NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

THE DEVELOPMENT OF PSEUDO-SIM/MRM AND RISK-BASED SCREENING METHODS FOR CHARACTERIZATION OF HUMAN EXPOSOME

YANG JUNJIE

SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING

2022

THE DEVELOPMENT OF PSEUDO-SIM/MRM AND RISK-BASED SCREENING METHODS FOR CHARACTERIZATION OF HUMAN EXPOSOME

YANG JUNJIE

School of Civil and Environmental Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2022

Statement of Originality

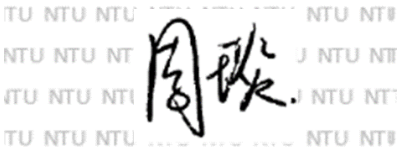
I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarized materials, and has not been submitted for a higher degree to any other University or Institution.

2022/03/28
Date

YANG JUNJIE

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.



2022/03/28
Date

Zhou Yan

Authorship Attribution Statement

This thesis contains material from [1] paper published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as Yang, J. J, Han, Y. Mah, C. H. Wanjaya, E. Peng, B. Xu, T. F. Liu, M. Huan, T. and Fang, M. L. (2020). "Streamlined MRM method transfer between instruments assisted with HRMS matching and retention-time prediction." Analytica Chimica Acta, Vol. 1100, pp. 88–96. DOI: https://doi.org/10.1016/j.aca.2019.12.002.

The contributions of the co-authors are as follows:

- A/Prof Fang Mingliang provided the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by A/Prof. Huan Tao, Dr. Han Yuan, Dr.Mah Chin Hao, Dr. Xu Tengfei, Dr. Liu Min, Wanjava Elvy, and Peng Bo.
- I co-designed the study with A/Prof Fang Mingliang and performed all the laboratory work at the school of civil and environmental engineering.
- All samples analysis by liquid chromatography coupled with mass spectrometry and sample preparation were conducted by me.
- Dr. Han Yuan assisted in the experimental trouble shooting.

2022/03/28
Date

YANG JUNJIE

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Zhou Yan, for her support and guidance in my thesis writing and my previous supervisor, Prof. Mingliang Fang, for his support and guidance throughout my thesis work. With their instructions, I have gained deep insights into my research field and critical experiment skills for good lab practices. I also would like to thank Prof. Grzegorz Lisak and Prof. Xunchang Fei for their guidance during my TAC committee meetings.

My gratitude extends to the School of Civil and Environmental Engineering for the funding opportunity to undertake my studies at the Nanyang Technological University. I would also like to acknowledge the technical staff in the Environmental Lab and NEWRY: Mr. Ong Chee Yung, Mr. Tan Han Khiang, Mrs. Lim-Tay Chew Wang, Ms.Maria Chong Ai Shing, Ms. See Shen Yen, Pearlyn, Dr. Lv Yunbo, Ms. Elvy RianiWanjaya, and Ms. Koh Danyu for their support in the laboratory.

I would like to thank my lab mates and friends: Mr. MengJing Wang, Ms. Seang Lidet Yean, Mr.Yao Sun, Ms. Haoyang Wang, Mr. Jiazu Zhou, Mr. Bo Xu, Mr. Lulu Zhang, Mr. Yi Tu, Ms. Peng Bo, and Dr. Jie Zheng, Dr. Cheng Zhou, Dr. Chinhao Ma, Dr. Fanrong Zhao, and Dr. Yingdan Zhang for a cherished time spent together in the lab and social settings. It is their kind help and support that have made my study and life in Singapore a wonderful time.

Finally, my deep appreciation goes to my family and my parents for their support and comfort throughout my thesis study. Their support and understanding help me survive the downtimes in the past several years and overcome all obstacles along the way.

TABLE OF CONTENTS

SUMMARY

Exposome, which studies all the environmental exposure during the lifelong period, is the new era of human health research and a rapidly expanding research area. Tracking chemical exposures in the environment or the human body has been performed by targeted analysis using multiple reactions monitoring or single ion monitoring (MRM/SIM) method and suspect screening in non-targeted analysis. However, the constraints of analytical platforms, such as high application costs, detection of poor throughput, restricted chemical coverage, and reliance on chemical standards for method developments, have hampered the growth of exposome characterization. One of the strategies to face the limitations in the targeted analysis is to develop computational optimization methods and further develop MRM/SIM spectral databases for different instruments. Meanwhile, another strategy is to develop high-throughput workflows with multiple screening filters.

Here, this thesis demonstrates the development of several proof-of-concept methodologies in targeted analysis and non-targeted analysis to tackle challenges in the characterization of human exposomes. MRM methods in mass spectrometry coupled with liquid chromatography (LC-MS) can be optimized computationally. The instrument-dependent settings of collision energy can be generalized for applications in different instruments. A reliable and fast screening platform is developed with generalized MRM methods. When coupled with retention time prediction and high-resolution mass spectrometry (HRMS), the conventional targeted analysis workflow can identify environmental exposure without chemical standards. Beyond CE optimization, this thesis replaces the experimental optimization of MRM/SIM transitions with an in-silico optimization strategy. Databases of pseudo-MRM/SIM spectra are developed by a pseudo-spectra algorithm using existing public spectral databases. The databases provide optimized MRM transitions with high selectivity for over 300,000 exogenous chemicals. Furthermore, this thesis develops a novel sensitive and high-throughput exposome analytical platform (CIL-ExPMRM) by isotope labeling urinary biomarkers to increase the detection of chemicals at trace levels. The CIL-pseudo-MRM exposome database consists of environmental pollutants and their transformation products for 110,000 compounds. The platform has been well incorporated with automatic MRM generation, dynamic MRM optimization, and data analysis. Meanwhile, this thesis proposes a new non-target

analysis workflow for environmental chemical screening using multiple screening filters and risk-based chemical prioritization. Retention time prediction and spectral matching can provide structural elucidation in identification. Toxicity prediction links the MS fragments to chemical toxicity. Risk-based prioritization highlights the candidates of high threat to human health. Overall, this thesis found the generalized method for optimizing collision energy in MRM methods. The pseudo-MRM/SIM spectral database and CIL-pseudo-MRM spectral databases provided alternatives for MRM transition optimization in the targeted analysis community. The automatic and integrated platform rendered high-throughput suspect screening for non-targeted analysis of human exposomes.

LIST OF PUBLICATIONS

First-author paper

- **Junjie Yang,** Yuan Han, Chin Ho Mah, Elvy Wanjaya, Bo Peng, Tengfei Xu, Min Liu, Tao Huan, and Mingliang Fang (2020). "Streamlined MRM method transfer between instruments assisted with HRMS matching and retention-time prediction." Analytica Chimica Acta, Vol. 1100, pp. 88–96. DOI: https://doi.org/10.1016/j.aca.2019.12.002.
- **Junjie Yang,** Fanrong Zhao, Jie Zheng. Mengjing Wang, and Mingling Fang. "An Automated Toxicity Prioritization Framework for Fast Characterizing Exposome in Non-Targeted Analysis." Analytica Chimica Acta. (Under review)
- **Junjie Yang,** Fanrong Zhao, Haoduo Zhao, Jie Zheng, Mengjing Wang, Jigang Wang, and Mingling Fang. "Integrative chemical proteomics-metabolomics approach revealing direct molecular targets of BADGE." (To be submitted)

Co-author paper

- Fanrong Zhao; Li Li; Penghui Lin; Yue Chen; Shipei Xing; Huili Du; Zheng Wang; **Junjie Yang**; Tao Huan; Cheng Long; Limao Zhang. M.L. Fang. "HExpPredict: In Vivo Exposure Prediction of Human Blood Exposome using A Random Forest Model and Its Application in Chemical Risk Prioritization." Environmental Health Perspectives. (Under Review)
- Jie Zheng[#], **Junjie Yang[#]**, Fanrong Zhao, Bo Peng, Xu Liang, Cheah Yeong Cheng, Jingtao Zhang, Yulan Wang, Mingliang Fang "CIL-ExPMRM: A User defined CIL-Pseudo-MRM Exposome Platform." (To be submitted)

LIST OF FIGURES

spectrometer using real sample on-column injection. (b) Well-matched extracted ion chromatography (EIC) from the analysis of spiked compounds in the synthetic water sample from HR-MS and LC-MS/MS. (c) Performance of the multiple linear regression model for retention time prediction. (a) Reference CE was derived from the averages from MRM archives and stepwise optimization was conducted. " " represents the CE values whereby we obtained the maximal intensity. (b) EIC of 9 representative compounds from analysis of spiked compounds in the synthetic water sample from HR-MS and LC-MS/MS. (c) Pearson r^2 between the predicted retention time and the observed retention time over training set as red dots, and test set as blue dots was 0.66 ($p < 2.2e-16$), and 0.63 ($p = 1.23e-8$) respectively.

Figure 3.4. Method validation and performance in validation cases. (a-b) Examples of scenarios in the analysis of pre-spiked small molecules in complex samples. Highlighted box-shape area is the predicted retention time window from the QSRR model. (c) Extracted ion chromatography (EIC) of MRM transitions and HR-MS of BPA, BPAS, and BPAG. (a) Putative signals of ampicillin in HR-MS, positive signal in LC-MS/MS, and predicted retention time in the sludge water sample. RTHR-MS = 7.2/8.7 min, RTLC-MS/MS= 7.3 min, RTPredicted = 6.4-10.4 min. (b) Putative signals of ciprofloxacin from HR-MS, LC-MS/MS, and predicted retention time in the urine sample. RTHR-MS = 1.4/7.5 min, RTLC-MS/MS= 1.4/7.5 min, RTPredicted = 6.9-10.9 min. (c) Putative signal of bisphenol A in HR-MS, LC-MS/MS, and predicted retention time in the cell extract. RTHR-MS = 11.5/19.0 min, RTLC-MS/MS= 11.5/19.1 min, RTpredicted = 9.5-13.5 min; Putative signal of BPA glucuronide (RTHR-MS = 8.3/9.7 min, RTLC-MS/MS= 1.3/9.8 min, RTpredicted = 8.2-12.2 min) and BPA monosulfate (RTHR-MS = 19.0 min, RTLC-MS/MS= 19.2 min, RTpredicted = 13.1-17.1 min) in cell extract from HR-MS, LC-MS/MS, and retention time prediction, as depicted in highlighted area.

Figure 4.1. Overview of the pseudo-SIM/MRM database development. A) workflow for developing a pseudo-SIM database. B) workflow for developing a pseudo-MRM database.

Figure 4.2. Chemical space and spectra repository of NIST EI-MS and MoNA database

Figure 4.3. Data cleaning for GC/MS and LC/MS spectra from NIST EI-MS database and MoNA LC-MS/MS database.

chemical standards. C) Candidates lists prioritized by predicted retention time and matching scores from library matching in waters progenesis QI.

Figure 6.4. A) Predicted toxicity at ORLD50 endpoint. B) The toxicity ranking of 28 spiked chemicals among total candidates. C) Six toxicity endpoints and their prediction accuracy by EPA TEST software and ToxPi.

Figure 6.5. A-C) 1475 candidates with 26 spiked chemicals were prioritized by deltaRT, and MS2 spectra matching scoring, and toxicity levels. 26 spiked chemicals were classified in the top 3 RTMS2 levels, with most of them showing toxicity prioritization at Tox_level 2.D) A ranking pyramid of 4 tiers of prioritization combining retention time, MS2 spectra, and toxicity endpoints levels. E) A case study of prioritization process for candidates' hit list at peak m/z=238.0851 with RT = 7.6 min.

LIST OF ABBREVIATIONS

Name	Abbreviations
Gas chromatography	GC
Liquid Chromatography	LC
Mass spectrometry	MS
Selected ion monitoring	SIM
Multiple ion monitoring	MRM
Low resolution mass spectrometry	LR-MS
High resolution mass spectrometry	HR-MS
Triple quadrupole	QqQ
Time-of-flight	TOF
Quadrupole time-of-flight	QTOF
Electrospray interface	ESI
Tandem mass spectrometry	MS2 OR MS/MS
Collision energy	CE
Exposome wide association studies	EWAS
Retention time	RT
Retention index	RI
Quantitative structure-retention relationship	QSRR
Multiple linear regression	MLR
Non-Target Analysis	NTA
Data independent acquisition	DIA
Data dependent acquisition	DDA
Molecular descriptors	MDs
Simplified Molecular Input Line Entry System	SMILES
Chemical isotope labelling	CIL
Pharmaceuticals/personal care products	PPCPs
Polychlorinated biphenyls	PCBs
Per-/polyfluoroalkyl substances	PFAS
Organochlorine	OC
Polybrominated diphenyl ethers	PBDEs
Polycyclic aromatic hydrocarbons	PAHs

Volatile organic compounds	VOCs
Trimethylamine N-oxide	TMAO
Acetonitrile	ACN
Ampicillin	AMP
Tris(2-chloroethyl) phosphate	TCEP
Bisphenol A	BPA
Triclosan	TCS
Triclocarban	TCC
Ethyl paraben	EP
Methyl paraben	MP
Methylphenyl ethylamine	MPEA
Triphenylphosphine	TPP
2,2’-dithiodipyridine	DPDS
Dimethyl aminopyridine	DMAP
Trimethylamine	TEA
Ethyl acetate	EA
Mililiter	mL
Milimolar	mM
Milimeter	mm

CHAPTER 1: INTRODUCTION

1.1 Background

Exposome is the study of the totality of environmental exposures from birth onwards (Wild 2012). The chemical exposures of a wide range under investigation require multiple monitoring methodologies, including targeted, non-targeted (untargeted), and semi-targeted methods based on mass spectrometry (Dennis et al. 2017). Targeted analysis based on low-resolution mass spectrometry coupled with gas/liquid chromatography (GC/LC-MS) provides sensitive detection for exposures present at low concentrations. GC-MS-based targeted analysis using selected ion monitoring mode (SIM) and LC-MS-based targeted analysis using multiple ion monitoring modes (MRM) are widely used to monitor environmental exposures, including volatile organic compounds, persistent and non-persistent pesticides, and industrial chemicals (Woodruff et al. 2011). Non-targeted (untargeted) approaches using high-resolution mass spectrometers (HR-MS) can offer high-throughput detections of environmental exposure covering a wide range of chemical entities. Its application in the analysis of small molecules and macromolecules biomarker in human specimens plays a critical role in exposome-wide association studies (ExWAS).

Overall, targeted and non-targeted analyses are widely employed for characterizations of human exposomes yet still face multiple overarching challenges. Targeted analyses are available to most researchers for exposomic characterization yet are limited to known compounds. Once without available chemical standards, one common practice of method optimization for targeted analysis is out of the option, resulting in a shortage of detection of contaminants at a low concentration level. One way to bypass the necessity of chemical standards is to *in silico* optimize the targeted methods by data mining existing curated databases. However, current optimization methods have highly relied on real-time data collection for MS spectra. The accuracy and specificity of in silico predicted MS spectra need more improvements because matrix effects can make quantification difficult and mask the signal from compounds with extremely low concentrations, such as endogenous and exogenous metabolites. On the other hand, even with existing optimized methods from publications, the optimization result is questionable when one method is transferred to another laboratory for application on a different instrument. There is a lack of conversions of targeted methods across different platforms. Recent studies have shown the benefits

of combining LC-QQQ-MS with HRMS for compound identification at a high confidence level, even without chemical standards. However, there is a lack of standardized procedures and standardized platforms for generalized applications.

The coverage of detection for chemical components is inherently limited in the targeted analysis due to the necessity of chemical standards and prior knowledge for spectral fragmentation of the target compounds. Non-targeted analysis, however, enables us to measure more than thousands of chemical components of human exposomes in a single run. Data deconvolution is challenging because compound identification required data processing, peak analysis, and candidate screening from large datasets that are generated in non-targeted analysis. One solution for the fast deconvolution of large datasets is the *in-silico* prediction of retention time for compound identification. Recent studies resort to integrated platforms using reference spectra from curated MS/MS libraries, retention time prediction, and metadata from existing exposome databases. However, integrated platforms require bioinformatic tools and in-house databases for development. On the other hand, prioritization of detected features has been in practice for a long time, as one of the strategies in semi-targeted analysis. Suspect targets of foremost concerns, such as toxic compounds and concentrated pollutants, can be prioritized for further quantification. It is challenging to prioritize more than thousands of candidates in current studies because of the limited integration of comprehensive environmental chemical databases.

1.2 Purpose and scope

The critical role of exposome characterization in exposome studies requires monitoring approaches with less dependence on standards, higher sensitivity, higher efficiency, and better generalization. To further overcome the challenges that current exposome researchers encounter, this thesis aims to achieve four purposes:

Aim 1: This thesis aims to develop a targeted analysis platform with less dependence on chemical standards by using multiple techniques. The aim is to build a reliable and fast-screening platform based on mass spectrometry coupled with liquid chromatography (LC-MS) using multiple reactions monitoring (MRM) methods with a combination of retention time (RT) prediction, *in silico* optimization of collision energy (CE), and HRMS.

Aim 2: The current targeted analysis is highly dependent on chemical standards for optimization of MRM pair ions, which is impossible to fulfill with

the *in-silico* prediction of RT and CE. This thesis aims to provide *in-silico* SIM and MRM pair ions optimization without standards using pseudo-SIM/MRM databases.

Aim 3: On top of the method developments, the matrix effect often masks the detection of exogenous and endogenous compounds of low concentrations. This thesis aims to develop a novel sensitive and high-throughput exposome analytical platform by isotope labeling urinary biomarkers to enable sensitive detection of chemicals at trace levels, with a combination of a pseudo-SIM/MRM database.

Aim 4: To further explore the unknown compounds beyond the detection of targeted analysis, this thesis aims to use risk-based chemical prioritization for a new non-targeted analysis workflow for environmental chemicals screening by providing software platforms for easy processing of large data.

1.3 Dissertation Overview

This dissertation consists of 7 chapters. Chapter 1 and Chapter 2 are the introduction and literature review. Chapter 8 includes conclusions and recommendations. The other chapters are as follows:

Chapter 3: I have proposed a reliable and fast screening platform based on mass spectrometry coupled with liquid chromatography (LC-MS) using the multiple reactions monitoring (MRM) method. With the assistance of retention time prediction models based on quantitative structure-retention relationships (QSRR), I successfully used conventional targeted methods for environmental exposure characterization without chemical standards.

Chapter 4: I further scaled up the suspected screening method by developing a pseudo-single ion monitoring (GC-SIM) and LC-MRM databases (https://github.com/YANGJJ93MS/Pseudo-SIM-MRM.git) with a powerful algorithm to increase sensitivity and specificity for over 300,000 exogenous chemicals with available MS/MS fragmentation in public databases.

