Supporting Information for: From Centroided to Profile Mode: Machine Learning for Prediction of Peak Width in HRMS Data

Saer Samanipour,[*,†,‡,¶] Phil Choi,[‡,§] Jake W. O'Brien,[‡] Bob W.J. Pirok,[†] Malcolm J. Reid,[¶] and Kevin V. Thomas[‡]

†*Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands*

‡*Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Woolloongabba, QLD, 4102, Australia*

¶*Norwegian Institute for Water Research (NIVA), Økernveien 94, 0579 Oslo, Norway*

§*Water Unit, Health Protection Branch, Prevention Division, Queensland Department of Health, Brisbane City QLD 4000, Australia*

E-mail: s.samanipour@uva.nl

Number of pages: 12

Number of figures: 13

Number of tables: 2

Contents

S1 Instrumental Conditions

The detailed information related to the datasets used in this study are provided in Table S1.

Table S1: The list samples, ionization mode, vendor, and the associated reference.

nr	Sample Type	Ionization Mode	Vendor	Reference
1	Wastewater Influent	Positive	Sciex	[1]
2	Wastewater Influent	Negative	Sciex	[2,3]
3	Produced Water	Positive	Waters	[4]
4	Produced Water	Negative	Waters	[4]
5	Surface Water Extract	Positive	Agilent	unpublished[a]
6	Surface Water Extract	Negative	Agilent	unpublished[a]
7	Biosolids	Positive	Waters	[5]

[a] Samples were prepared following extraction[4] and analysis[5] procedures detailed elsewhere.

S2 Self Adjusting Centroiding Algorithm

S2.1 Centroiding Parameters

Table S2: The list of parameters, their description, and the used value for centroiding of the data.

nr	Input	Description	value
1	raw data	raw data in mzXML format	-
2	min intensity	minimum absolute intensity for signal	1000
3	resolution	nominal resolution	20000
4	R^2 threshold	threshold for goodness of fit	0.8
5	signal to background	the ratio of the apex to the median signal in the window	1.5

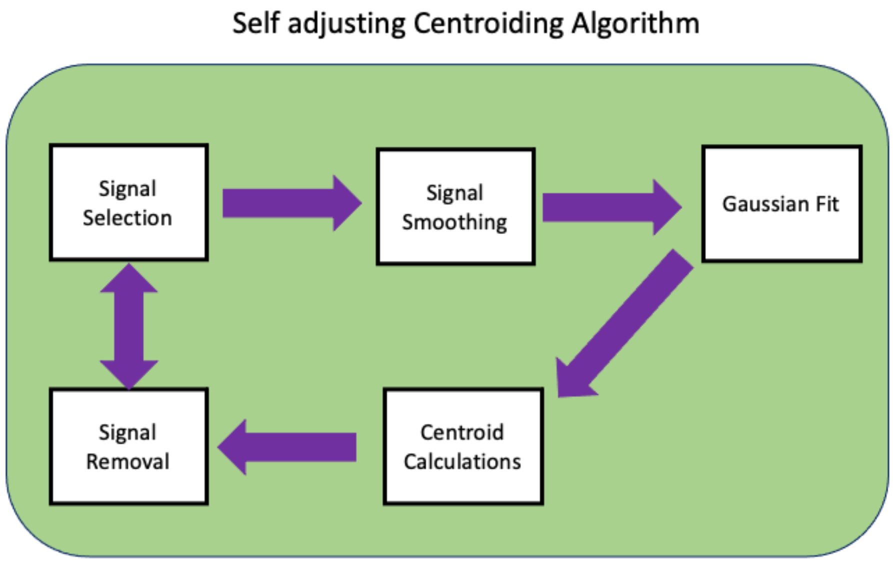

Figure S1: The workflow of the self adjusting centroiding algorithm.

