MS screening approach to evaluate signals of interest potentially missed in the previous targeted data processing.

Workflow Discussion

The summed IMS-MS screening approach has several main steps, as outlined in **Figure 2**. Following LC-IMS-MS analysis for samples of interest, data were imported into Agilent's IM-MS Browser, which allows for interactive browsing and visualization of LC-IMS-MS data. Within this software, summed 2D IMS-MS nested spectra or heatmaps can be viewed at different retention times, providing a snapshot of the size and mass of molecules occurring at different timepoints. These individual nested IMS-MS spectra can be summed together, providing a singular 2D spectra and showing the entire molecular profile of the sample over the course of the analysis. While this ultimately collapses the LC dimension, by including chromatography in the experimental workflow, we are able to retain the LC analytical benefits such as reducing ionization suppression and filtering out early and late eluting contaminants. This allows for a more comprehensive picture of molecules present and of interest in the samples.[7] Additionally, following the screening analysis, the collected LC data can be revisited to aid in molecular annotations.

Following collapse of the LC dimension and IMS-MS summation, the final summed IMS-MS spectra for each sample are exported in the form of a matrix with drift time and *m/z* values as the corresponding row and column labels and abundances comprising the individual matrix cells. These matrices were then imported into R for general data cleaning, model construction, and evaluation. The first of these steps was binning the data, which entailed taking the column-wise sum for *m/z* values within 2 Da of each other and the row-wise sum for drift times within 0.5 ms of each other. This step was essential to ensure that all samples had consistent coordinates, and therefore were readily comparable. Bin sizes were chosen based on instrument precision with the goal of features fitting into individual coordinates. However, due to the inherent discretization in this binning process, some features might be split between coordinates.. As such, modifying bin

size parameters as well as the binning function itself to minimize feature splitting is a possible future direction.

Following binning and other data cleaning steps, classification models were constructed. As the number of coordinates greatly outnumbered the samples in the training data (42,036 coordinates vs. 42 samples for the passive sampler data and 28,499 vs. 87 for the pregnancy data), performing coordinate selection prior to building these models was imperative to reduce the risk of overfitting. This was achieved by first combining unique *m/z* and drift time pairs to create coordinates, and then applying Lasso logistic regression, which encourages a sparse model by imposing a penalty on regression coefficients. Selection with this method was initially performed on the entire training dataset, and all coordinates with a nonzero coefficient were retained. However, we noticed that the coordinates selected were sensitive to variation in the size of the training set, as well as the random seed used. Therefore, to improve the stability of the selected coordinates, we opted for a bootstrapped Lasso approach. With this method, 1,000 datasets were randomly generated by sampling the training data with replacement and then Lasso regression was applied to each dataset. From each analysis, selected coordinates were noted, and a running tallying was kept across all iterations. Thus, to be retained for final analysis, a coordinate had to be selected at least 200 out of the 1000 runs. This threshold was set to ensure a balance between retaining informative coordinates and excluding those less consistently relevant. The training data was then subset to only selected coordinates, and classification models were constructed using support vector machines and subsequently evaluated.