Chapter 5: I have developed a novel sensitive and high-throughput exposome analytical platform (CIL-ExPMRM) by isotope labeling urinary biomarkers to increase the detection of chemicals at trace levels. I further built up a CIL-pseudo-MRM exposome database of environmental pollutants and their transformation products for 110,000 compounds. The platform has been well incorporated with automatic MRM generation, dynamic MRM optimization, and data analysis.

Chapter 6: I further proposed one new non-target analysis workflow for environmental chemical screening using risk-based chemical prioritization. It can help detect environmental exposure with the highest risk in samples with top priority.

Chapter 7: I revisited the summary of those developed exposome characterization approaches and provided future directions for exposome studies. Overall, the results of this thesis have overcome the barriers and data gaps to characterize human exposome with a standard-free acquisition, higher sensitivity, coverage, and accuracy with a combination of targeted analysis, suspected screening, and prioritized non-targeted approaches.

CHAPTER 2: LITERATURE REVIEW

2.1 Environmental exposure and human exposome

Environmental exposure has been recognized as a large part of risks factors, apart from the genetic factors, contributing to the chronic disease, such as cancer, diabetes, and vascular diseases (Vineis, 2004; Hindorff et al., 2009; Cosselman, Navas-Acien and Kaufman, 2015). Early in 2000, scientists concluded that inherited genetic factors contribute minor (10%) to susceptibility to most types of neoplasm and the environment has the principal role (90%) in causing sporadic cancer (Lichtenstein et al., 2000). A later study reported that 70%-90% of chronic diseases, including colon cancer, stroke, coronary heart disease, and diabetes, were probably attributed to environmental influences (Willett 2002). Improvements in the environment were reported to be the key to the prevention of a certain fraction of cancers. (Tomasetti, Li and Vogelstein, 2017). The principal role of environmental factors in developing chronic diseases suggests major knowledge gaps in human exposure to the environment.

Environmental exposures contributed to 51% of chronic-disease mortality globally in 2010. Most of the exposures were identified as airborne particles, active and passive smoking, high plasma glucose, high sodium, alcohol, and high cholesterol (Fig. 2.1A) (Rappaport et al. 2014). Over 1,400 small molecules have been summarized in human blood exposomes in this review article. For example, 52 chemicals in six different chemical classes were identified in pregnant women in a cohort study (Fig. 2.1B). Those identified chemicals included cotinine, metals, organochlorine (OC) pesticides, phthalates, brominated flame retardants (PBDEs), and Polycyclic aromatic hydrocarbons (PAHs) (Woodruff, Zota and Schwartz, 2011). In addition, personal daily exposures to the environment are dynamic and complex. Longitudinal and spatial variations in the environment generate extensive and various exposures to chemicals (Jiang et al. 2018). Such variety and complexity of chemical exposure were recognized to be the attribution for extra toxicity to human health and challenges of exposure identification (Escher et al. 2020).

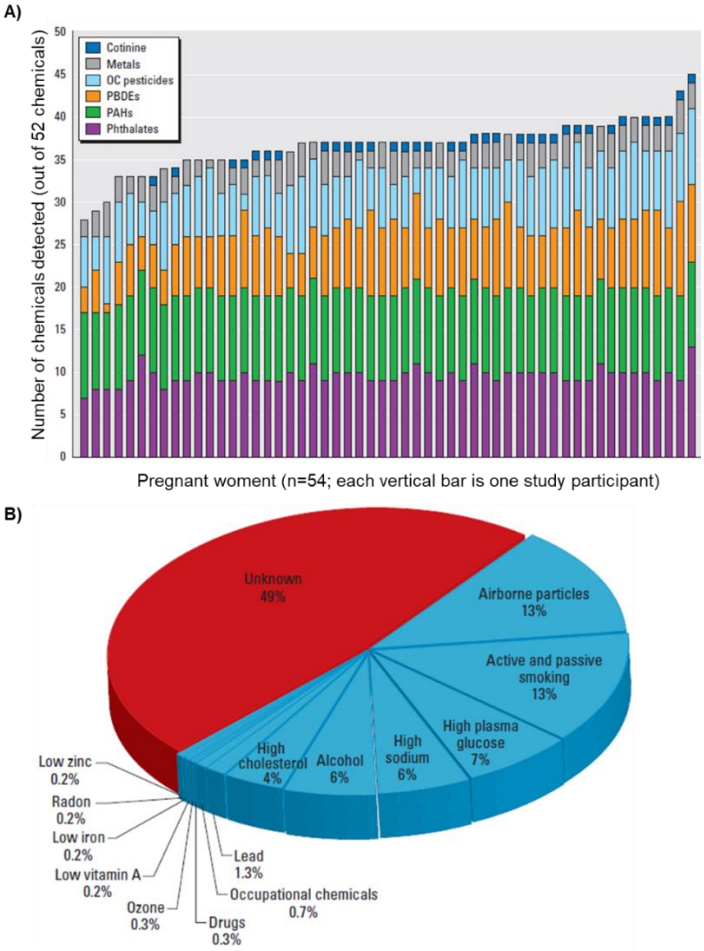


Figure 2.1. Human exposure to chemicals. A) Number of chemicals detected by chemical class in U.S. pregnant women (Woodruff et al. 2011). B) Risk factors for exposures that contribute to chronic-disease mortality (Rappaport et al. 2014).

Exposures come from the environment we live in, including physical and social activities. They can be from external environments and internal environments of the human body (Fig 2.2). External environmental exposures originate from ecosystems (radiation and pollution), personal lifestyle, personal social activities (stress and anxiety), infections, medication (drugs), personal diet, and other physical-chemical environments. Internal environmental exposures refer to those detected inside the human body, including xenobiotics, inflammation, preexisting disease, lipid peroxidation, and so forth (Rappaport and Smith, 2010; Vermeulen, Emma L Schymanski, et al., 2020). The broad range of environmental exposures makes it challenging to unravel the association between exposure and human health. Exposome, as a new concept, is defined to investigate exposure in holistic methods. It takes a long way for traditional biomonitoring research to evolve into new exposome studies.

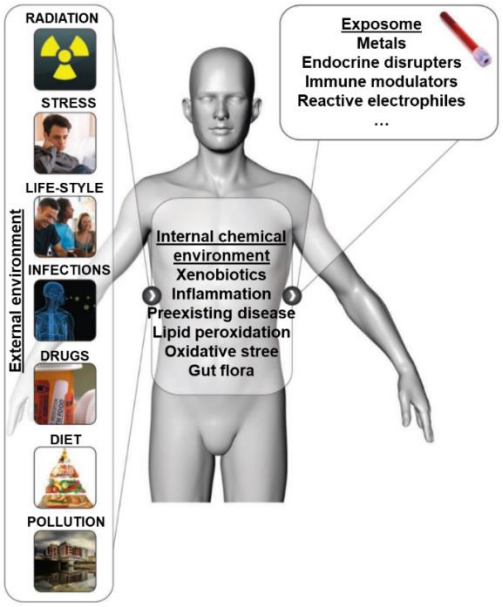


Figure 2.2. Exposures from external and internal environments and the concept of exposome (Rappaport and Smith 2010).

2.2 Characterizing approaches for human exposome

During the last two decades, the concept of exposome has developed in a more comprehensive way to provide more tangible results of environmental exposure assessment. Exposome research takes advantage of advanced mass spectrometers, bioinformatics, and wearable sensors, broadening the exposome horizon with a fusion of omics methods. Characterizing approaches of exposomes include biological approaches, chemical approaches (measure exogenous and endogenous biomarkers to describe human body changes in response to exposure to chemical equipment), and other special approaches regarding different subjects such as wearable devices for personal exposure monitoring. Among all the exposome factors, chemical characterization is the most challenging one. Most current studies focus on the investigation of the chemical component of the human exposome and the development of methodologies to characterize the chemical exposures (Chung et al. 2021; González-Domínguez et al. 2020; Grashow et al. 2020; Hu et al. 2021; Jiang et al. 2018; Maitre et al. n.d.; Perera 2017; Robinson et al. 2018; Schymanski, Singer, et al. 2014).

Characterization approaches in the current exposome studies rely on mass spectrometry. They can be grouped into three major categories (Fig 2.3A-B): (1)

conventional targeted approaches focusing on monitoring the known compounds present in environments and human specimens introduced by exposures, using low-resolution or high-resolution mass spectrometry (LRMS/HR-MS); (2) non-target approaches focusing on unknown compounds from the exposure or unknown metabolites from the exposure-induced biological response, using high-resolution mass spectrometry (HR-MS) in a high throughput manner, and (3) semi-targeted approaches utilizing the discovered compounds list by non-targeted approaches and quantifying the discovery by targeted approaches.

Mass spectrometry-based analytical techniques have been the fundamental tool in current exposome research. The mass spectrometer, often coupled with gas/liquid chromatography separation techniques (GC/LC-MS), expands our analytical workflows from external environmental exposures to endogenous exposures, facilitating a holistic and systematic assessment of the totality of environmental impact on human health. It is a predominant analysis approach for the simultaneous monitoring of numerous chemicals from a large variety of media, from human specimens to environmental samples. Mass spectrometry-based exposome characterizations evolve with the development of mass spectrometry technologies, in terms of detection sensitivity, specificity, and dynamic range (Andra et al. 2017b). Low-resolution mass spectrometry (LR-MS) is widely used in targeted exposome studies for monitoring known compounds at a trace level. High-resolution mass spectrometry (HR-MS) is developed to discover more unknown chemicals present in samples in non-targeted approaches.

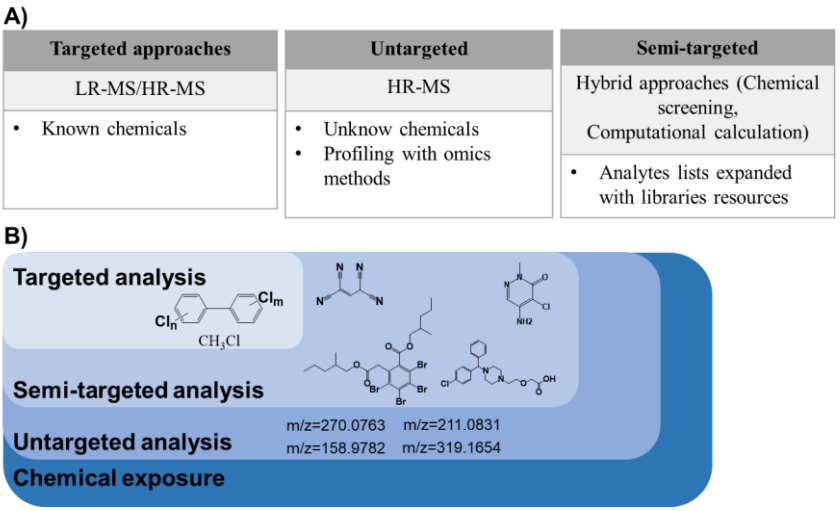


Figure 2.3. A) Targeted, non-targeted, and semi-targeted analysis components and B) their detection coverage (Escher et al. 2020).

2.3 Targeted analysis in the exposome characterization

Tremendous exposome studies heavily rely on the LR-MS-based targeted approach because of its stability, sensitivity, and availability for environmental monitoring and biological monitoring. For example, National Health and Nutrition Examination Study (NHANES) relies primarily on targeted analysis. Volatile organic compounds (VOCs), PBDEs, PAHs, and other common pesticides were found by GC/LC-MS/MS method in pregnant women in NHANES 2003-2004 project (Woodruff et al. 2011). The environmental fate of widely spread pollutants is well monitored by targeted methods. Commercialized plasticizers such as bisphenols are well characterized by LC-MS/MS method in the external and internal environments, even at a very low concentration (Chen et al. 2016). The relationship between chronic diseases with environmental exposures from dietary habits is well studied based on the identification by targeted analysis. Metabolites from dietary lipid phosphatidylcholine-choline, trimethylamine N-oxide (TMAO), and betaine were identified and related to risk for cardiovascular disease via metabolic profiling by targeted methods (Z. Wang et al. 2011).

Multiple reaction monitoring (MRM) is currently the gold standard for small-molecule quantitative analysis, which can be designed to provide accurate quantitation results of targeted environmental pollutants with great sensitivity. For example, Miossec *et al.* combined SPE with LC-MS/MS to analyze 44 pharmaceuticals in environmental samples (Cai et al. 2009). In MRM mode analysis, analytes with targeted m/z (precursor ion) are ionized and selected for collision and sequential dissociation, resulting in multiple fragments (product ions) with unique m/z and intensity. The retention time and the ratio of the intensity of the mass transitions are used to identify chemicals accordingly. Two or more abundant product ions are selected for analytes quantification and qualification (Fig 2.4A). Since the parameter setting of MRM transitions was highly dependent on the authentic standards, the method was not suitable for screening many analytes. The lack of internal standards also resulted in many false positive peaks and inaccurate quantitation analysis results. Therefore, targeted detection by the MRM method with similar coverage to the non-targeted method is a wish for scientists in the field of the exposome.

Researchers in exposome studies seek to overcome drawbacks in targeted analysis through variable efforts. Chemical coverage was increased by multiple

samples run across different mass spectrometry platforms (LC-MS/MS and GC-MS) (Chung et al. 2018) and advanced QqQ instruments with dynamic MRM methods (Dresen et al. 2010; Mueller et al. 2005). Relatively high-throughput and scalable targeted analysis was achieved by fast and reliable express liquid extraction together with HR-MS (GC/LC) for volatile chemicals (Hu et al. 2021). Apart from these breakthroughs in experimental practices, the data mining of existing biomonitoring data gains attention from investigators, especially the chemical mass spectral databases.

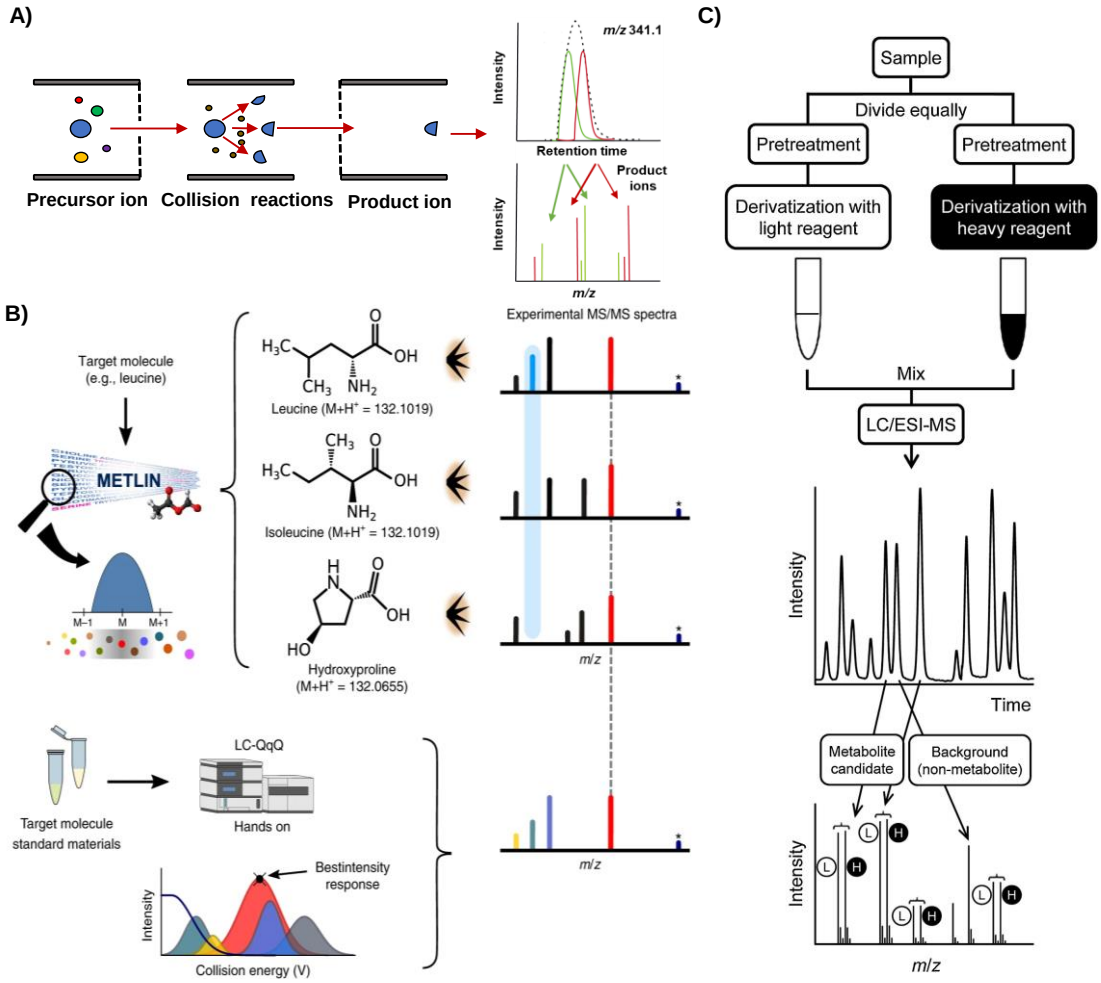


Figure 2.4. The schematic diagram of the MRM methods in the targeted analysis of human exposomes. (A) Extracted ion chromatogram (EIC) and total ion chromatogram (TIC) from MRM fragmentation. (B) The computational and experimental optimization of MRM methods in METLIN MRM database. (C) Scheme of metabolites profiling procedure based on chemical isotope labeling (Jones; Domingo-Almenara et al.; Higashi and Ogawa).

2.3.1 Computational optimization of MRM method without chemical standard

One key step for circumventing the necessity of standards is to determine MRM transitions with optimized MS instrumental parameters that work well with existing instruments and provide high selectivity. Collision energy (CE) is the most important instrument parameter that is optimized to increase fragmented ion intensity and different instrument vendors provide users with various CE parameters. MRM transitions and CE of many targeted analytes have been widely reported across research laboratories. For example, METLIN MRM database provides MRM transitions for more than 15,500 molecules and facilitates data sharing across different instruments and laboratories (Fig 2.4B). CE is optimized in experiments using chemical standards. CE with the best intensity response is recorded in the database. However, it is questionable whether those parameters are transferrable between instruments. There is a lack of computational optimization of CE across different instruments.