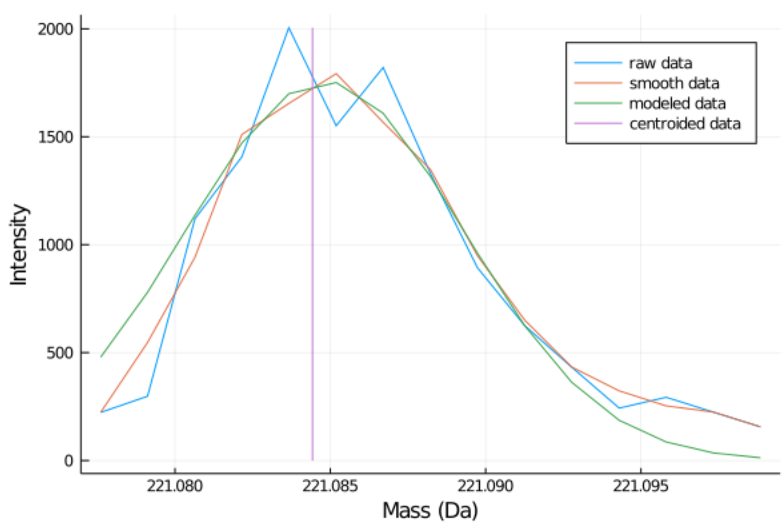

Figure S2: The signal of a successfully detected and centroided peak.

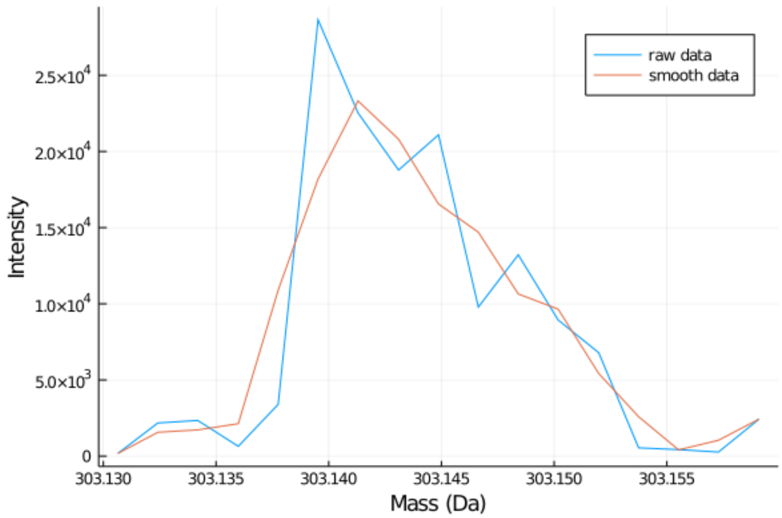

Figure S3: The signal of false negative peak where the algorithm fails to detect and centroid the peak.

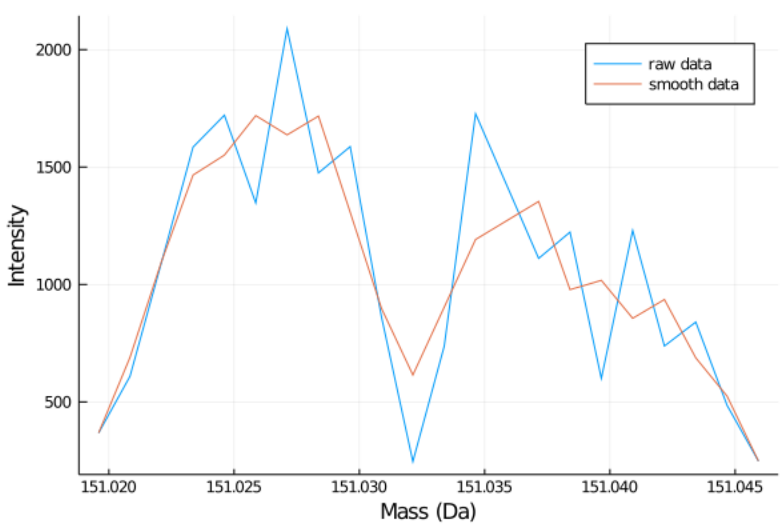

Figure S4: The signal of a true negative assessment by the algorithm. The signal does not belong to a peak, and has been assessed as such.