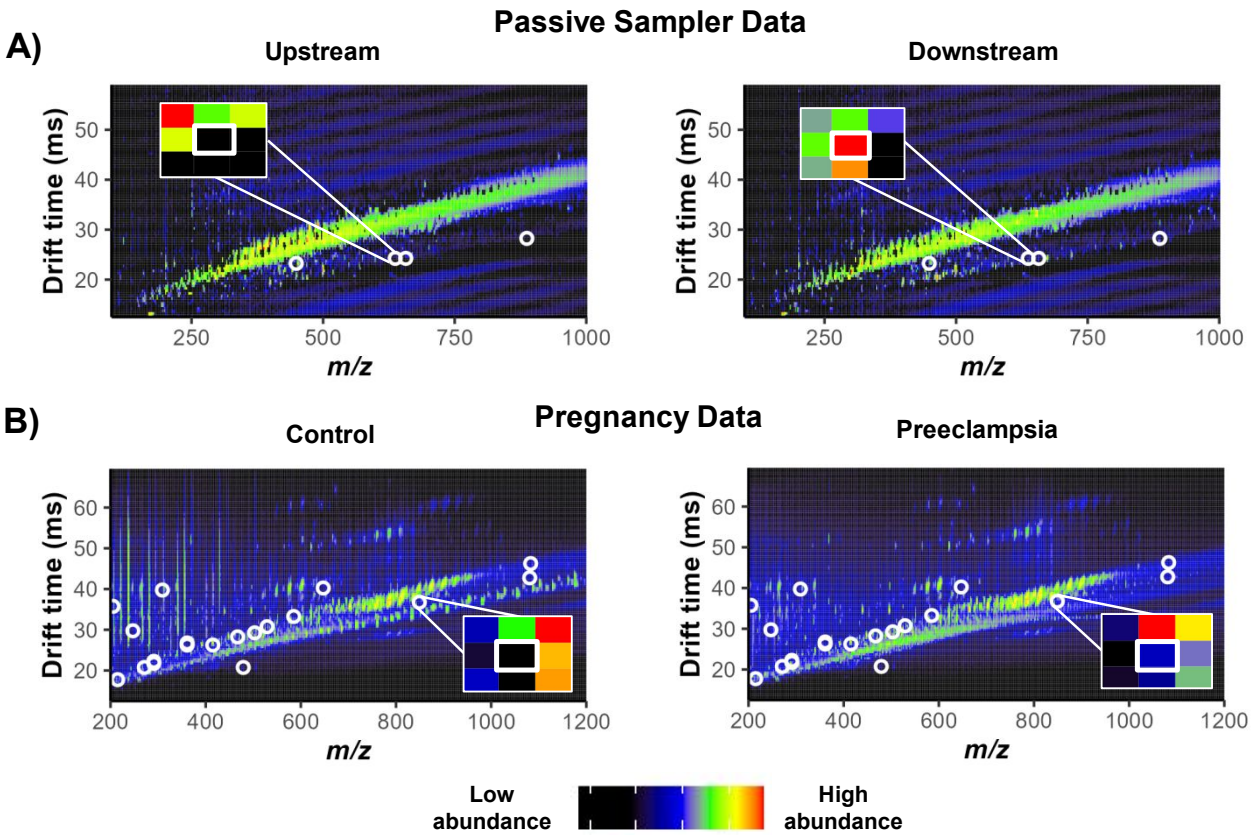


Figure S1: Example sample heatmaps for select **A)** passive sampler data and **B)** pregnancy data. Individual coordinates are colored based on their respective abundances. Circles highlight coordinates that were selected by the bootstrapped Lasso, and a zoomed in example of coordinates within these circles is shown in the upper left of each heatmap.

Lipid Coordinates	*m/z* bin	DT bin	observed *m/z*	RT (if 1-2 peaks)	type	annotation/notes
	1082-1084	46-46.5	1083.803		precursor	no ID
	646-648	40-40.5	647.506		fragment	neutral loss fragment of any FA from TGs with two 20:4 FAs
	466-468	28-28.5	466.336	9.4	precursor	LPE O-18:1 M+H
	204-206	35.5-36	205.041	12.5	fragment	likely fragment of 621.316 (no ID - also has fragments 215, 223, 271 and 279)
	270-272	20.5-21	270.006	2.7	fragment	M+2 of 269.001 (no ID)
	214-216	17.5-18	215.127	2.7	fragment	multiple contributors at same RT - 215.127 (high/low energy) and 215.0907 (high energy); maybe oxidized FAs
	478-480	20.5-21	478.332		fragment	neutral loss fragment of any FA from PCs with a 16:0 FA originating from a doubly charged lipid
	246-248	29.5-30	247.206	6.7	fragment	probably polymer fragment
	290-292	22-22.5	29.233	4.9	precursor	M+1 and M+2 of FA 16:0;O2 (289.2303 - theoretical is 289.2373)
	290-292	21.5-22	29.233	4.9	precursor	Same as above
	528-530	30.5-31	528.394	8.1, 8.5	precursor	M+4 of LPC 18:0 M+H (524.374)
	360-362	26-26.5	360.362	8.5	precursor	in-source fragment of LPC 18:0
	360-362	26.5-27	360.362	8.5	precursor	Same as above
	414-416	26-26.5	414.322	7	precursor	m/z match to AC 16:1;O
	502-504	29-29.5	502.374	6.8	precursor	m/z match to AC 20:1;O3
	308-310	39.5-40	309.281		fragment	neutral loss fragment M-HG-X for PC O-18:0/X M+Na
	1080-1082	42.5-43	1080.803		precursor	no ID
	848-850	36.5-37	849.29	19.1	precursor	M+3 and M+4 of 845.296; maybe SM
	584-586	33-33.5	581.2466	7.3	fragment	fragment