2.3.2 The development of MRM and pseudo-MRM spectral databases

The requirement for chemical standards and prior knowledge of MRM transitions prevents conventional targeted analysis from using MRM covering more chemicals and high throughput screening. To overcome these limitations, the establishment of MRM databases has been trending in the mass spectrometry field. MRM databases are developed to provide a compendium of MRM assays for small molecules and streamline the targeted qualification and quantification workflow. Through MRM data sharing across research laboratories around the world, larger coverage of chemical space and easy access to MRM transition of targeted analytes can be achieved, facilitating the MRM optimization procedures.

Tremendous effort has been pioneered in the compiling of optimized MRM reactions for targeting chemicals. Large commercial libraries such as NIST and Wiley are available with curated spectra and enriched contents. Freely open libraries such as MoNA, Massbank, and GNPS, have developed and specifically focus on data sharing and community efforts. Commonly used 20 MS libraries cover a wide range of chemicals varied from lipids to environmental small molecules. LipidBlast records the most compounds and MS/MS spectra: 119200 compounds, and 212516 MS/MS spectra. MetaMS records the least compounds and MS/MS spectra: 150 compounds,

and 150 MS/MS spectra (Fig 2.5). MRM transitions for over 15,000 small molecules across different instruments are provided via cloud-based XCMS and METLIN MRM databases (Domingo-Almenara et al. 2018b). The identification of metabolites with matrix interferences can be confirmed by the retention time data in MRM databases (Fig 2.4B). The optimized MRM transitions can be selected using fragmentation data in the METLIN-MRM database (Xue et al. 2021).

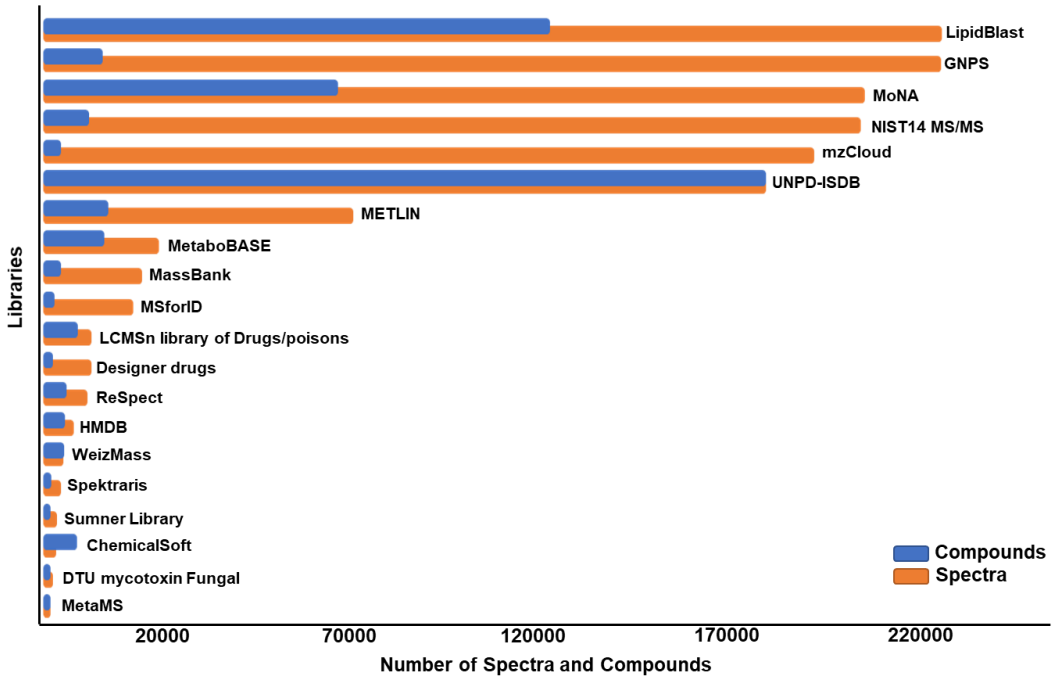


Figure 2.5. The number of compounds and their MS/MS spectra available in databases

Pseudo-multiple reaction monitoring (Pseudo-MRM) merges the advantages of both non-targeted and targeted detection methods. It is usually developed by transforming a non-targeted compound profiling method into a pseudo-targeted MRM method. In detail, the ion pairs of pseudo-MRMS of these methods were acquired from the real samples through non-targeted tandem MS/MS by HRMS or directly from MS/MS compound database (Xu et al. 2019). Pseudo-MRM has been widely used in metabolomics due to its high sensitivity, high coverage, and low dependency on chemical standards. For example, Zheng *et al.* realized a semi-quantitative analysis of almost 1,000 metabolites by pseudo-MRM after their MRM transitions were collected under product ion mode by UHPLC-HRMS at different collision energies (F. Zheng et al. 2020).

2.3.3 The application of chemical isotope labeling (CIL) in MRM methods of targeted analysis

The concentrations of environmental pollutants are much lower than those of metabolites in biological samples, which makes it difficult to acquire efficient MS/MS information from real samples. In addition, there are few powerful MS2 databases of environmental pollutants, and the current database can only cover less than 10% of the known environmental chemicals (Vermeulen et al. 2020b); their MS2 fragmentations are not easy to predict due to various chemical structures when corresponding standards are not available. Peak assignment and peak integration are other challenges, especially when faced with thousands of peaks in many samples, which is too time-consuming (Melnikov et al. 2020). To date, there is no effect pseudo-MRM method in the exposome field.

The chemical isotope labelling-LC-MS (CIL-LC-MS) strategy can solve the above problems to some extent. In CIL-LC-MS methods, a pair of isotopes labeled reagents are used to tag with reactive functional groups before generated products are subjected to LC-MS analysis (Qi et al. 2014). The strategy can further increase the detection sensitivity of analytes, improve their chromatographic peak shapes, and provide one-to-one internal standards to reduce false positive rates (Winter et al. 2018). For example, differentiating between background peaks and the peaks of targeted metabolites is challenging. CIL strategy is one of the effective solutions for this difficulty. The test sample was split into two equal aliquots, as seen in Fig. 2C, and derivatized using either the light or heavy reagents, such as the $^{13}C_0$- or $^{13}C_4$-coded reagent. The aliquots were then combined and injected into the LC/MSMS for analysis. In contrast to background peaks and underivatized metabolites, all derivatized metabolites (true metabolites) produced peak pairs in the mass spectra. The true metabolite peaks can be easily identified because the $^{13}C_0$- or $^{13}C_4$-coded metabolites coeluted with the characteristic mass difference (in this case 4 Da × number of functional groups (derivatized sites) in the metabolite). Therefore, the CIL method can make it easier to find and recognize metabolite peaks in the mass spectra amid other peaks.

To date, it has been used for the detection of exposure biomarkers (Luo et al. 2018). For example, dansyl chloride (DnsCl) was used by Chang *et al.* to develop the CIL-LC-MS method of 9 hydroxylated polybrominated diphenyl ethers (OH-PBDEs).

After DnsCl derivatization, the peak shape of OH-PBDEs was improved and the detection limit was pg/L in wastewater samples (Chang et al. 2012). Another advantage of CIL is that the introduction of derivatization reagents to the original analytes makes corresponding derivatization products possess similar chemical skeletons, resulting in consistent MS/MS fragmentations (Guo et al. 2017; P. Liu et al. 2014).

2.3.4 Circumventing the necessity of chemical standards with the assistance of HRMS and retention time prediction

Besides MRM in LC-QqQ-MS/MS, LC-HRMS-based unknown identification has become a trending area of environmental studies in past years and many unsuspected chemicals have been reported in the environmental samples (Andra et al. 2017a; Chindarkar et al. 2015; Schlittenbauer et al. 2016). However, one of the biggest challenges is to acquire high-quality MS/MS fragmentation of low abundant parent ions in environmental or biological samples (Cao et al. 2015a; Hagiwara et al. 2010). For most xenobiotic molecules in complex samples such as wastewater and biological fluids, the concentration is usually in the part per trillion (ppt) or part per billion (ppb) levels and therefore, the acquisition of MS/MS fragmentation with high quality is highly challenging. On the other hand, isobaric ions in the isolation window interfere with the MS/MS fragmentation of targeted molecules. Though the MS/MS spectrum of environmental chemicals remains limited, multiple reaction monitoring (MRM) transitions are widely reported in the scientific literature or commercial databases (e.g., Agilent Pesticide Dynamic MRM Compound Database and Waters Pesticide Dynamic MRM Compound Database). Nevertheless, it should be noted that the sensitivity of MRM is commonly at least several folds higher than the HRMS in complex samples (Yap 2011; Zisi et al. 2017). LC-QqQ-MS/MS is often accessible in many research laboratories but not necessarily from the same vendors.

In addition to accurate mass and fragmentation patterns, retention time (RT) prediction serves as another tool to qualify small molecules. For example, RT prediction models based on quantitative structure-retention relationship (QSRR) modeling are commonly used in metabolome analysis and unknown identification. In QSRR modeling for RT prediction, experimental RT was correlated with molecular descriptors (MDs) by a function developed by Multiple Linear Regression (MLR) and other machine learning algorithms such as genetic algorithm, neural network, and

random forest regression tree (Beyer et al. 2018; Horning and Weber 1985; Warth et al. 2018). MLR has been widely accepted in QSRR modeling yet suffers difficulty in searching for good predictors from a huge dataset of MDs unless given a long time for processing the data. Random forest decision trees, on the other hand, is capable of handling a large number of variables such as MDs, which are either continuous or discrete. Therefore, data mining of retention time from analyzed compounds using feature selection by random forest and RT prediction by MLR modeling of the RTs between selected molecular descriptors (MDs) will be of great help to further rule out false-positive signals without using standards.

2.4 HR-MS-based non-targeted analysis of human exposome

As target analysis requires prior knowledge of known compounds, NTA approaches using high-resolution mass spectrometer coupled with liquid chromatography (LC-HR-MS) have been increasingly used to monitor micro-pollutants with more chemical coverage (M. Dong et al. 2021; Hollender et al. 2017b; Newton et al. 2018; Plumb et al. 2006; Ruff et al. 2015). The feasibility of monitoring short-lived endogenous and exogenous compounds via high-throughput non-targeted approaches is critical to exposome studies(Jones 2016). HR-MS-based non-targeted analysis can measure thousands to tens of thousands of chemical features in one analytical run, even though most of the features are unannotated, suitable for exposome-wide association studies (EWAS).

For instance, in the environmental and biological monitoring, the non-targeted workflow typically starts with (1) the method development of appropriate sample preparation (solid-phase extraction and liquid-liquid extraction); (2) LC-HR-MS analysis involves MS acquisition in full MS scan mode, containing mostly protonated or deprotonated molecular ions, plus MS/MS or MS^n data, where the fragmentation patterns are extracted for devolution in data pre-processing; (3) Data pre-processing include peaks detection, annotation, and alignments of peaks across samples to subtract compounds signals present in blanks; (4) Prioritization of specific profiles for further evaluation is conducted through sample-wise profiles by peaks alignment, principal component analysis, and clustering analysis with gradients of time, space, or treatment; and (5) finally, identification of prioritized profiles relies on MS or MS/MS

spectra library searching and matching the experimental analytical properties with *in-silico* predicted values for further confirmation (Fig 2.6).

The rich data sets from the unknown chemical screening of various human specimens (such as urine, blood, and hair) offer a path for discovering health-impairing exposures. A dynamic exposome interaction network between humans and other ecosystems was developed, which revealed a previously unrecognized type of hazardous exposure detected in the air (Jiang et al. 2018). Persistent chemicals, including polychlorinated bisphenols, organochlorine pesticides, and non-persistent chemicals, including pesticides, surfactants, and personal care products, have been well characterized by HR-MS-based unknown chemical screening (Broecker et al. 2011; Helfer et al. 2015; Hernández et al. 2009; Roca et al. 2014). Non-targeted metabolomics studies widely rely on this approach to detect endogenous and exogenous chemicals, making it a critical component of exposome characterization (Niedzwiecki et al. 2019).

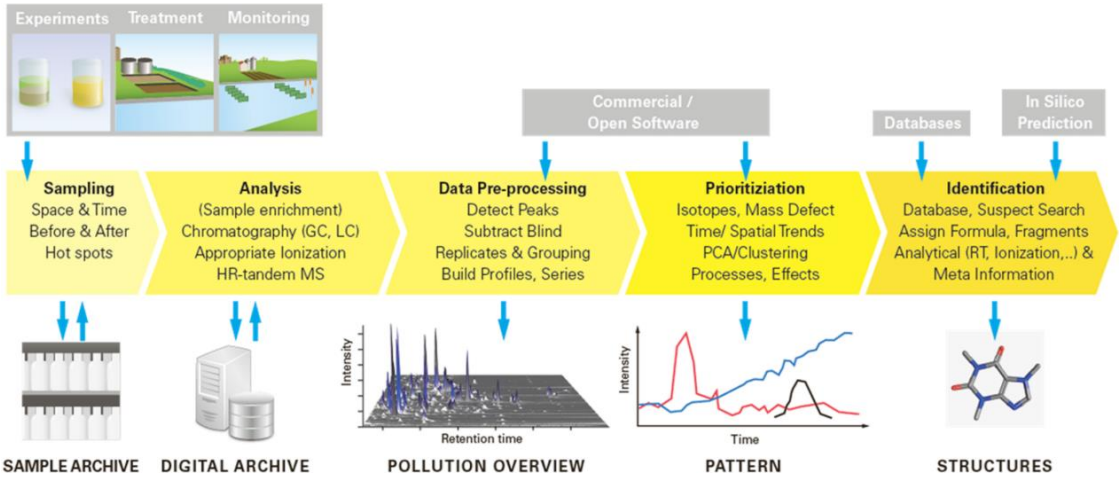


Figure 2.6. Generic workflow for non-targeted analysis (Hollender et al. 2017a).

There are typical or emerging non-targeted and suspected workflows for exposomic characterization: (1) suspect chemical screening (biased non-targeted analysis) and (2) unknown chemical screening (unbiased non-targeted analysis). For example, non-targeted profiling based on high-resolution mass spectrometry (HRMS) generally aims at detecting as many environmental pollutants as possible in biological samples (David et al. 2021; Dodds et al. 2021; Takahashi et al. 2018). In Kiefer's study, an LC-HRMS in a data-dependent acquisition mode was used to obtain data in both positive and negative ionization modes to discover new pesticides and their

transformation products (Takahashi et al. 2018). Though its wide detection coverage and powerful resolution, the NTA method can be challenging from several perspectives, depending on the availability of chemical standards and mass spectra libraries (Bendik et al. 2021; Dhungana et al. n.d.; Ulrich et al. 2019). Besides, NTA methods also suffer from limited sensitivity, complex data processing, and narrow linear range.

2.4.1 Improving the confidence of identification using multiple filters in non-targeted analysis for exposome characterization

In suspect chemical screening, suspect compound lists with prior knowledge of compound structure information are required. The required information includes molecular formula, physical and chemical properties for retention time reference, and molecular structure for MS/MS fragmentation pattern inference. The prior knowledge serves as multiple screening filters to improve the confidence level of compound identification. MS/MS fragmentation and isotope patterns of precursor ions are used to identify compounds by structure elucidation. The confidence of unknown compounds identification can be summarized in five different levels regarding the amount of evidence provided by the features matching (Schymanski, Jeon, et al. 2014). MS and MS/MS spectra matching alone can only achieve compound identification with medium or low-level confidence. For example, as seen in Fig 2.7, many detections were acquired by LCMS analysis, in this case, 12051 suspect compounds. By RT correction and isomer distinction, the final analytical sample consisted of 4791 suspect candidates that matched 620 suspect chemicals.

As a supporting filter, retention time screening facilitates identification by matching the observed retention time with the in-silico retention time predicted by machine learning models. From our previous work in retention time prediction, the random forest algorithm is often used for its excellent prediction power in chemical and physical properties like retention time for small molecules based on SMILES structure and molecular descriptors (Yang et al. 2020). However, accurate retention time prediction is demanded as a retention time in most references was predicted based on local LCMS conditions (Bonini et al. 2020; Cao et al. 2015b; Zdravković et al. 2018). Tedious models' modification is required for applications.

2.4.2 Risk-based prioritization in the non-targeted analysis of chemical exposures

Compound identification by structural information alone can be difficult to highlight the pollutants which pose the most threat to human health. Another filter is required to prioritize those pollutants among chemical mixtures present in the environmental samples (Escher et al. 2020). Risk-based prioritization in non-targeted characterization has been performed for years. To this date, most chemical prioritization strategies are based on data-driven techniques, such as simple ranking of signal intensity, frequency of occurrence in a data set, and searching for masses of expected compounds in the sample (Hug et al.; Thurman et al.; Schymanski et al.). For example, as seen in Fig 2.7, the filter list of 620 suspect candidate chemicals was further screened by multiple prioritization criteria, such as the detection frequency difference between different groups of samples (in this case, female firefighters, and office workers) and peak area (an indicator of higher relative concentration). Finally, the candidate list of 620 candidates was narrowed down to 71 for further confirmation.

On the other hand, toxicity prioritization can apply to library matching hits based on identification confidence and compound toxicity to highlight environmental pollutants that pose the most harm to human health. Toxicity ranking in NTA methods is driven by controlled laboratory experiments, using biological effect tests to prioritize unknown components associated with specific toxic effects for identification. However, this process could be very time consuming and offers little understanding of the relationship between effects and compounds. Chemical prioritization based on each criterion alone might be misleading. There is still a lack of high-throughput prioritization platforms to link the fragmentation information and chromatography behaviour to toxicity effects (Jonker et al.; Helbling et al.; Weiss et al.). Meanwhile, in a recent study, toxicity prioritization is performed by identifying the specific MS/MS fragments that can be representative of common toxicity using deep learning methods, resulting in a prioritized list for identification (Meekel et al. 2021). However, we still do not have sufficient knowledge to link the fragments with the toxicity, especially considering the different toxicity endpoints.