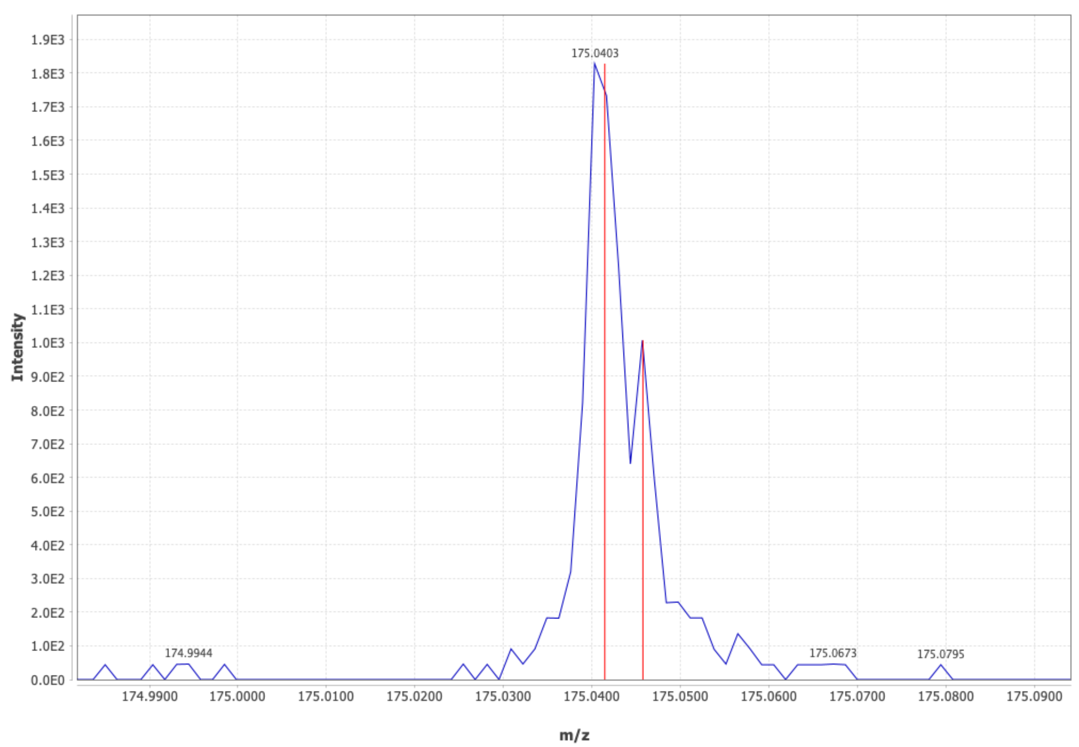

Figure S5: Shows the signal of a true positive (the main peak) and a false positive (the shoulder peak) detected by Centroiding algorithm implemented via MzMine2.[6]