PFAS	*m/z* bin	DT bin	observed *m/z*	RT (if 1-2 peaks)	type	annotation/notes	CCS (Å²)
	448-450	23-23.5	448.954	10.6, 10.9	precursor	maybe fragment of PFNS (-C2F4)	173.8
	636-638	24-24.5	636.935	7.7	precursor	has M-H-CO2 at 592.947	177.9
	656-658	24-24.5	656.904	7, 7.5	precursor	likely related to 672.897, potentially an oxidation product	181.0
	886-888	28-28.5	(1) M+2 of 884.8482 (2) 886.8469	9.8, 10	precursor	likely dimer of 442.9192 containing Sulfur; also an isobar of PFO5DoA and NBP1 dimers	209.7

Table S1: Coordinate annotations for lipids and PFAS noted by *m/z* and drift time bins. Annotations are based on accurate mass, fragmentation spectra when available, and CCS.

A)

Passive Sampler Model Reduced Bin Size		
	Training	Testing
# Samples	42	13
Accuracy (%)	100	100
Sensitivity (%)	100	100
Specificity (%)	100	100
Kappa	1	1

B)

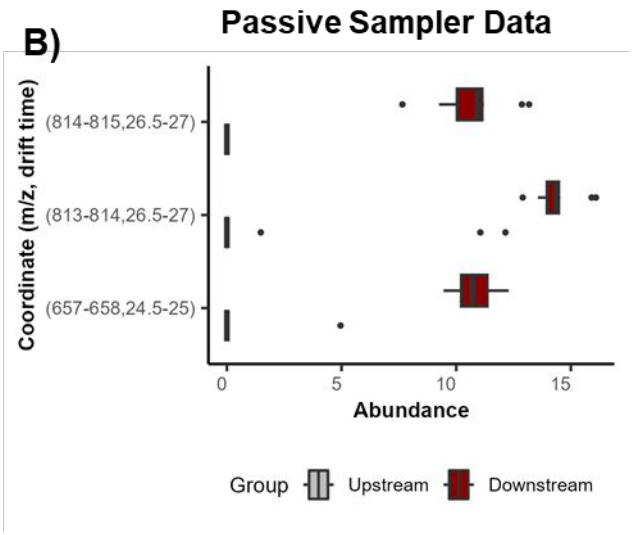
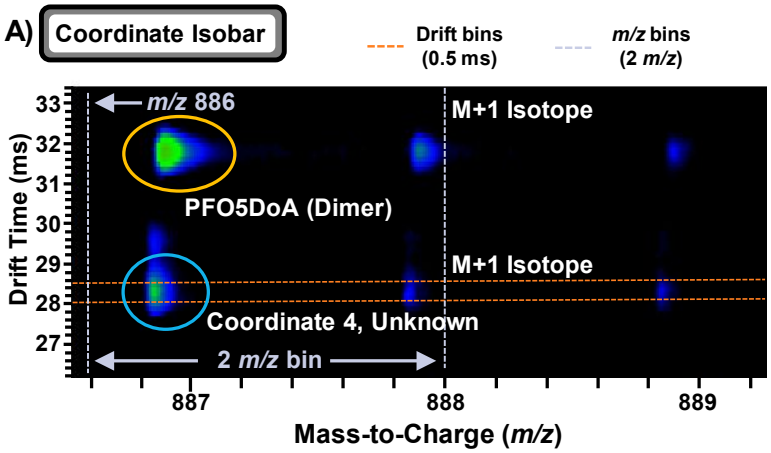