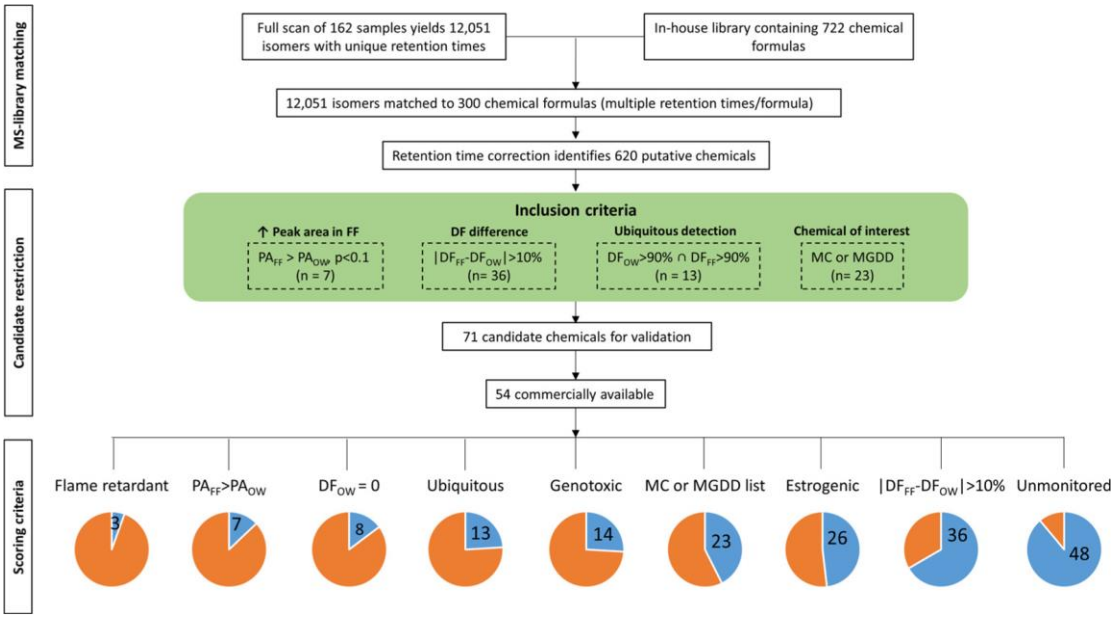


Figure 2.7. Non-targeted analysis using multiple filters for exposome characterization (Grashow et al.).

2.5 Summary of Literature Review

In summary, the literature review reveals the following knowledge gaps:

1) Traditional targeted approaches based on LR-MS can only apply to a limited range of chemical exposures and encounter multiple limitations: (i) the number of targeted chemicals is limited due to the limited coverage in one sample run; (ii) Extensive sample cleanups are required to remove matrix effect from samples; (iii) Chemical standards of target compounds are required for quantification in samples and MRM optimization; and (iv) target chemicals are required so that MRM transition optimization is feasible.

2) The current non-targeted analysis encounters multiple limitations: (i) the identification of chemical exposures requires comprehensive strategies and integrated platforms for data processing; (ii) the existing prioritization strategies do not link the identification of chemicals to their risk of exposure.

CHAPTER 3: THE DEVELOPMENT OF MRM METHOD USING IN SILICO OPTIMIZATION FOR FAST SCREENING OF ENVIRONMENTAL CHEMICALS

3.1 Summary

Chapter 3 first proposed a generic methodology for MRM methods transfer between different LC-QqQ-MS/MS platforms to analyze small molecules in environmental and biological samples. The method was based on a collective strategy including building up a library comprised of MRM transitions and CE values, peak matching between LC-QqQ-MS/MS and LC-HR-MS, and QSRR retention time prediction with random forest feature selection (Pearson r^2 = 0.63, prediction error = ±1.35min). Conversion equations for CE values were developed, whereby the CE values of the same compounds employed on different LC-QqQ-MS/MS platforms could be converted into nearly optimized values for the local equipment. The *in-silico* prediction of optimized collision energy renders reliable and fast access to MRM method optimization across different instruments. It was successfully applied with conventional targeted methods for environmental exposure characterization without chemical standards.

3.2 Introduction

Liquid chromatography-triple quadrupole-tandem mass spectrometry (LC-QqQ-MS/MS) or high-resolution mass spectrometry (LC-HRMS) with the electrospray ionization interface (ESI) has been the most widely used methods for identification of small molecules and their metabolites in environmental, agricultural, biological, and medical samples (Kolpin et al. 2002; E Naegele 2013; Rodil et al. 2005). In LC-QqQ-MS/MS analysis, MS/MS data acquisition is performed in multiple reaction monitoring (MRM) to achieve higher sensitivity and selectivity, whereby two or more abundant product ions are selected for quantification and qualification. The optimization of MRM parameters needs to be performed with standards traditionally, whereas standards are often costly yet not always available.

There is great potential to make use of the high sensitivity of LC-QqQ-MS/MS for the qualification and quantification of small molecules and establish a method with good transferability across different platforms without chemical standards. Various limitations need to be attended to for facilitating the application of LC-QqQ-

MS/MS. First of all, the collision energy of the MRM method requires multiple steps of optimization for local equipment using chemical standards. Experimental settings from existing publications are not able to meet the demand for fast implementation of MRM-based detection across different platforms. Second, spectra searching by MRM ion pairs could end up with multiple hits and therefore, additional filters are needed to rule out the false-positive signals.

In this study, we aim to develop a practical platform that applies predicted CE values for specific MRM transitions from existing different instruments to identify small molecules in complex environmental samples with the assistance of LC-HRMS and RT prediction. Firstly, we collected MRM transitions and their CE values for up to 152 small molecules across various instruments and derived linear regression equations to convert CE between platforms. Secondly, we acquired and optimized CE values of MRM transitions using the on-column injection of real samples. We further utilized peak matching between LC-QqQ-MS/MS and LC-HRMS together with a QSRR retention time prediction model to rule out isobaric candidates. Lastly, the developed platform was validated by identifying 20 typical small molecules in complex matrix samples including sludge water, human urine samples, and cell extracts.

3.3 Methods and materials

3.3.1 Chemicals and reagents

All chemicals (see Table S3.1 and Table S3.2) and solvents were purchased from Sigma Aldrich (Singapore) with a purity > 97% unless stated otherwise. Ultrapure water was obtained by Milli-Q integral water purification system (Merck, Singapore). A sludge water sample was collected from the supernatant of the waste active sludge from a water reclamation plant. Stock solutions of 20 micropollutants were prepared by dissolving a certain amount of each substance separately in a mixture of water and methanol (1:1, v/v). The mixture solution of 20 standards was prepared by mixing the stock solution. The pharmaceutical compounds were selected from an FDA-approved drug library purchased from ApexBio (ApexBio, USA). Two vials of mixture solution comprising 85 and 48 drugs standards were prepared in a mixture of water and methanol (1:1, v/v).

3.3.2 LC-MS instrumentation

The chromatographic separation was performed using Agilent 1290 Infinity II Binary LC system coupled with Agilent 6460 triple quadrupole (QqQ) mass spectrometer and Agilent 1290 Infinity I Binary LC system coupled with Agilent 6550 iFunnel quadrupole time-of-flight (QTOF) mass spectrometer (Agilent, USA). Both mass spectrometers were fitted with electrospray ionization (ESI) interfaces in positive and negative modes with Agilent jet stream technology. The analytes were separated on a Waters Atlantis T3 reverse phase column (3 µm, 100 mm × 2.1 mm, Waters, USA) for MRM analysis in ESI positive and negative modes using mobile phase: water with 0.1% formic acid (A) and acetonitrile with 0.1% formic acid (B). The linear gradient elution started from 95% A (0-3.0 min) to 95% B (14.0-17.0 min), and then reverted to 95% A (17.1-25.0 min). The flow rate was 0.2 mL/min, and the temperature of the column compartment was set at 30 °C.

In MRM transition modes, ESI source conditions were set as follows: gas temperature 300 °C, drying gas flow rate 10 L/min, nebulizer 25 psi, sheath gas temperature 250 °C, sheath gas flow rate 10 L/min, delta EMV 200 volts (±), and capillary voltage 3500 volts in positive and negative mode. In QTOF, gas temperature 200 °C, drying gas flow rate 14 L/min, nebulizer 35 psi, sheath gas temperature 350 °C, sheath gas flow rate 11 L/min, capillary voltage 3500 volts, nozzle voltage 1000 volts, fragmentor voltage 175 volts, and skimmer voltage 65 volts in ESI positive and negative mode. The instrument was set to acquire over the mass-to-charge ratio(m/z) range 100-1100, with the MS acquisition rate of 1 spectra/s. The specific MRM transitions and ESI modes for 20 chemical standards were selected based on previous studies and our optimization results (See Table S3.1).

3.3.3 Retention time prediction modeling

A quantitative structure-retention relationship (QSRR) was developed from a data set consisting of experimental retention time data of 133 FDA-approved drugs for the reverse phase LC-MS analytical platform and the molecular descriptors calculated by PaDEL-Descriptor. PaDEL-Descriptors is open-source software dedicated to molecular descriptors calculations. Up to date, it allows users to retrieve overall 1,875 descriptors based on the input of molecules information in the sdf format (Yap 2011). Important molecular descriptors were selected by random forest

feature selection with 10-fold cross-validation. A multilinear regression QSRR prediction model was selected based on its superior performance. The modeling process was conducted in R environment.

The entire data set was randomly divided into a training set (n = 93) and a test set (n = 40). In-house development of a predictive QSRR model started with collecting retention time data of 133 drugs mixture from accurate mass measurement on LC-QTOF (See Supplementary file). The calculation of molecular descriptors (MDs) of 133 drugs was then performed. Specifically, canonical SMILES structural representations of 133 drug standards were obtained from the drug vendor (ApexBio, USA). SMILES of each molecule in its neutral form was further used to acquire its information in the sdf file format by the R package "rcdk".

A total number of 1445 1D and 2D MDs were calculated for each molecule. Many of those MDs were either highly correlated with each other, constant and none values. Therefore, the reduction of the size of descriptors was performed manually in R. Molecular descriptors that are highly correlated (Pearson correlation no less than 0.9), nonvalues, and constants were removed. As a result, 489 molecular descriptors were retained, which could still lead to the over-fitting issue. Thereafter, we further applied a feature selection using a random forest with 10-fold cross-validation to determine if any subsets with better predictivity were available to offer a better explanation for the observed retention time of standards (Cao et al. 2015a). Feature selection by the random forest in the R package "randomForest" was performed to find the most significant molecular descriptors. The process was applied to training set data to generate predictors for our final model. MDs were evaluated based on root mean square error (RMSE) of 10-fold cross-validation and their contribution to the model performance calculated by random forest. Finally, the sizes of variables for our model were allowed to be as small as 10 variables because when the size of variables reached above 10 (see Table S3.3). No significant improvement was achieved according to the RMSE value. According to the importance rating by random forest, 10 molecular descriptors (XLogP, BCUTp.1h, AATS1i, AATS3i, GATS1e, CrippenLogP, AATSC0p, ETA_EtaP_B, AATS4i, AATS5i) were retained because they made the top contribution to mean square error (MSE) than the rest of the other MDs. A prediction model was trained based on 93 training data sets. The model was assessed based on its RMSE value acquired by 10-fold cross-validation with 40 test

data sets. In the application case, canonical SMILES structural representations of 20 micropollutants, bisphenol A, and two of its metabolites were extracted from the PubChem (https://pubchem.ncbi.nlm.nih.gov/).

3.3.4 Preparation of synthetic and sludge water sample

Synthetic water sample with a pH value of 7.4 was obtained by dissolving 4 inorganic reagents and humic acid into MILLI-Q® water, detailed in Table S3.2 (Horning and Weber 1985). A total mixture of 20 compound standards was spiked into 1 liter of water with the environmentally relevant level at 0.2 ppb. The original wastewater sample was prepared from the supernatant of the waste active sludge. After filtration by 0.45 μm PTFE filters, the solution was spiked with 20 compound standards at a final concentration of 0.2 ppb similarly. Both water samples were concentrated by 1,000 times using solid phase extraction by Oasis HLB Cartridges (6 mL) and stored at -40 °C till analysis.

3.3.5 Bisphenol A and its metabolites in HepG2 cell line

HepG2 was grown in DMEM (Gibco Invitrogen, Singapore) containing 10% heat-inactivated fetal bovine serum (FBS; Gibco Invitrogen, Singapore), and maintained in a humidified 37 °C incubator with 5% CO_2. For the experiment, cells were seeded in 100 mm petri dishes containing 5 mL medium. After 12 hours, cells were either incubated either with common medium containing 0.1% DMSO as control or 10 μM BPA for another 48 hours. After medium was aspirated, cell was washed with 1 mL ice-cold phosphate-buffered saline (PBS; Gibco Invitrogen, Singapore) and harvested with a cell scraper in 2 mL ice-cold quenching mixture solution of methanol, acetonitrile and Milli-Q water (2:2:1, v/v/v). Samples were further extracted by three cycles of freeze-thaw method in liquid nitrogen with sonication (Beyer et al. 2018). Samples were then stored at -20 °C for 1 h and subsequently centrifuged at 13,000 rpm to precipitate out proteins. The supernatant was dried in a vacuum concentrator and reconstituted in ACN: H2O (1:1, v/v) (Warth et al. 2018).

3.3.6 Urine sample preparation

A urine pool sample (n = 33, a pooled mixture of urine samples from 33 participants) was from our previous study (M. Liu et al. 2019). Briefly, enzymatic

deconjugation was performed before liquid-liquid extraction. 5 mL of a urine sample was transferred into a 15 mL polypropylene tube followed by addition of 3 mL of 100 mM sodium acetate buffer which contained beta-glucuronidase. After incubation with shaking at 37 °C for 18 hours, the urine sample was extracted three times with ethyl acetate and concentrated to near dryness under vacuum in a Speed vac (Labconco, Inc.Singapore). The sample was further reconstituted to 200 μL methanol:water (1:1, v/v). The concentrated urine matrix was further spiked with the selected 20 compound standards to 0.2 ppm at human relevant level (i.e., 8 ppb in original urine sample).

3.3.7 Statistical Analysis

SPSS 12.0.1 was used for the statistical analyses. Collision energy (CE) values for all tested compounds were tested for normality using the Shapiro–Wilk test. The correlation between different instrumental platforms were conducted using Pearson correlation analysis.

3.4 Results and discussion

3.4.1 The workflow of MRM transfer and its application in small molecule identification coupled with HR-MS and RT prediction

The workflow of small molecules identification and MRM method transfer is shown in Figure 3.1. In Step 1, parameters for MRM and corresponding CE used in LC-QqQ-MS/MS were extracted from at least three previous literatures if available. In Step 2, a general LC-QqQ-MS/MS method was set up to optimize CE using real samples with chromatographic separation. Furthermore, a preliminary identification was achieved by matching the similarity of peak shape and retention time of multiple product ions (if available) generated in MRM modes in Step 3, which worked as the first filter for screening putative signals. In Step 4, we made a further confirmation of the detected signals by using LC-HR-MS as the second filter. Retention time matching and peak alignment based on the similarity between LC-HR-MS, and LC-QqQ-MS/MS proved to be another solid filter to determine positive detections from putative results. Finally, we performed an additional filter on false-positive signals by using a home-developed QSRR RT prediction model in Step 5. Those signals that passed the afore-mentioned filters were considered as positive identifications. In

summary, the workflow aims to transfer MRM methods between instruments and analyze small molecules in a manner with the circumvent of the authentic standards.

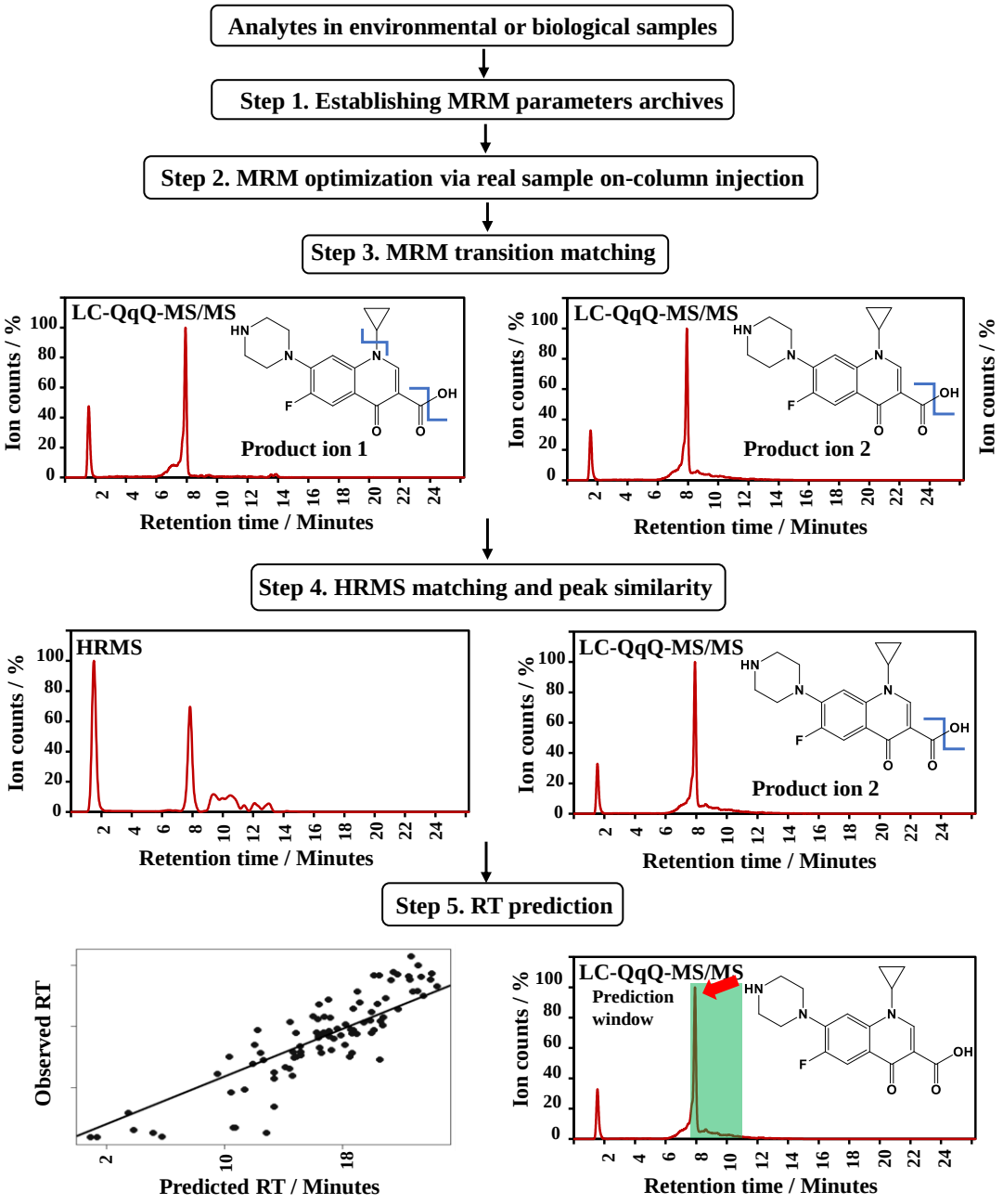


Figure 3.1. General workflow for MRM transfers between different LCMS platforms**.** Ciprofloxacin was selected as an example in the workflow. Highlighted green area denotes predicted retention time window from the QSRR prediction model. "RT" represents retention time.