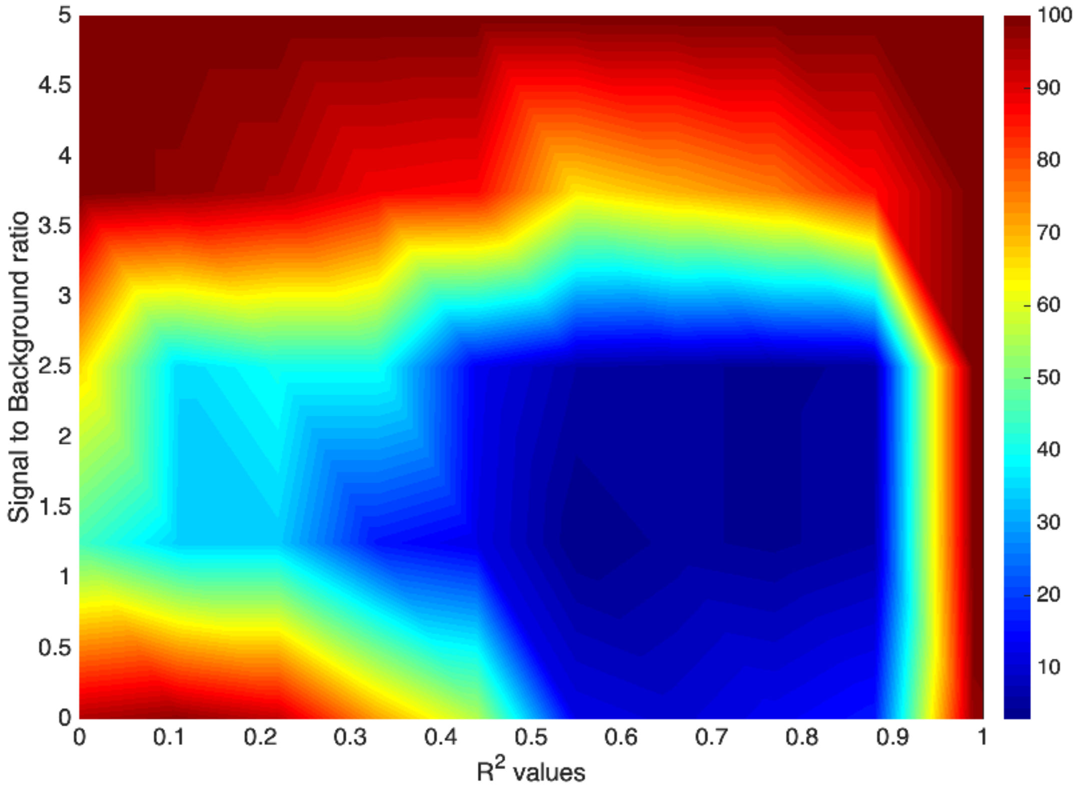

Figure S6: shows the total number of false detection (i.e. the sum of false positives and false negatives) as a function of R^2 and the signal to background ratio.

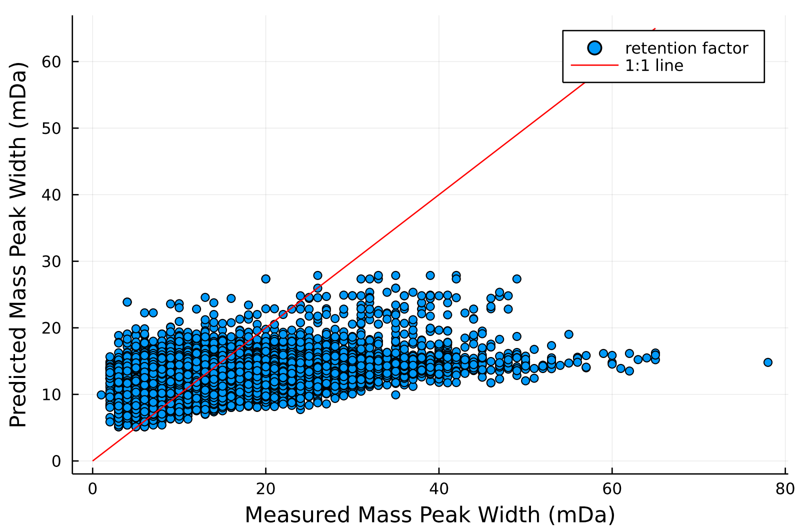

Figure S7: shows the random forest model based on 10000 randomly selected retention factors.

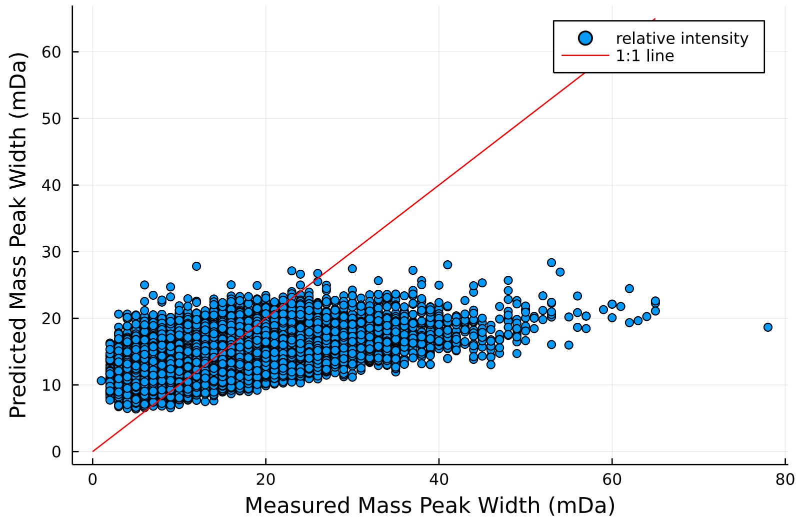

Figure S8: shows the random forest model based on 10000 randomly selected relative intensities.