Figure S2: Analysis of passive sampler data with reduced *m/z* bin size (1 Da and 0.5 ms). **A)** Results of a classification model using passive sampler data **B)** Abundance distributions for selected training sample coordinates shown as (*m/z* range (Da) and drift time range (ms)) for the passive sampler data.



B) Coordinate Descriptors

* **PFO5DoA (Dimer)**

Chemical Formula	$C_{14}H_2F_{26}O_{14}$ $[M-H]^-$
Observed *m/z*	886.8969 ~2 ppm error
Exp. CCS Value	233.7 Å^2 (0.2% error)
Retention Time	10.05 minutes

* **Coordinate 4, Unknown**

Chemical Formula	Unknown
Observed *m/z*	886.8557
Exp. CCS Value	207.1 Å^2
Retention Time	9.8 minutes

Figure S3: Illustrative depiction of coordinate assessment of the passive sampler dataset. **A)** Coordinate binning process for both the mass and drift time dimension. **B)** Molecular annotators for known PFAS detected in passive samplers and suspect coordinates.

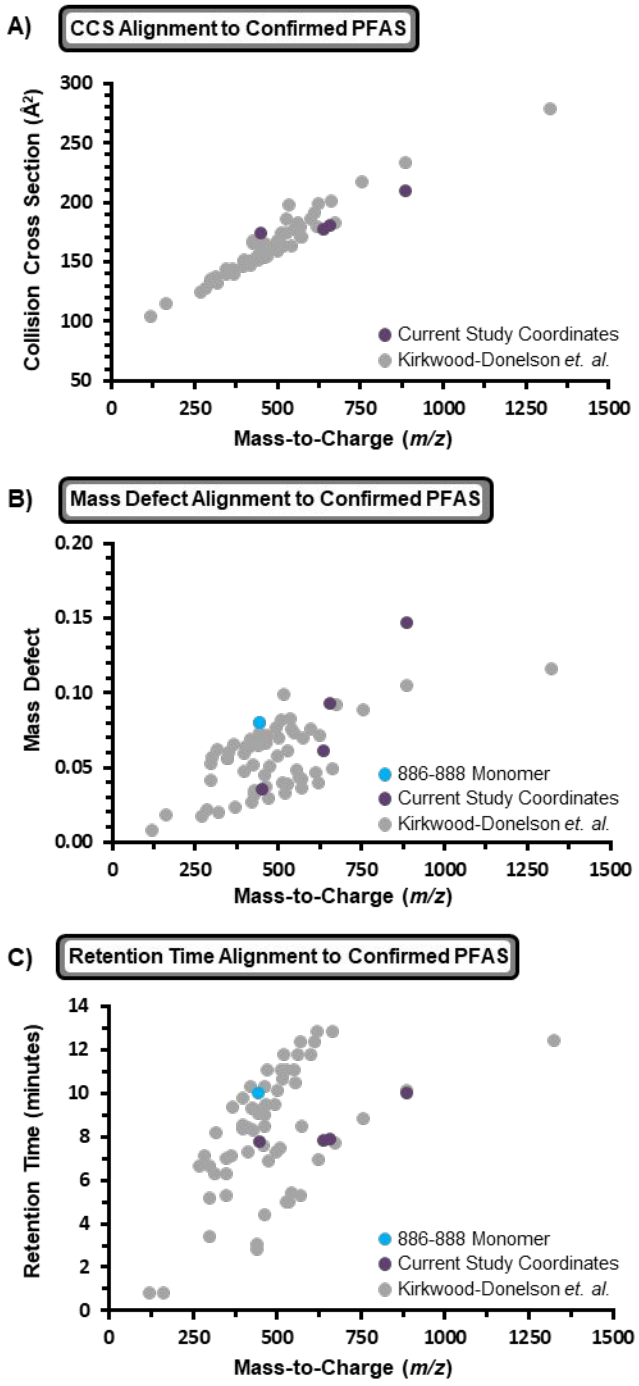


Figure S4: Chemical evidence to support coordinate annotations as potential PFAS. **A)** Coordinates of this model were observed to possess CCS values per mass to charge ratio (m/z) consistent with previously known PFAS. **B)** Mass defect values were also in the same range as previously logged PFAS entries. **C)** Support for dimer notation as retention times between dimers and monomers were conserved.

References:

(1) Kirkwood-Donelson, K. I.; Dodds, J. N.; Schnetzer, A.; Hall, N.; Baker, E. S. Uncovering per- and polyfluoroalkyl substances (PFAS) with nontargeted ion mobility spectrometry-mass spectrometry analyses. *Sci Adv* **2023**, *9* (43), eadj7048. DOI: 10.1126/sciadv.adj7048 From NLM.