3.4.2 Collection of the MRM transitions, CE values from existing publications and MRM transfer

MRM transitions of 20 small molecules were collected from previous publications as an example to validate the platform (Table S3.1). As shown as examples in Figure 3.2a, the most abundant product ion (product ion 1) in MRM transitions of 7 representative compounds was acquired in multiple studies, including ampicillin (AMP), tris(2-chloroethyl) phosphate (TCEP), bisphenol A (BPA), triclosan (TCS), triclocarban (TCC), ethyl paraben (EP), and methyl paraben (MP). The second most abundant product ion (product ion 2) in MRM transitions of 6 model compounds except triclosan was listed in Figure 3.2.b. Overall, identical MRM transitions of each selected small molecules were commonly used in different studies on various instruments. For example, identical product ions are used for MRM transitions of triclosan (Anumol and Snyder 2015; Boleda et al. 2013; Imma Ferrer 2008; Ren et al. 2016; Rodil et al. 2009; R. S. Zhao et al. 2011), triclocarban (Anumol and Snyder 2015; Klein et al. 2010; R. S. Zhao et al. 2011), methyl paraben (Geens et al. 2009; Jakimska et al. 2013; Ren et al. 2016; Vela-Soria et al. 2014) and ampicillin (Granelli et al. 2009; Imma Ferrer 2008; Kearney n.d.). For ethyl paraben, the variation of its MRM transitions in different studies is because of the low resolution of LC-MS/MS. As detailed in Table S3.1, product ion with m/z of 137 and 136 is the same ion fragment, as the typical quadrupole mass resolution is 0.7 Da (Domingo-Almenara et al. 2018a). For tris(2-chloroethyl) phosphate (TCEP), identical product ion with m/z of 63 was selected as the first product ion in 3 out of 4 studies and m/z of 99 was selected in 2 out of 4 studies. In studies of ciprofloxacin, 4 out of 5 studies selected product ion with m/z of 314 as the first product ion. For MRM transitions of tetracycline, 5 out of 7 studies selected m/z of 410 as the second product ion. In summary, identical product ions are commonly selected for small molecules and thus MRM transitions from previous studies could be directly applied in MRM mode optimization, saving great efforts for users in determining MRM transitions experimentally.

MRM transitions performed on optimized CE could provide high sensitivity and generate data with high reproducibility. It is of great interest to investigate the CE variation and their transferability across different instruments. As such, we acquired CE values for analysis of 152 pharmaceutical compounds on Agilent 6460 QqQ and

Thermo TSQ, 62 pharmaceutical compounds on Waters TQS, 136 pharmaceutical compounds on ABS qtrap, and 116 pharmaceutical compounds on Waters TQD. As shown in Figure 3.2c-d, the correlation analysis demonstrated that CE values on Agilent QqQ and Thermo TSQ were closely correlated (Pearson $r^2 = 0.78$, $p < 0.0001$, Equation 1). Similarly, a high linear correlation was also revealed between CE values on Agilent QqQ and Waters TQD (Pearson $r^2 = 0.66$, $p < 0.0001$, Equation 2). CEs employed on Agilent QqQ were also moderately correlated with those from Waters TQS (Pearson $r^2 = 0.48$, Figure S3.3a-b). The correlation could be described in the following two equations:

$$y = 0.68\ x + 9.58\ (r^2 = 0.78, p < 0.0001) \quad (1)$$

$$y = 0.60\ x + 9.17\ (r^2 = 0.66, p < 0.0001) \quad (2)$$

However, the correlation of CE employed on Agilent QqQ and ABS qtrap was found to be minor (Pearson $r^2 = 0.10$, Figure 3.3a), indicating a poor reproducibility of CEs between these two instruments. This poor correlation of CE values was expected due to different mass analyzers applied in two systems (Provencher et al. 2014b). Overall, our study revealed that the optimized CE performed on different LC-MS/MS instrument platforms especially the quadrupole system can be inter-convertible and two regression equations were provided for conversion of CEs between instrumental platforms.

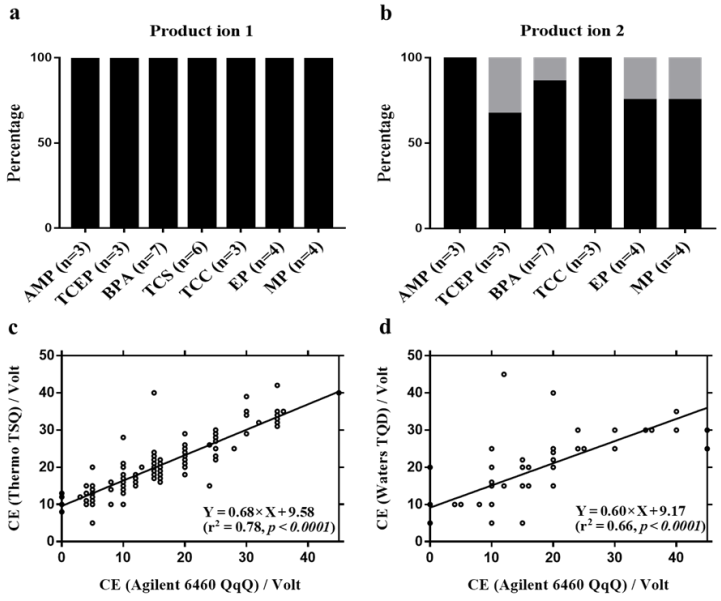


Figure 3.2. (a-b) The overlapping (%) of the two most abundant product ions (a: Product ion 1; b: Product ion 2) for selected environmental chemicals (as examples)

across different studies. (c-d) Correlation analysis of collision energy (CE) adopted across several instrumental platforms in MRM transitions. (a-b) The similarity of selected product ions for each environmental chemicals were depicted in percentage. The number of publications (n) where MRM transitions were extracted was denoted in parenthesis. (c) Linear regression analysis of CE adopted on Agilent 6460 QqQ and Thermo fisher TSQ (Pearson $r^2 = 0.88$, n = 152). (d) Linear regression analysis of CE adopted on Agilent 6460 QqQ and Waters Xevo TQD (Pearson $r^2 = 0.81$, n = 116).

3.4.3 Optimization of collision energy using an on-column real sample injection

The MRM optimization of small molecules using real samples is of great challenge. Unlike optimization with pure standards, it is challenging to optimize MRM parameters using direct flow injection by changing MS parameters such as CE and fragmentor voltage (FV) stepwise for analysis of small molecules in complex samples. The matrix effect such as ion suppression masks the analytes without chromatographic separation. Therefore, we must explore the method of MRM optimization using on-column real samples injection of matrix-spiked synthetic water. Tetracycline, malathion, triclocarban, and triclosan were selected as the model compounds to optimize this method (Fig 3.3a). Based on collision energy from the previous studies (Imma Ferrer 2008; Edgar Naegele n.d.; Pareja et al. 2011), we applied stepwise optimization during one sample running using a wide CE range of 0.2×, 0.4×, 0.6×, 0.8×, 1.0×, 1.2×, 1.5×, 2.0×, and 2.5× original value from references. For tetracycline, the signal with maximum MRM intensity was obtained when performed at the original value of CE in the published references (Fig 3.3a). The intensity of response varied in a small range, from 81% to 100% of the maximum intensity. Similarly, the maximal signal was also obtained on the original reported CE value for malathion. We achieved a maximal response of MRM transition for triclosan and triclocarban at 1.5×, and 0.8× the reported CE values, which were 114% and 112% of signal intensity acquired at the originally reported values; respectively.

In summary, our result demonstrated that maximum signal intensity, with acceptable variation, could be achieved when the MRM transition was performed on the CE obtained from references. CE library from references was a useful starting value for the MRM optimization. The optimization method via on-column infusion was an efficient and effective way to optimize the MRM transition in real samples.

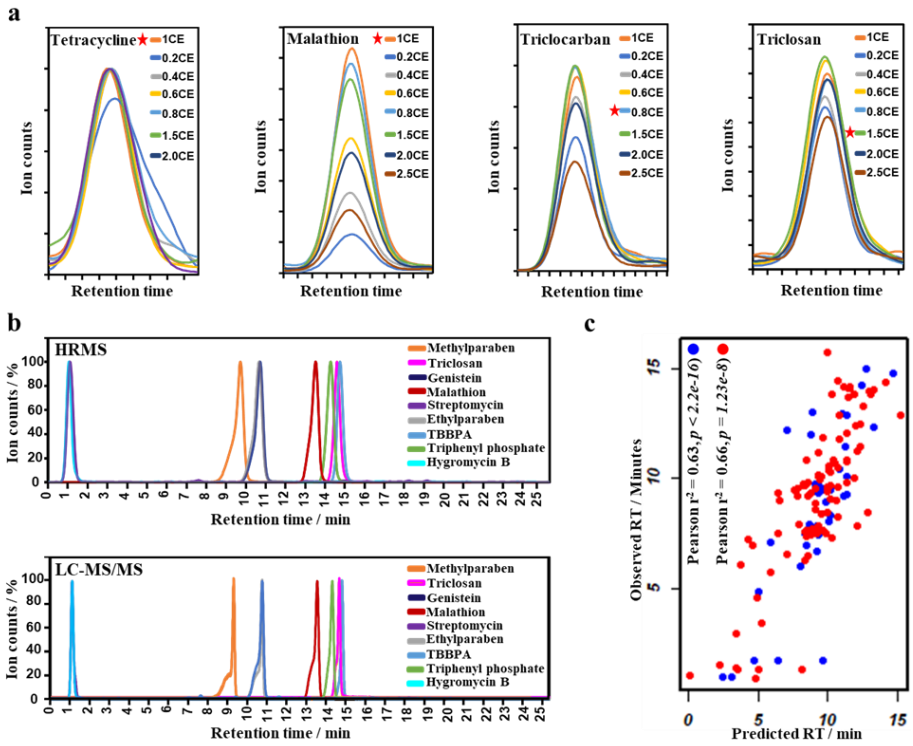


Figure 3.3. CE optimization of 4 selected compounds and performance of the prediction model. (a) CE optimization for 4 compounds performed on QqQ mass spectrometer using real sample on-column injection. (b) Well-matched extracted ion chromatography (EIC) from the analysis of spiked compounds in the synthetic water sample from HR-MS and LC-MS/MS. (c) Performance of the multiple linear regression model for retention time prediction. (a) Reference CE was derived from the averages from MRM archives and stepwise optimization was conducted. " " represents the CE values whereby we obtained the maximal intensity. (b) EIC of 9 representative compounds from analysis of spiked compounds in the synthetic water sample from HR-MS and LC-MS/MS. (c) Pearson r^2 between the predicted retention time and the observed retention time over training set as red dots, test set as blue dots was 0.66 ($p < 2.2E-16$), and 0.63 ($p = 1.23E-8$), respectively.

3.4.4 Peak alignment by shape similarity matching and retention time matching between MRM and HR-MS

As mentioned above, we managed to optimize the parameters of MRM transitions using the on-column real sample injection onto LC-QqQ-MS/MS. However, the extraction of one MRM transition usually results in multiple peaks with a complex matrix. To further eliminate those false positive signals, we used peak

matching between LC-HR-MS and LC-QqQ-MS/MS under the identical chromatographic condition. We included two criteria for peak matching in this study: geometric similarity of peaks' shape and well-matched retention time. Peak matching based on geometric similarity could be achieved by comparing the shape of the peak after overlapping acquired signals using vendor software. As depicted in Fig 3.3b, the signal of each compound from LC-QqQ-MS/MS and LC-HR-MS displayed a similar geometric distribution, indicating similar chromatography behavior under identical chromatographic conditions.

The retention time shift between the LC-QqQ-MS/MS and LC-HR-MS system is of great concern during peak alignment and compound identification. The time shift needs to be compensated if it is significant before cross-validation of identified signals between different systems when an LC method is transferred from one system to another. To characterize and quantify the retention time differences, we evaluated the retention time shift between two Agilent LC systems (Agilent 1290 Infinity II LC system in LC-QqQ-MS/MS and Agilent 1290 infinity I LC system in LC-HR-MS) through two strategies: calculation of the theoretical value of retention time shift based on instrumental parameters; characterization of retention time shift via peak alignment of 20 authentic standards between two systems under an identical chromatographic condition. The RT shift can be explained by the dwell volume differences between the two LC systems (Hong and Mcconville 2016), consisting mainly of the gradient mixer, the tubing between the pump and the injector, and tubing between the injector and the column (Table S3.4). According to our calculation based on instrumental parameters provided by the vendor, the dwell volume in QqQ system was 64 μL larger than that in HR-MS system, leading to a retention time difference of 0.3 min with the mobile phase speed at 0.2 mL/min. For the second part, we acquired the retention time of 20 authentic compounds using identical afore-mentioned LC methods. The data was summarized in Appendix Figure 3.4, which shows a box plot of the retention time shift of 20 selected compounds. The retention time data exhibited deviation in a narrow range between LC-QqQ-MS/MS and HR-MS systems. The maximum, minimum and median retention time shift was 0.057 min, -0.002 min, and 0.0415 min, indicating a minor shift during a total running time of 25 min. Overall, the experimental retention time shift was lower than the calculated value, indicating a minor retention time difference between these two in-house LC systems

with similar instrumental configuration. However, for mass spectrometry systems equipped with different LC systems (e.g., binary VS quaternary pump), a more conservative cutoff of RT shift should be considered in the peak alignments.

Due to the same ionization mode (ESI), MRM and accurate MS1 of the targeted compounds are supposed to be found on LC-QqQ-MS/MS and LC-HR-MS systems; respectively. However, the sensitivity difference of two LC-MS systems is another concern, which could cause the missing capture of the target compounds on one instrument platform and fail the peak alignment. To avoid this possibility, we used low and high injection volume in LC-QqQ-MS/MS and LC-HR-MS systems (e.g., 2 and 10 µL). The result showed that 20 targeted compounds can be found on both the MRM mode of LC-QqQ-MS/MS and MS1 scanning of LC-HR-MS.

3.4.5 QSRR retention time prediction and robustness validation

To achieve a reliable retention time prediction of small molecules, we have used a library of FDA-approved pharmaceutical compounds to develop a QSRR prediction model. A significant correlation between predicted RT and experimental RT was observed. A regression model was achieved as the following equation:

$$RT = 0.20 \times XLogP - 0.21 \times BCUTp.1h - 0.19 \times AATS1i + 0.05 \times AATS3i - 0.01 \times AATS5i - 0.10 \times AATS4i - 1.94 \times GATS1e + 0.26 \times CrippenLogP + 5.56 \times AATSC0p + 126.96 \times ETA\_EtaP\_B + 60.05 \quad \text{.....(3)}$$

As a result, the final MLR model has a multiple r^2 of 0.66 and an adjusted R^2 of 0.61. The top contributors of molecular descriptors are ETA_EtaP_B, AATSC0p, CrippenLogP, and XLogP, as depicted in Appendix Table S3.3. The predicted retention time (rtPred) was in a high correlation with the observed retention time(rtObserved) in the training set (r^2 = 0.66, p < 2.2E-16) (Figure 3.3c). As for the test set, lower correlation was obtained (r^2 = 0.63, p = 1.23E-8) (Figure 3.3c). The median absolute prediction error (|rtPred-rtObserved|) for the training set, test set, and the entire sample set was 1.28 min, 1.42 min, and 1.35 min; respectively. Therefore, a prediction window was determined as predicted retention time ± 2.0 min, which was 8% and 20% of error for a compound eluting at 25 min and 10 min, respectively, suggesting higher accuracy for the prediction of more hydrophobic compounds. Finally, molecular descriptors of 20 small molecules and BPA, BPA monosulfate, BPA glucuronide were calculated as described previously and the retention time was

predicted based on a QSRR model, as detailed in Appendix Table S3.6 and Appendix Table S3.9.

3.4.6 Method validation by micropollutant identification in sludge water and urine samples

The performance of our proposed platform was validated in the detection of 20 pre-spiked model micro-pollutants in sludge water and human urine at environmentally or human-relevant levels (Appendix Table S3.6). The sludge water offers a simple simulation of a complex environmental matrix in sample analysis. The urine sample provides the complex matrix effect and interferences that researchers commonly encounter in the case of biological samples. In our preliminary test, very few of the spiked chemicals had high-quality MS/MS using LC-HR-MS in those two types of samples. To illustrate our methodology in facilitating the identification of target compounds from multiple false-positive signals, the results from our study were categorized into three scenarios for discussions (Appendix Table S3.6). In the scenario I, good peak alignment across LC-QqQ-MS/MS and LC-HR-MS with further retention time prediction offered additional confirmation of compound identification. In Scenario II, multiple signals were observed in LC-QqQ-MS/MS and only a single signal was detected in HR-MS or vice versa. In either way, peak matching between LC-QqQ-MS/MS and LC-HR-MS offered a filter ruling out false-positive signals. Retention time prediction from MLR model provided further confirmation with higher confidence. In Scenario III, multiple putative signals were obtained both in LC-QqQ-MS/MS and LC-HR-MS for a target compound and we could make use of the MLR retention time prediction model for confirmation. Additionally, the last scenario also included compounds that were observed only in either LC-HR-MS or LC-QqQ-MS/MS.

The result showed that 19 out of 20 compounds were observed in the sludge water sample. Specifically, 11, 3, and 5 compounds fell into Scenario I, II and III; respectively. For ampicillin, multiple signals were observed on the LC-HR-MS platform at 7.2 and 8.7 min while the only single signal was observed in LC-QqQ-MS/MS at 7.3 min. The prediction window offered by our QSRR model ranged from 6.4 to 10.4 min. Therefore, there was good confidence to denote peak at 7.3 min as a positive signal (Fig 3.4a). Overall, positive signals of 17 compounds in all scenarios

were determined with further confirmation by results in LC-QqQ-MS/MS, LC-HR-MS and the MLR prediction model (Equation 3).

Similarly, 19 compounds were detected in the urine sample, along with 9, 6, and 4 compounds were observed in Scenario I, II, and III; respectively. Among them, positive signals of 18 compounds were confirmed by all three filters. Analysis of ciprofloxacin demonstrated a good application of our proposed method. Multiple putative signals were obtained both in LC-MS/MS at 1.4 and 7.5 min while two putative signals were observed in HR-MS at 1.4 and 7.5 min. The predicted retention time was calculated by the MLR model, RTPredicted = 8.9 ± 2.0 min, indicating that the signal observed at 7.5 min was positive and the others were false-positive signals (Fig 3.4b).