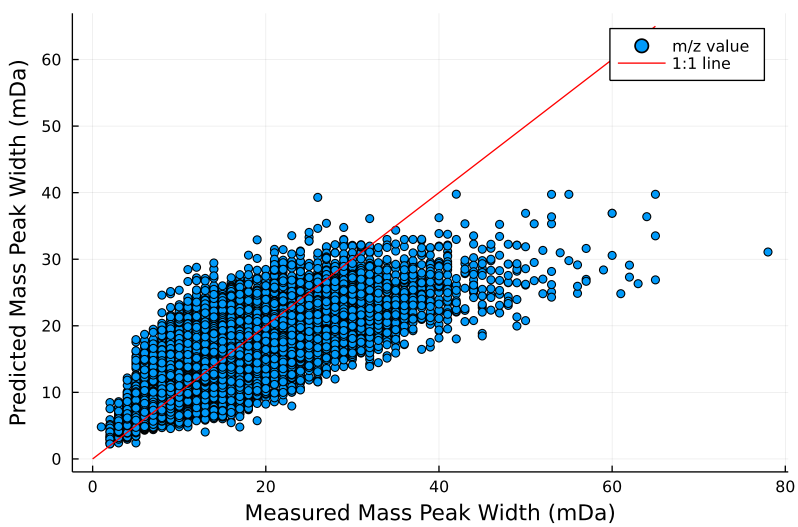

Figure S9: shows the random forest model based on 10000 randomly selected m/z values.

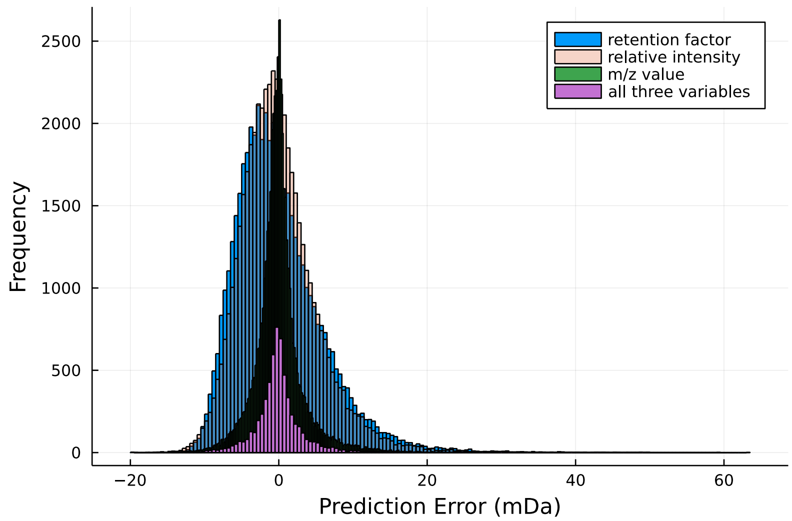

Figure S10: shows the prediction error (mDa) distribution of four models using individual variables as well as all three variables together.

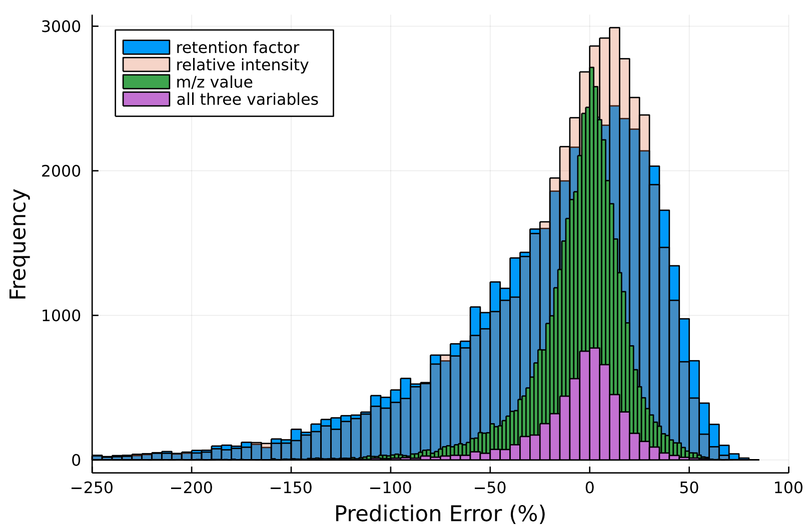

Figure S11: shows the prediction error (%) distribution of four models using individual variables as well as all three variables together.