(2) Hopkins, Z. R.; Sun, M.; DeWitt, J. C.; Knappe, D. R. U. Recently Detected Drinking Water Contaminants: GenX and Other Per- and Polyfluoroalkyl Ether Acids. *Journal AWWA* **2018**, *110* (7), 13-28. DOI: https://doi.org/10.1002/awwa.1073 (acccessed 2023/11/07).
(3) Sun, M.; Arevalo, E.; Strynar, M.; Lindstrom, A.; Richardson, M.; Kearns, B.; Pickett, A.; Smith, C.; Knappe, D. R. U. Legacy and Emerging Perfluoroalkyl Substances Are Important Drinking Water Contaminants in the Cape Fear River Watershed of North Carolina. *Environmental Science & Technology Letters* **2016**, *3* (12), 415-419. DOI: 10.1021/acs.estlett.6b00398.
(4) Pétré, M. A.; Salk, K. R.; Stapleton, H. M.; Ferguson, P. L.; Tait, G.; Obenour, D. R.; Knappe, D. R. U.; Genereux, D. P. Per- and polyfluoroalkyl substances (PFAS) in river discharge: Modeling loads upstream and downstream of a PFAS manufacturing plant in the Cape Fear watershed, North Carolina. *Science of The Total Environment* **2022**, *831*, 154763. DOI: https://doi.org/10.1016/j.scitotenv.2022.154763.
(5) Dodds, J. N.; Hopkins, Z. R.; Knappe, D. R. U.; Baker, E. S. Rapid Characterization of Per- and Polyfluoroalkyl Substances (PFAS) by Ion Mobility Spectrometry–Mass Spectrometry (IMS-MS). *Analytical Chemistry* **2020**, *92* (6), 4427-4435. DOI: 10.1021/acs.analchem.9b05364.
(6) Odenkirk, M. T.; Stratton, K. G.; Gritsenko, M. A.; Bramer, L. M.; Webb-Robertson, B.-J. M.; Bloodsworth, K. J.; Weitz, K. K.; Lipton, A. K.; Monroe, M. E.; Ash, J. R.; et al. Unveiling molecular signatures of preeclampsia and gestational diabetes mellitus with multi-omics and innovative cheminformatics visualization tools. *Molecular Omics* **2020**, *16* (6), 521-532, 10.1039/D0MO00074D. DOI: 10.1039/D0MO00074D.
(7) Annesley, T. M. Ion Suppression in Mass Spectrometry. *Clinical Chemistry* **2003**, *49* (7), 1041-1044. DOI: 10.1373/49.7.1041 (acccessed 1/18/2024).