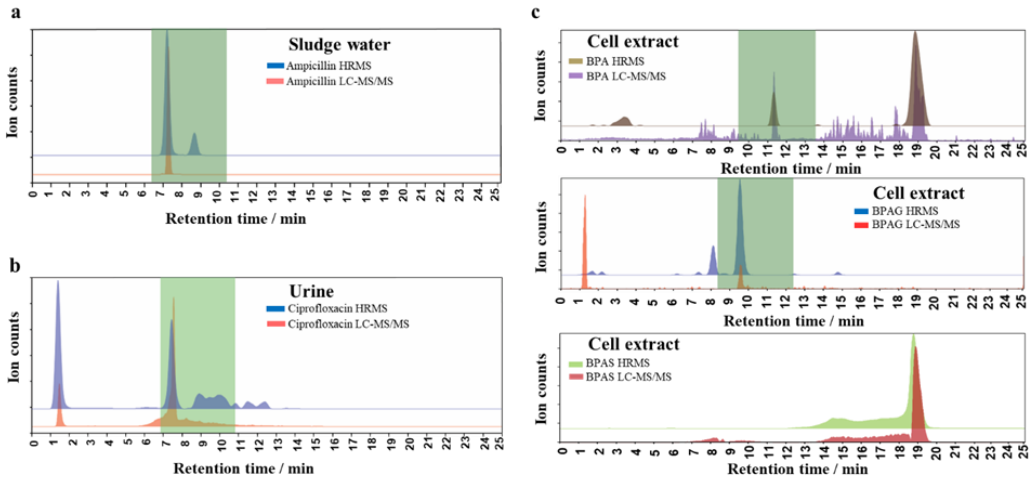


Figure 3.4. Method validation and performance in validation cases. (a-b) Examples of scenarios in the analysis of pre-spiked small molecules in complex samples. Highlighted box-shape area is predicted retention time window from the QSRR model. (c) Extracted ion chromatography (EIC) of MRM transitions and HR-MS of BPA, BPAS, and BPAG. (a) Putative signals of ampicillin in HR-MS, positive signal in LC-MS/MS, and predicted retention time in the sludge water sample. RTHR-MS = 7.2/8.7 min, RTLC-MS/MS= 7.3 min, RTPredicted = 6.4-10.4 min. (b) Putative signals of ciprofloxacin from HR-MS, LC-MS/MS and predicted retention time in the urine sample. RTHR-MS = 1.4/7.5 min, RTLC-MS/MS= 1.4/7.5 min, RTPredicted = 6.9-10.9 min. (c) Putative signal of bisphenol A in HR-MS, LC-MS/MS and predicted retention time in the cell extract. RTHR-MS = 11.5/19.0 min, RTLC-MS/MS= 11.5/19.1 min, RTpredicted = 9.5-13.5 min; Putative signal of BPA glucuronide (RTHR-MS = 8.3/9.7 min, RTLC-MS/MS= 1.3/9.8 min, RTpredicted = 8.2-12.2 min)

and BPA monosulfate (RTHR-MS = 19.0 min, RTLC-MS/MS= 19.2 min, RTpredicted = 13.1-17.1 min) in cell extract from HR-MS, LC-MS/MS, and retention time prediction, as depicted in highlighted area.

3.4.7 Method validation by the analysis of xenobiotic transformation metabolites in cell extracts

The methodology we proposed here demonstrated its advantage in analyzing three major targeted small molecules (BPA, BPA monosulfate (BPAS) and BPA glucuronide (BPAG)) in BPA metabolism inside HepG2 cell. MS/MS transition information and instrument parameters of the three compounds were first collected from previous publications (Appendix Table S3.7) (Battal et al. 2014; Provencher et al. 2014a). The original CE values of bisphenol A metabolites were adopted from the Waters Xevo TQ-S platform in our study and transformed into the Agilent platform using the mentioned equations above. After instrument parameters were collected, MS/MS transition optimization was performed to obtain the maximal signal at different CE. The highest response was observed when fragmentation was performed either at original CE or at a value close to the original CE (0.6× for BPAG), as detailed in Appendix Table S3.8.

The predicted retention time window for bisphenol A and its common metabolites was calculated by our MLR prediction model (Equation 3, Appendix Table S3.9). For analysis of bisphenol A (Figure 3.4c), the prediction model was applied to rule out false positives. Sample analysis in LC-HR-MS obtained two signals at 11.5 min and 19.0 min respectively. Likewise, two signals were observed at 11.5 min and 19.1 min in LC-QqQ-MS/MS. The predicted retention time of bisphenol A fell into the range from 9.5 min to 13.5 min, as highlighted in the green area in Figure 3.4c. Therefore, the signal observed at 11.5 min was denoted as the positive peak while other peaks obtained at 19 min could be excluded (which was generated from the in-source fragmentation of BPA metabolites). Analysis of bisphenol A standard in LC-QqQ-MS/MS and LC-HR-MS validated the former peak as a positive result (see Appendix Table S3.9).

For BPA glucuronide, MRM analysis detected two putative signals at 1.3 min and 9.8 min. Results obtained from LC-HR-MS indicated putative signals at 8.3 min and 9.7 min. Predicted retention time indicated that BPA glucuronide should elute

from 8.2 to 12.2 min, as highlighted in the green area in Figure 3.4c. Peak alignment and predicted retention time suggested that the observed signal at 9.7 min should be the positive peak. Further analysis of the standard confirmed the identification (see Appendix Table S3.9). In the analysis of BPA monosulfate, no false-positive signals were detected in LC-QqQ-MS/MS and LC-HR-MS. Good peak alignment indicated a solid identification from LC-QqQ-MS/MS and LC-HR-MS. However, the predicted retention time was much smaller than the experimental value. The inaccurate RT prediction of phase II metabolites could be due to the strong ion-exchange interaction and retention of highly charged functional groups such as sulfate on the column, which was not considered in the retention time prediction model. Overall, examining BPA and its multiple metabolites by our platform successfully solidates that the standard-free MRM method is rigorous for sample analysis with complex background.

3.5 Conclusion

We have proposed a generic methodology for MRM method transfer between different LC-QqQ-MS/MS platforms to analyze small molecules in environmental and biological samples based on a collective strategy including building up a library comprised of MRM transitions and CE values, peak matching between LC-QqQ-MS/MS and LC-HR-MS, and QSRR retention time prediction. We have demonstrated via direct on-column real sample injection that the direct employment of instrument parameters for MS/MS transitions offers an efficient way for method optimization without demanding standards. CE values for the same compounds employed on different LC-QqQ-MS/MS platforms share a statistical correlation, which could be converted into nearly optimized values using a proposed conversion formula (Equation 1,2). Peak alignments combine the high sensitivity of LC-QqQ-MS/MS and high accuracy of LC-HR-MS whereby false-positive signals of small molecules could be ruled out. Furthermore, an established QSRR model (Equation 3) for retention time prediction offers further confirmation of the identification. The developed platform enables users to exploit MRM archives with thousands of small molecules that have been previously analyzed. The proposed method offers relative qualification and quantification of the small molecules or their metabolites in complex samples. It suffices for many applications, including metabolomics, environmental contamination analysis and chemical removal in wastewater treatments. This platform has great

potential to be applied in the fast screening of multiple small molecules in environmental samples and human samples.

CHAPTER 4: THE DEVELOPMENT OF PSEUDO-MRM/PSEUDO-SIM DATABASE FOR TARGETED EXPOSOEM CHARACTERIZATION

4.1. Summary

Following Chapter 3, to avoid the dependence on chemical standards, we further scaled up the MRM method by developing a pseudo-single ion monitoring (GC-SIM) and pseudo-multiple reaction monitoring (LC-MRM) database with a powerful algorithm to optimize SIM and MRM transitions for exogenous and endogenous chemicals with available MS/MS fragmentation in public databases. To imitate the fragmentation patterns found in already-published experimental spectra, we chose MS/MS fragments (pseudo spectra) for the target compounds that caused the fewest fragmentation interferences from other molecules. We created a pseudo-SIM database and a pseudo-MRM database to boost sensitivity and specificity for targeted analysis by GC/LC-MS for over 70,295 unique compounds from the NIST EI library and over 10,353 unique compounds from MoNA LC-MS/MS library (https://github.com/YANGJJ93MS/Pseudo-SIM-MRM.git).

4.2. Introduction

Gas/Liquid chromatography coupled with mass spectrometry (GC/LC-MS) has been widely used for the analysis of various compounds due to its high sensitivity and availability. To date, there have been some methods for the detection of environmental pollutants in both non-targeted and targeted aspects assisted with GC/LC-MS. Interestingly, targeted methods based on triple quadrupole mass spectrometry (QQQ MS) have been well adopted in characterizing human exposomes (Mueller et al. 2005; Xue et al. 2021). Single ion monitoring (SIM) and multiple reaction monitoring (MRM) are currently the gold standards for quantitative analysis of volatile molecules and small molecules, which can be designed to provide accurate quantitation results of targeted environmental pollutants. Though targeted detection by SIM/MRM with similar coverage to the non-targeted method is a wish for the scientists in the field of the exposome, its development has been hindered by the analytical platforms to measure the trace level of thousands of chemicals, regardless of the high cost, low throughput, and availability of standards (Escher et al. 2020).

In this chapter, to avoid the unavailability of chemical standards, we scaled up the SIM/MRM method by developing a pseudo-GC-SIM and LC-MRM database with

a powerful algorithm to increase sensitivity and specificity for over 300,000 exogenous chemicals with available MS/MS2 fragmentation in the public database. We hypothesize that the wide coverage of environmental chemicals by selected MS spectra databases can provide comprehensive interferences in targeted analysis like matrix effect in real samples. And in silico development and optimization of SIM/MRM transition can be used to generate reliable pseudo-SIM/MRM transitions for sample analysis.

MS/MS fragmentation patterns and retention time or retention index were required to generate pseudo-SIM or pseudo-MRM transitions. The pseudo libraries were established in four steps: (1) MS1 and MS2 spectra data cleaning to remove unqualified data; (2) Retention time prediction for compounds for spectra screening; (3) Calculation of specificity for each MS2 spectra by spectra optimization algorithm and MS2 transitions ranking and validation using chemicals standards. Finally, we increased sensitivity and specificity for 306,869 unique compounds in NIST electron ionization (EI) spectra library and over 100,000 unique compounds in Massbank of North America (MoNA) LC-MS/MS (MS2) library (Fig. 4.1A-B).

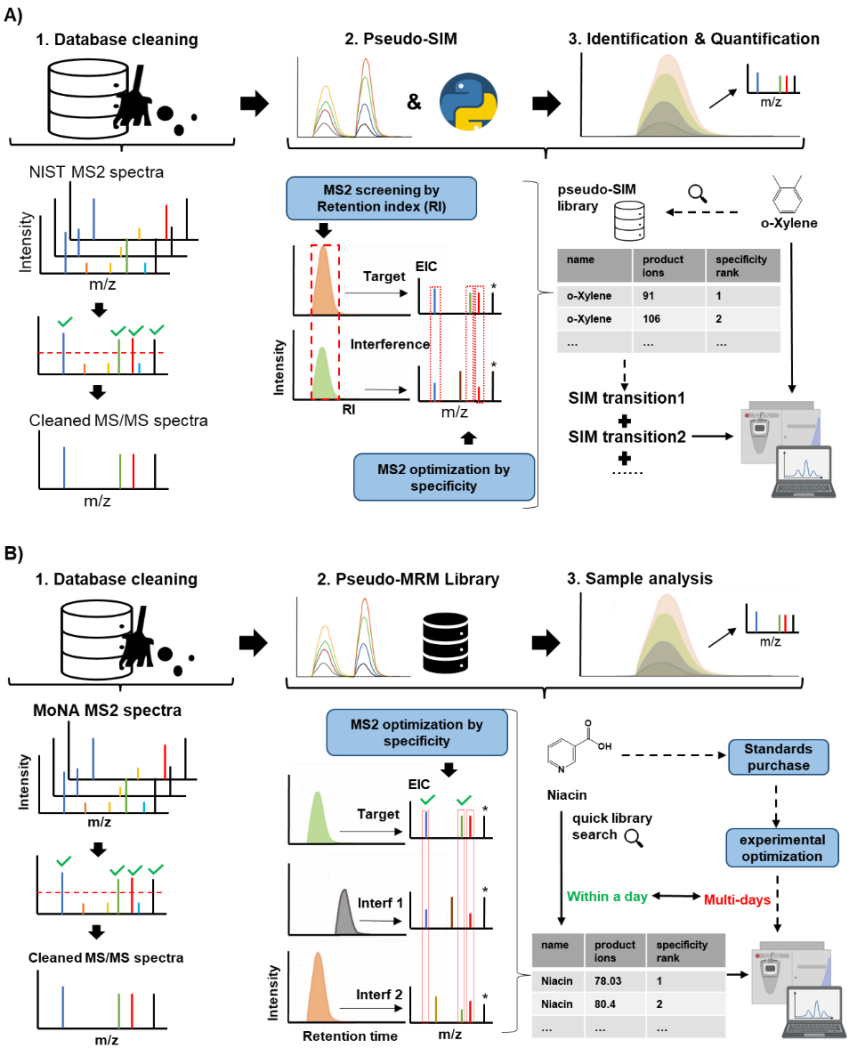


Figure 4.1. Overview of the pseudo-SIM/MRM database development. A) Workflow for developing pseudo-SIM database. B) Workflow for developing pseudo-MRM database.

4.3. Experimental section

4.3.1 Chemical reagents and sample treatments

FDA drug and all environmental pollutant standards were purchased from Sigma-Aldrich (Singapore). Metabolite standards and volatile compound standards were purchased from Supelco (USA). Detailed information was in SI. HPLC grade formic acid (FA), water (H_2O), ethyl acetate (EA), methanol (MeOH), and acetonitrile (ACN) were purchased from Thermo Fisher Scientific (Singapore). The stock solutions of all standards were dissolved in HPLC grade ACN or ultrapure H_2O at a concentration of 10 M. All stock solutions were stored at -80 ^oC. The volatile

compounds were diluted to 10 ppm in methanol and stored at -80 ºC for further analysis.

4.3.2 Mass spectrometry and LC/GC methods

Liquid chromatographic separation was achieved on an Acquity UPLC BEH C18 column (100 × 2.1 mm i.d., 1.7 μm; Waters, Milford, USA) at 30 °C. FA in water (0.1%, v/v, solvent A) and ACN (solvent B) were employed as the mobile phases for the analysis of derivatization products with a flow rate of 0.2 mL/min: 0−16 min 20% to 95% B, 16−19 min 95% B, 19−19.1 min 95% to 20% B, and 19.1−22.1 min 20% B. The auto-sampler was kept at 4 °C during LC-MS analysis. The injection volume was 10 μL and triplicate measurements were performed for every experiment. The confirmation of MRM transition signals was performed on Agilent 6490 Triple Quadrupole LC/MS, coupled with 1290 Infinity II 2D-LC System, equipped with an electrospray ionization (ESI) source operating in positive ionization mode in the ESI positive mode (Agilent, Singapore).

For gas chromatography separation, the injection volume is 1 uL. The samples were injected in splitless mode, with the injection port held at 250 °C. The initial oven temperature was held at 35 °C for 2 min, then ramped at 10 °C/min to 250 °C in 2 min, and further ramped at 20 °C/min to 320 °C and held for 28 min. Separation was performed on SLB-5 MS column (30 m × 25mm × 25mm). The post-run was 1 min to allow the oven cool down to 60 °C. The confirmation of SIM transition signals was performed on Agilent 5977B GC/MSD, equipped with electronic ionization source (EI) (Agilent, Singapore). MSD transfer line was held at 250 °C, ion source at 250 °C, and the mass analyzer at 150 °C. The GC-MS-SIM data were acquired in 35 min with 4 min solvent delay at a normal scan rate in the mass range 50-800 Da.

4.3.3 Database data conversion

Before the SIM/MRM optimization, two data files were needed for the data mining on the MS2 spectra data. NIST EI-MS library data were recorded repeatedly by task recorder. The entire GC-MS data were recorded in TXT files and parsed into python script for organization and data mining. LC-MS2 spectra in positive mode and negative mode were downloaded freely from MoNA mass spectrometry library (‘https://mona.fiehnlab.ucdavis.edu/downloads’). The LC-MS2 spectra were acquired in MSP format and were converted to TXT files by Editpad lite (Just Great Software,

Thailand). The GC-MS data files contain the chemical names, formula, database ID, and mass spectra information, including m/z and their intensity, retention index, and mass analyzer types. The LC-MS data files contain the chemical name, formula, exact mass, instrument type, InChiKey, and mass spectra data, including m/z and their intensity. The retention time was not included. The final pseudo-SIM and pseudo-MRM databases developed based on these two spectral databases were compiled into XLSX files for easy access.

4.3.4 Prediction for retention index in GC and retention time in LC

The implementation of the prediction model for the retention time index (GC) and retention time prediction followed the same procedures in session 3.3.3 in chapter 3. For GC-RI prediction model, the experimental retention time dataset of 3543 compounds with a wide spectrum across molecular weight and retention index was acquired from NIST EI-MS library. The training and testing splits were made in the ratio of 75:25. For retention time prediction in LC, an experimental retention time dataset of 1088 FDA-approved drugs was used for model development. The dataset for LCMS was acquired as described in 3.3.2. The entire dataset was divided into two subsets for model training and testing. The training and testing splits were made in the ratio of 75:25, respectively, with the caret package in R. 286 chemical descriptors were computed for each compound. For fast implementation and less demand for computation power, 2D-based chemical descriptors were applied in our package. Feature engineering was performed by eliminating none value, constant values, and highly correlated descriptors (Pearson $r^2 > 0.9$). The remaining 153 chemical descriptors were imported into to modeling function for hyperparameters optimization. The correlation R^2 values between observed and predicted retention times were used to indicate linear relationships and for global generalization of the prediction set. The top 18 important variables (descriptors) were selected by recursive searching and imported into the tree model as the final predictors (Fig S4.1). The hyperparameters for both models were tuned and optimized for the lowest mean absolute errors (MAEs) with 5-fold cross-validation. The fine-tuning of the model returned the tree number of 500 and the mtry value of 9.