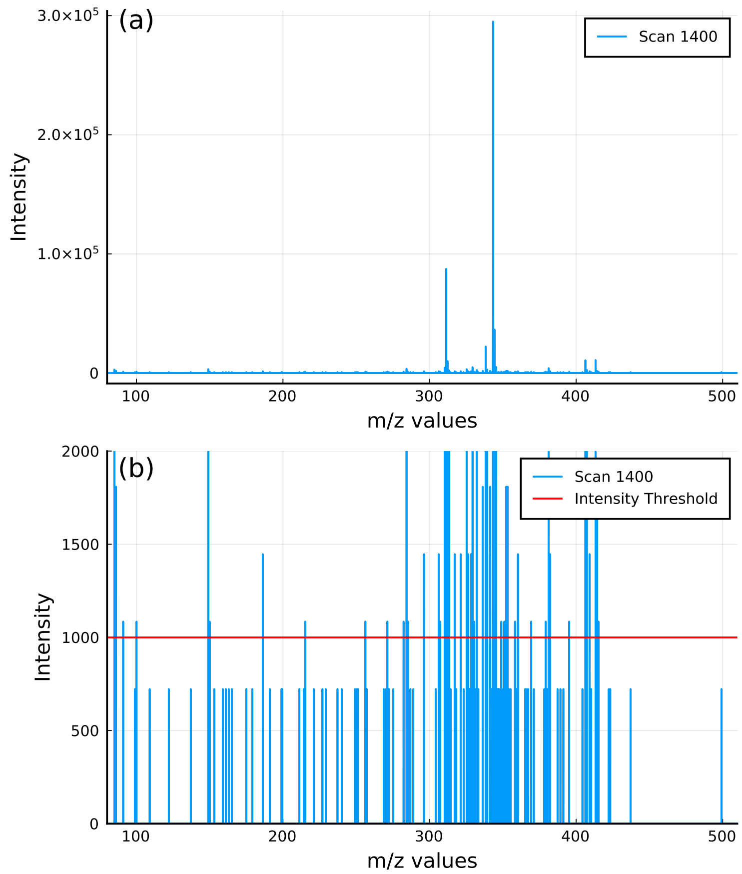

Figure S12: shows (a) the raw signal at scan 1400 and (b) the zoomed in around the set intensity threshold of 1000 counts per second.

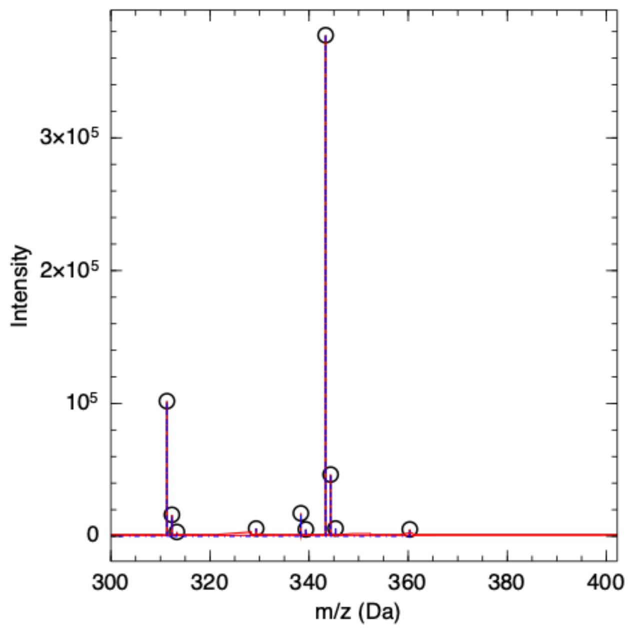

Figure S13: shows the signal of a successfully detected and centroided scan.

References

(1) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V. Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. *Anal. Chem.* **2019**, *91*, 10800–10807.

(2) Choi, P. M.; O'Brien, J. W.; Tscharke, B. J.; Mueller, J. F.; Thomas, K. V.; Samanipour, S. Population socioeconomics predicted using wastewater. *Environ. Sci. Technol. Lett.* **2020**, *7*, 567–572.

(3) Choi, P. M.; Tscharke, B.; Samanipour, S.; Hall, W. D.; Gartner, C. E.; Mueller, J. F.; Thomas, K. V.; O'Brien, J. W. Social, demographic, and economic correlates of food and chemical consumption measured by wastewater-based epidemiology. *PNAS* **2019**, *116*, 21864–21873.

(4) Samanipour, S.; Hooshyari, M.; Baz-Lomba, J. A.; Reid, M. J.; Casale, M.;

Thomas, K. V. The effect of extraction methodology on the recovery and distribution of naphthenic acids of oilfield produced water. *Sci. Total Environ.* **2019**, *652*, 1416–1423.

(5) Samanipour, S.; Reid, M. J.; Thomas, K. V. Statistical variable selection: an alternative prioritization strategy during the nontarget analysis of LC-HR-MS data. *Anal. Chem* **2017**, *89*, 5585–5591.

(6) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics* **2010**, *11*, 1–11.