4.4. Results

4.4.1 Overview of the pseudo-GC-SIM and LC-MRM database development

A typical workflow for developing pseudo-SIM and pseudo-MRM databases is elaborated in Fig. 4.1A-B. There are three steps involved in the development: (1) Mass spectra in TXT files were parsed into a python script. A spectra-cleaning step was performed by the python script and unqualified spectra were filtered. A compilation of cleaned MS/MS2 spectra was organized and stored in XLSX files. (2) A mass spectra screening based on the retention index and an MS spectra optimization algorithm was performed to produce SIM/MRM transitions for compounds. MS fragments from the precursor ions with similar m/z and retention index (GC) or retention time (LC) were extracted for comparison to highlight the product ions with the best selectivity, which makes SIM/MRM transitions spectra for target analytes easily differentiate from those of interference molecules in the sample matrix. (3) Finally, a targeted analysis of several volatile chemicals was performed to validate the pseudo-SIM transitions. Several common environmental pollutants were used in a targeted analysis to validate the pseudo-MRM transitions.

The pseudo-SIM/MRM spectra library can be applied in targeted analysis for compound identification and even qualification. It can also be applied to suspect chemical screening. For instance, O-xylene is a colorless liquid and easily escapes to the atmosphere from industrial sources through volatilization. Targeted analysis of O-xylene by GCMS can start with pulling prioritized SIM transitions from the pseudo-SIM library and using those transitions for monitoring with GC-MS/MS in SIM mode. For application in LCMS-based targeted analysis, taking niacin as an example, a query for MRM transitions from the pseudo-MRM library can return a list of the top 2 product ions for analysis within a day. Targeted analysis using the pseudo-MRM method can bypass experimental optimization of MRM transitions that requires multiple days.

4.4.2 MS spectra and chemical space covered by NIST EI-MS and MoNA library

NIST EI-MS spectra library contains experimental spectra obtained using the mass spectrometer coupled with gas chromatography and electron ionization (EI) interface. It covers a large chemical space of 306,869 unique compounds, including

but not limited to pesticide contaminants, drugs and their metabolites, surfactants, human metabolites, and industrial chemicals. 350,643 EI spectra and 447285 experimental retention time indexes are included in the library. However, only 112,253 compounds are recorded with both retention time and MS spectra. MassBank of North America (MoNA) is a metadata-centric, auto-curating repository designed for efficient storage and querying of mass spectra records. LC-MS/MS spectra can be downloaded from the library. 31,413 unique compounds with 145,349 MS2 spectra in positive and negative modes are available and can be downloaded from the MoNA website (Fig. 4.2A-B).

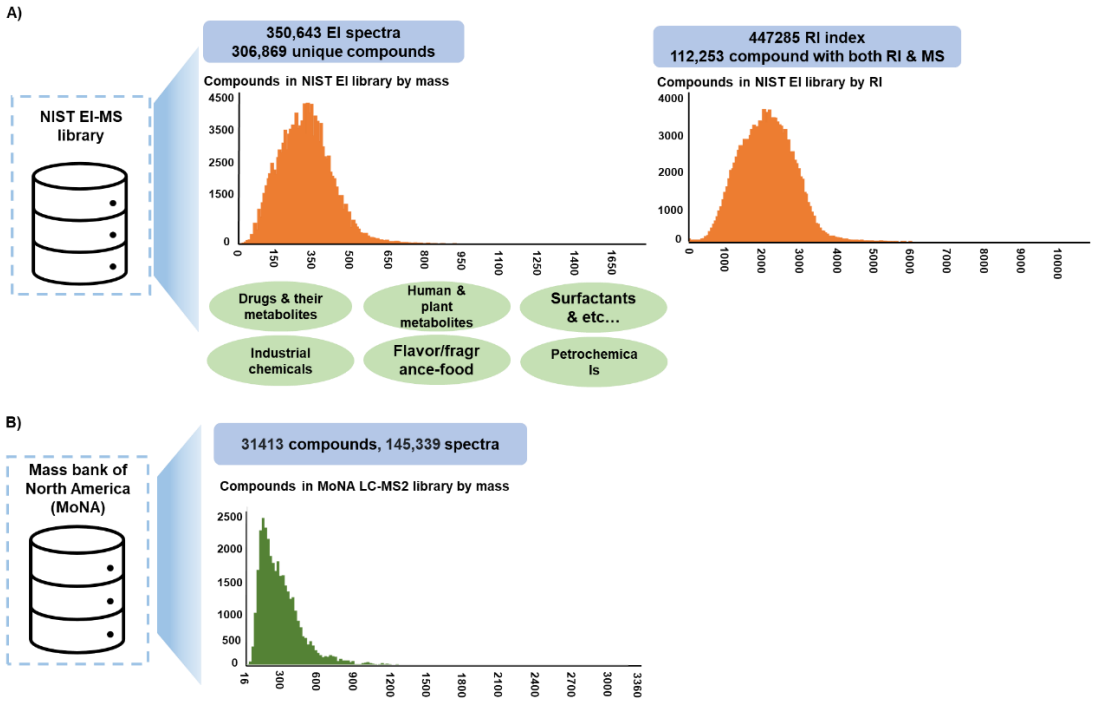


Figure 4.2. Chemical space and spectra repository of NIST EI-MS and MoNA database

Raw MS spectra data from original NIST EI-MS are massive and comprehensive yet challenging to utilize for deep data mining. Compound name, formula, molecular weight, exact mass, InChIKey, m/z of MS fragments and their relative intensity in column structure, and retention index are provided. The data format is not fully consistent, making it difficult for data extraction. Raw MS data from MoNA include compound name, InChIKey, formula, MW, exact mass, MS fragments, and their relative intensity in column structure, instrument type, precursor type, and precursor m/z value. Formats are not consistent, molecular weight, MW, precursor ion m/z, precursor ion type, and instrument type are not recorded for some compounds. MS2 fragments with no intensity are also recorded in the spectra data

columns, making it difficult for MS2 data extraction. Though tremendous data mining techniques are required to clean up raw mass spectra, the wide coverage for environmental pollutants offered by the NIST EI-MS library and MoNA LC-MS/MS library provides a foundation for *in silico* development of SIM/MRM transitions and their optimization.

4.4.3 Data cleaning for raw GC/LC-MS spectra

Different data cleaning processes were applied to raw MS spectra collected from NIST EI-MS library and MoNA library before MS spectra from databases can be used to generate pseudo-SIM/MRM transitions. There are five steps of data cleaning for GC-MS spectra collected from NIST EI-MS library. (1) mass spectra with available MS spectra and molecular weight were considered for further extraction. (2) The MS data for each compound was subjected to retention index prediction by a house random forest tree model if the experimental retention index was not available. (3) At this step, MS spectra list for each compound includes name, InChIKey, Formula, MW, MS2, abundance, estimated RI (EstRI), and predicted RI (predRI). (4) MS fragments with relative peak abundance low than 10% were filtered. (5) MS fragments with relative abundance higher than 10% were combined into final GC-MS spectra lists in CSV files (Fig. 4.3A). Finally, 350, 643 spectra from NIST EI-MS library were cleaned and refined into a clean MS spectra library of 70,295 spectra with 1048,575 SIM transitions.

Clean LC-MS/MS spectra from MoNA were generated by six steps: (1) MS2 spectra with precursor ion of $[M+H]^+$ were used to build a pseudo-MRM transitions library. (2) MS2 spectra must have precursor m/z or MW value in it. (3) MS2 list after screening includes name, InChIKey, Exact Mass, precursor ion, precursor ion m/z, MW, MS2 fragments, and their abundance. (4) The MS2 fragments were filtered. MS2 fragments with a relative abundance lower than 10% were not considered. (5) All filter MS2 spectra were combined into an MS2 spectra list. (6) Retention time prediction for each compound was conducted. The final clean LC-MS/MS spectra were curated in CSV files (Fig. 4.3B). Finally, 145,339 spectra from the MoNA LC-MS/MS library were cleaned and refined into a clean MS spectra library of 10,353 spectra with 180,001 SIM transitions.

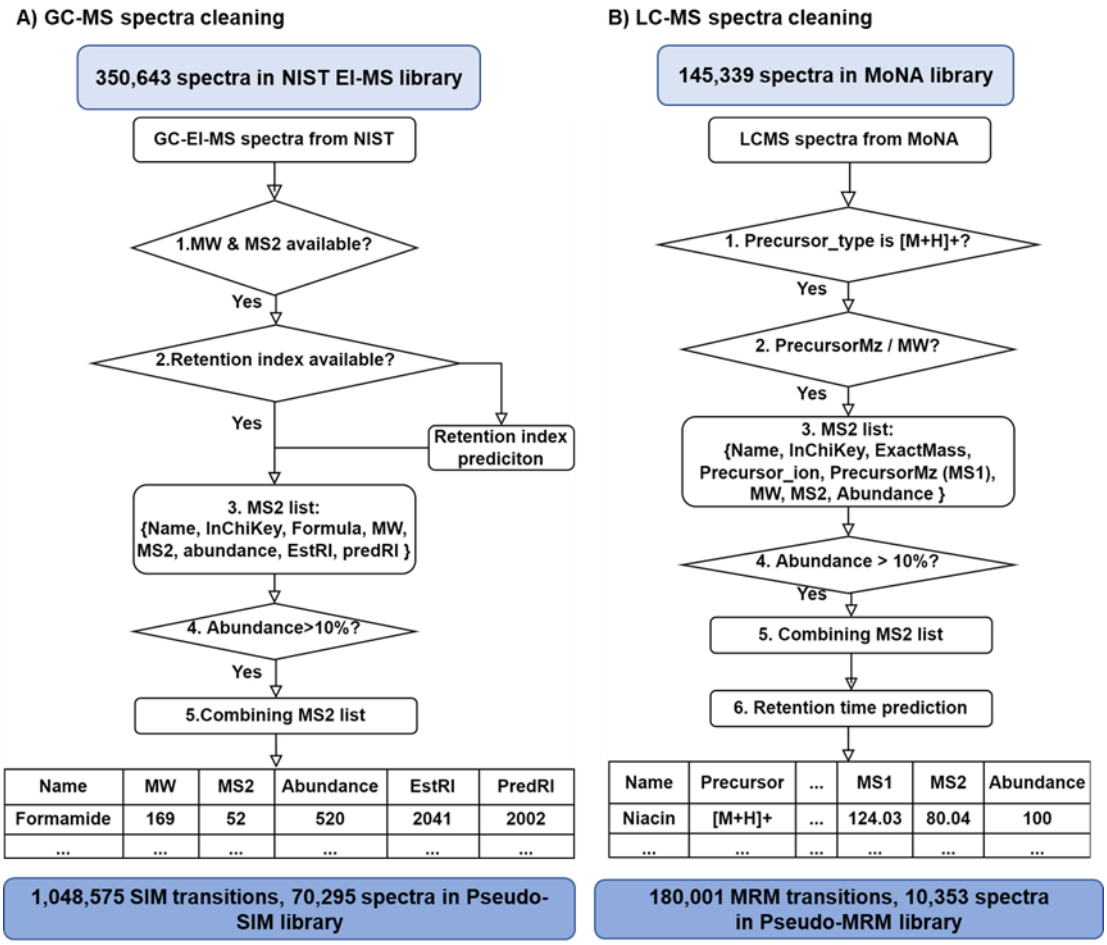


Figure 4.3. Data cleaning for GC/MS and LC/MS spectra from NIST EI-MS database and MoNA LC-MS/MS database.

4.4.4 Retention time index and retention time prediction

Retention time prediction by random forest regression models based on QSRR is widely accepted for non-targeted analysis (Bonini et al. 2020; Cao et al. 2015b; Zdravković et al. 2018). In our previous study, we achieved a comparable prediction accuracy by QSRR-based regression model with features selection by random forest algorithm (Yang et al. 2020). In this study, we apply a random forest algorithm for direct prediction of the retention index for GC and retention time for LC. Feature selection and hyperparameter optimization can be operated automatically without manual monitoring. The high efficiency of tree models allows fast development of prediction models from raw datasets.

For both RI and RT models, 286 molecular descriptors were computed for each compound. Specifically, for RI prediction, 153 chemical descriptors remained and were imported into to modeling function for hyperparameters optimization. The

top 18 important variables (descriptors) were selected by recursive searching and imported into the tree model as the final predictors (Fig S4.1A). The fine-tuning of the model returned the tree number of 500 and the mtry value of 9. Overall, the prediction model returned a mean absolute error (MAE) of 67.63, with a strong correlation (R=0.97, *p<2.2e-16*) between predicted RI and experimental RI. For retention time prediction in LC, 127 molecular descriptors remained after feature selections. The fine-tuning of the model returned the tree number of 500 and the mtry value of 9. The top 15 important variables (descriptors) were selected by recursive searching and imported into the tree model as the final predictors (Fig. S4.1B). The final prediction model returned a mean absolute error (MAE) at 1.31 min, with a strong correlation (R=0.91, *p<2.2e-16*) between predicted RT and experimental RT.

With tolerance prediction error at 67 RI units and 1.31 min for RT, we were allowed to set the retention index threshold for screening at 100. Candidates in the database with a deviation of RI from target compounds with less than 100 were considered interfering molecules. The retention time threshold for screening was set at 1.5 min. Candidates from the LC-MS/MS library with RT deviation from target compounds with less than 1.5 min were considered interfering molecules.

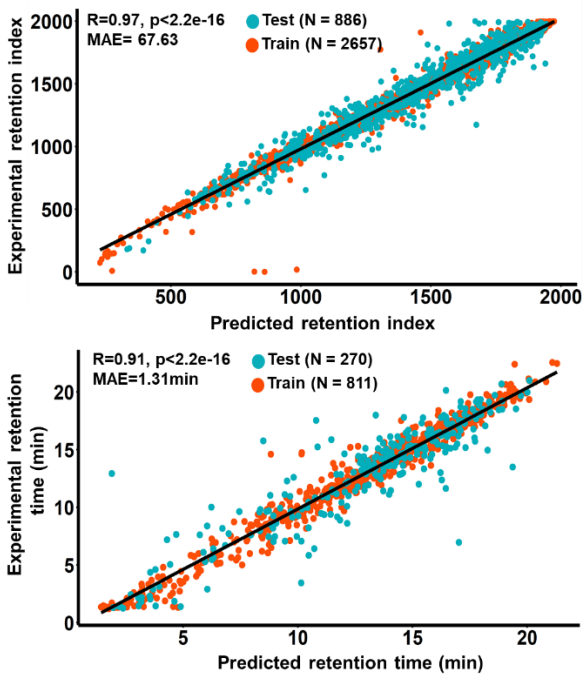


Figure 4.4. Performance of prediction model for RI and RT. A) Prediction model for RI reached MAE = 67.63 with Pearson correlation R=0.97. B) Prediction model for RT reached MAE = 1.31min with Pearson correlation R=0.91

4.4.5. MS spectra optimization for building pseudo-SIM transitions

Determination of the SIM ions is of utmost importance for targeted analysis using GC-MS-SIM. In the experimental optimization process, users need a scan-mode spectrum of the compounds to examine for candidates SIM ions following two general rules: (1) the ions are unique to the compound and not common in a wide of compound spectra; and (2) the ions are abundant in response. Ions of low abundance may not be found when compounds are at low concentrations. To meet these two general rules, we developed an optimization algorithm to determine the SIM ions from EI-MS spectra in NIST library. The pseudo-SIM library was built by extracting compound spectra one by one and comparing one compound spectrum with all other similar spectra using the optimization algorithm. All compounds' spectra with their ratings were compiled into a new database, the pseudo-SIM database.

Specifically, the algorithm started with the following process: (1) a molecule spectrum was extracted (for instance, precursor ion (MS1) m/z=200, retention index RI = 1000) as the target spectrum. Compounds with similar RI (|RI-Target RI|<=0) and MS1 (|MS1-targetMS1|<=1) were considered to be interfering molecules (interfering molecules 1,2,3) and their spectra were collected from clean spectra data (Fig. 4.5A, Fig. S4.2); (2) The occurrences of each fragments ion in interfering molecule spectra were counted and classified into four categories: true positive (present in target spectra), true negative (not present in interfering spectra), false positive (present in interfering spectra) and false negative (not present in target spectra). Three parameters were calculated for each ion based on the four types of occurrences: specificity, sensitivity, and accuracy, as the formula shown in Fig. 4.5B; and (3) The algorithm ranked all fragment ions according to the order of specificity and output the results in xlsx files. Fragment ions with the highest selectivity against interfering spectra (highest specificity value) were selected as pseudo-SIM ions.

For demonstration, the development of pseudo-SIM transitions and its application for targeted analysis of ethylbenzene (MW/MS1=106, RI=893) was presented in Fig. 4.5D. The EI-MS spectra of ethylbenzene included fragment ions 51, 65, 91, and precursor ion 106. Library search by our algorithm indicates 1.3% of total spectra in clean NIST EI-MS library shared the same retention index. 2.1% of those spectra share the same precursor ion m/z of 106. A graphical showcase of the calculation of ions' specificity between ethylbenzene spectra and two interfering

compounds spectra was described in Fig. 4.5D. Ions with m/z 65 and 51 were found to be the top 2 ions with less interference from other coeluting compound spectra. Finally, three ions were selected as pseudo-SIM transitions, with 65 and 51 as the top 2 SIM ions in the final targeted analysis. The experimental chromatogram by pseudo-SIM transitions showed that all transitions were observed in the mixture sample. Targeted analysis using pseudo-SIM transitions for spiked ethylbenzene in a mixture sample was successful.

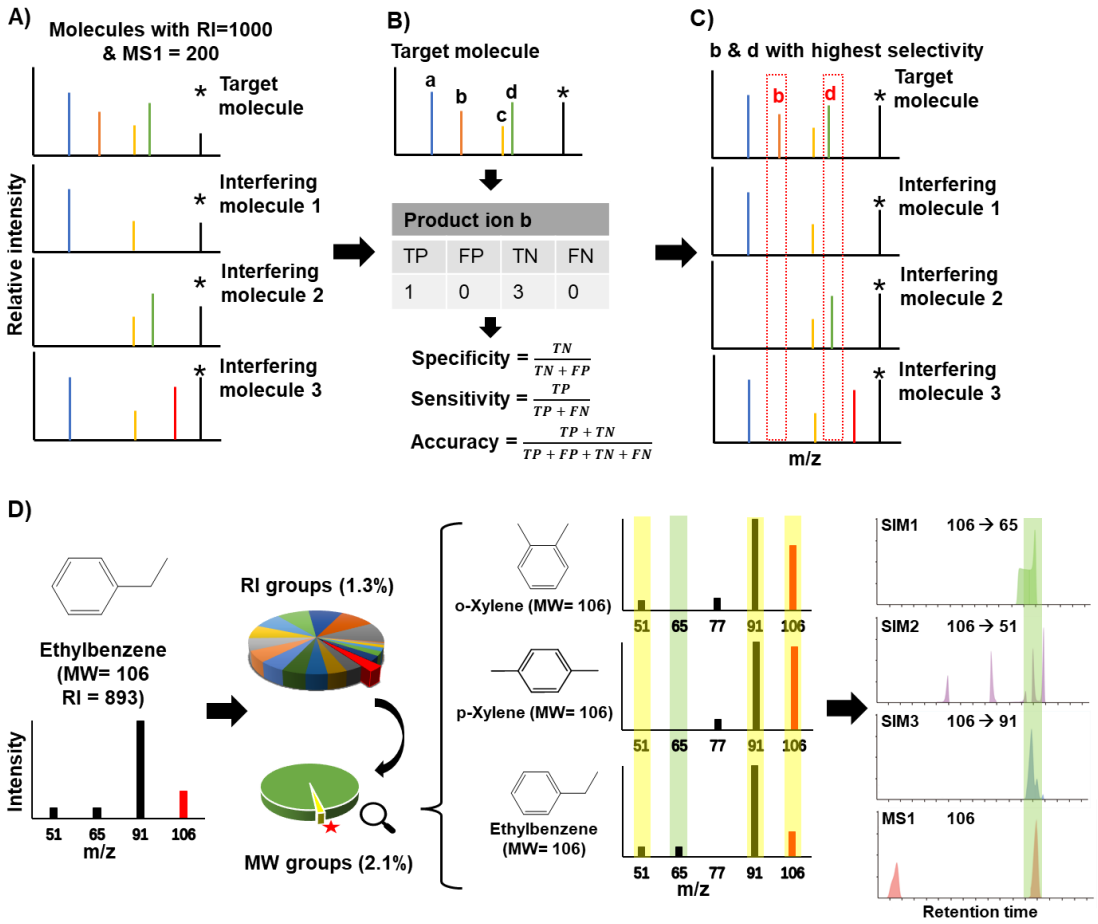


Figure 4.5. A-C) Algorithms to search for unique SIM transitions for target compounds. D) SIM transitions development of O-xylene.

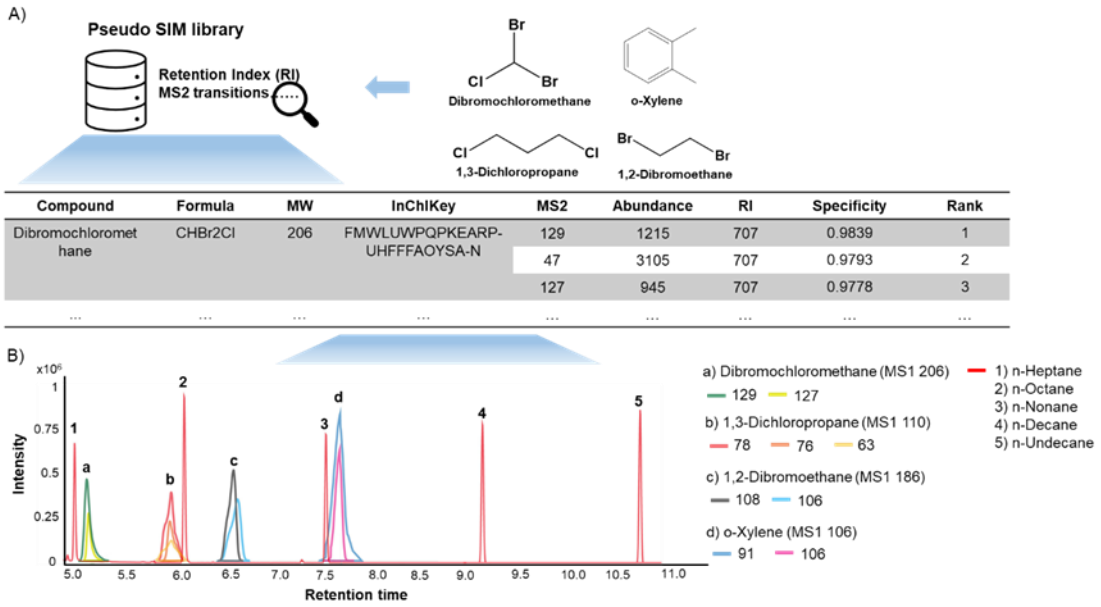
4.4.6 MS spectra optimization for building pseudo-MRM transitions

The optimization algorithm was modified for developing a pseudo-MRM transition library due to the different data structures of LC-MS/MS spectra from LC-MS/MS spectra. LC-MS/MS spectra from MoNA were high-resolution mass spectra. The threshold for differentiating interfering spectra from other spectra was lower than EI-MS spectra. Compound spectra from the library were defined as interfering spectra

for target spectra with closer retention time (|RT-Target RT|=1.5 min) and closer m/z value (|MS1-targetMS1|<=0.7). There are three steps involved in the pseudo-MRM spectra optimization: (1) a molecule MS2 spectrum was extracted (for instance, precursor ion (MS1) m/z=124.03, RT =9.0min) as the target spectrum. Compounds with similar RT (|RT-Target RT|<=1.5min) and MS1 (|MS1-targetMS1|<=0.7) were interfering molecules (interfering molecules 1,2,3) and their spectra were collected from clean spectra data (Fig. S4.3A, Fig. S4.4). (2) The occurrences of each fragments ion in interfering molecule spectra were counted and classified into four categories: true positive (present in target spectra), true negative (not present in interfering spectra), false positive (present in interfering spectra) and false negative (not present in target spectra). Three parameters were calculated for each ion based on the four types of occurrences: specificity, sensitivity, and accuracy, as the formula shown in Fig. S4.4B. (3) the algorithm ranked all fragment ions according to the order of specificity and output the results in xlsx files. Fragment ions with the highest selectivity against interfering spectra (highest specificity value) were selected as pseudo-SIM ions (Fig. S4.3C, Fig. S4.4).

4.4.7 Validation of pseudo-SIM transitions with VOCs

We performed a pseudo-SIM-based targeted analysis for four volatile organic carbons for validation in the indoor environment. Exposure to VOCs in a long term might cause detrimental health effects, especially to the respiratory system(S. Wang et al. 2007). Monitoring of VOCs in indoor air is important and largely performed using GC-MS. Analyzing VOCs in environmental samples with GC-MS in SIM mode requires chemical standards and SIM method optimization. Using a pseudo-SIM database for searching optimized SIM transitions might enable a high-throughput qualification and quantification of VOCs. Here we applied SIM transitions from our pseudo-SIM database to identify 4 VOCs (Dibromocholoromethane, 1,3-Dichloropropane, 1,2-Dibromoethane, and O-xylene) in the mixture sample of 76 VOCs.



Compound	Formula	MW	InChIKey	MS2	Abundance	RI	Specificity	Rank
Dibromochloromethane	CHBr2Cl	206	FMWLUWPQPKEARP-UHFFFAOYSA-N	129	1215	707	0.9839	1
				47	3105	707	0.9793	2
				127	945	707	0.9778	3
...	...	...	...	...	...	...	...	...

Figure 4.6. EI-MS spectra pseudo-SIM transition. A) SIM inquiry by compound ID in the pseudo-SIM library. B) GC-MS-SIM chromatograms of 4 VOCs for validation.

The InChIKey list of 4 targeted VOCs was imported to the pseudo-SIM library. Multiple SIM transitions were retrieved from the database with the calculated specificity value of each SIM transition and their ranking regarding the specificity (Fig. 4.6). Top 3 pseudo-SIM transitions of each compound were selected for MS detection. Experimental or predicted RI were provided for the pseudo-SIM transitions. The retention index for each transition was then converted to retention time and imported to equipment parameters to ensure enough time for MS acquisition of each SIM ion. Finally, dibromochloromethane, 1,2-dibromoethane, and O-xylene were identified by two pseudo-SIM transitions. 1,3-dichloropropane was identified by three pseudo-SIM transitions. Compound identification was confirmed by the similar retention time and signal abundance of pseudo-SIM transitions for each compound.

4.5 Conclusion

We generated a pseudo-SIM database and a pseudo-MRM database that increased sensitivity and specificity for targeted analysis by GC/LC-MS for over 70,295 unique compounds with 1,048,575 pseudo-SIM transitions from NIST EI library and over 10,353 unique compounds with 180,001 pseudo-MRM transitions from MoNA LC-MS/MS library in positive and negative mode. With these two databases, we can perform targeted analysis but require no chemical standards for MS method development and optimization, which speeds up the targeted analysis

procedure. Pseudo-SIM transitions of four VOCs were validated by the identification of the selected compounds in a mixture sample. To further increase the chemical size of our pseudo-SIM and pseudo-MRM library, additional MS spectra from other public EI-MS or LC-MS/MS libraries will be incorporated.

CHAPTER 5: USING CHEMICAL ISOTOPE-LABELING ON THE TOP OF PSEUDO-MRM METHOD FOR CHARACTERIZING CHEMICALS AT TRACE LEVELS

5.1 Summary

In Chapter 5, we further tackle the unavailability of standards, low detection rate, and laborious quantification processes in targeted analysis. we have developed a novel sensitive and high-throughput exposome analytical platform (CIL-ExPMRM) by isotope labeling urinary biomarkers to increase the detection of chemicals at trace levels. We built up a CIL-pseudo-MRM exposome database of environmental pollutants and their transformation products for 110,000 compounds. The platform has been well incorporated with automatic MRM generation, dynamic MRM optimization, and data analysis. Finally, almost all the above-mentioned processes were integrated by a computational pipeline and eventually built into a user-friendly website. The performance of this platform has been validated with highly successful and low false positive rates across several instrumental platforms. Using this one-stop platform, we can complete the suspected screening analyses of tens to hundreds of exposomic chemicals within a short time.

5.2 Introduction

The detection of environmental pollutants in biological samples is vital for the study of the exposome. However, one challenge has been the ever-evolving and growing list of environmental pollutants along with the development of human society (Manzetti et al. 2014). Exposome characterization has been challenged by the vastly large number of chemicals and low concentrations. Up to 2021, there are more than 900,000 chemicals have been registered as shown in the Toxcast database (Williams et al. 2021). There are many chemical classes of environmental pollutants, such as per-/polyfluoroalkyl substances (PFAS), pesticides/herbicides, polychlorinated biphenyls (PCBs), and pharmaceuticals/personal care products (PPCPs), most of which concentrations are too low for being detected (Dodds et al. 2021). To cope with the increasing number and volume of environmental pollutants, a new analytical platform must also be developed to assess these environmental pollutants and their transformation products.

The MRM transitions of derivatized environmental chemicals could likely be directly generated from the parent ions to product ions. Therefore, we speculated that CIL-LC-MS combined with a pseudo-MRM (CIL-pseudo-MRM) strategy can be used to develop one exposome platform. In the study, we aimed to develop a novel sensitive and high-throughput exposomic analytical platform (CIL-ExPMRM) by chemical isotope labeling assisted LC-pseudo-MRM-MS. Specifically, methylphenylethylamine (MPEA) and DnsCl were used to derivatize carboxyl and hydroxyl compounds, respectively. A pseudo-MRM database of > 110,000 environmental pollutants and their transformation products were built up. A simple dynamic MRM optimization algorithm was generated by using collision energy, dwell time optimization, and retention time prediction models. Automatic MRM alignment and data statistical analysis were achieved by comparing peak intensity, which shows an excellent correlation with peak area after derivatization. Finally, all the above-mentioned processes were integrated by a computational pipeline https://github.com/YANGJJ93MS/CILMRM.git.

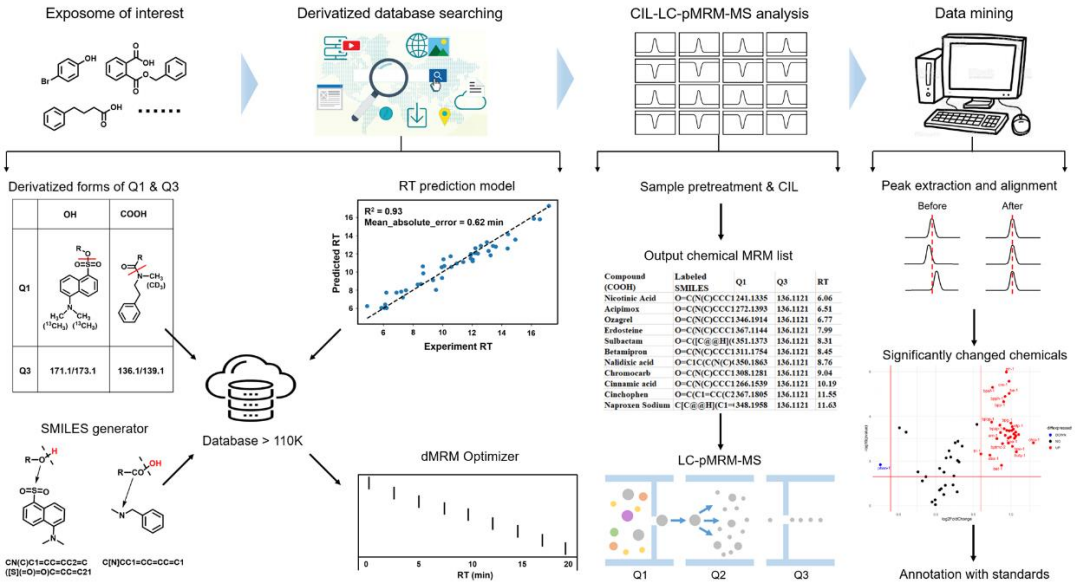


Figure. 5.1. Schematic framework of the user defined CIL-pseudo-MRM exposome platform.

5.3 Experimental Section

5.3.1 Chemical Reagents and Stock Preparation

FDA drug, an environmental pollutant, metabolite standards were purchased from Sigma-Aldrich (Singapore). The derivatization reagents DnsCl, MPEA,

triphenylphosphine (TPP), 2,2'-dithiodipyridine (DPDS), dimethylaminopyridine (DMAP), trimethylamine (TEA) were purchased from Sigma-Aldrich (Singapore) with a purity > 95%. The isotope reagent $^{13}C_2$-danysl chloride ($^{13}C_2$-DnsCl) was purchased from TMIC (Edmonton, Canada) and d_3-MPEA was synthesized in our lab according to the previous work (J. Zheng et al. 2021). HPLC grade formic acid (FA), water (H_2O), ethyl acetate (EA), and acetonitrile (ACN) were purchased from Thermo Fisher Scientific (Singapore). The stock solutions of all standards were dissolved in HPLC grade ACN or ultrapure H_2O at a stock concentration of 10 mM. DnsCl/$^{13}C_2$-DnsCl and DMAP were at the concentrations of 100 mM in ACN; MPEA/d_3-MPEA, TPP, DPDS, and TEA were prepared at the concentrations of 20 mM in ACN. All stock solutions were stored at -80 ^{o}C.

5.3.2 Sample Pretreatment

A pooled human urine sample was prepared by mixing the first morning voids of 10 volunteers (healthy, ages 24−42) which were immediately frozen at −80 °C after sampling. Informed written consent was obtained from all participants. Institutional Review Board approval for human specimen analysis was obtained from Singapore (IRB-2017-02-023). Five hundred μL pooled urine was used for sample pretreatment. First, 6 M HCl was added to urine samples to adjust pH to 1.0. Then, EA (400 μL × 2) was continuously added and vortexed for 30 s. Next, the organic layer was collected, combined, and dried at -80 °C (Peng and Li 2013). The obtained residues were redissolved with 100 μL ACN for subsequent chemical derivatization.

5.3.3 Derivatization steps

For the derivatization reaction between DnsCl and hydroxyl compounds, 100 μL ACN containing hydroxyl compounds were mixed with DMAP (10 μL), TEA (10 μL), and DnsCl/$^{13}C_2$-DnsCl (10 μL). After vortexing and spinning the sample down, the mixture was incubated at 60 °C for 60 min. Then, the reaction was subsequently quenched by NaOH and FA solutions. Detailed optimization details could be found in previously published methods (Jia et al. 2019). For the derivatization of MPEA and carboxyl compounds, TPP (10 μL) and DPDS (10 μL) as the catalysts were added into ACN solution containing carboxyl compounds. Then, MPEA/d_3-MPEA (10 μL) was added into the above reaction system. The mixture was vibrated under 40°C for 30

min followed by drying under N_2. The obtained residues were redissloved with 40%ACN for LC-MS analysis. Optimization details could be found in the published method (J. Zheng et al. 2021).

5.3.4 Instrumentation and Analytical Conditions

Liquid chromatographic separation was achieved on an Acquity UPLC BEH C18 column (100 × 2.1 mm i.d., 1.7 μm; Waters, Milford, USA) at 40 °C. FA in water (0.1%, v/v, solvent A) and ACN (solvent B) were employed as the mobile phases for the analysis of derivatization products with a flow rate of 0.4 mL/min. A post-gradient with the following proportions (v/v) of solvent B for the detection of DnsCl-OH derivatization products was applied: 0−16 min 20% to 95% B, 16−19 min 95% B, 19−19.1 min 95% to 20% B, and 19.1−22.1 min 20% B. For MPEA-COOH derivatization products, the corresponding post-gradient was 0−3 min 10% B, 3−16 min 10% to 95% B, 16−19 min 95%B, 19−19.1 min 95% to 10% B, and 19.1−22.1 min 10% B. The auto-sampler was kept at 4 °C during LC-MS analysis. The injection volume was 10 μL and triplicate measurements were performed for every experiment.

The confirmation of all derivatization products was performed under full scan on Waters ACQUITY UPLC system-Xevo G2-XS hybrid quadrupole-time-of-flight (QTOF) mass spectrometer (Waters, Milford, MA, USA), equipped with an electrospray ionization (ESI) source operating in positive ionization mode. All underivatized standards were analyzed by full scan by QTOF in the ESI negative mode. To explore the applicability of the MRM method, we have tested the performance on three types of instruments, Waters ACQUITY UPLC system-Waters Xevo™ Triple quadrupole (TQ) mass spectrometer (Waters, Milford, MA, USA), Agilent UPLC 1290 system-Agilent iFunnel 6495 QqQ (Waldbronn, Germany) and Agilent UPLC 1290 system-AB QTRAP 6500 (USA) in the ESI positive mode. Then, MRM transitions were developed by screening precursor and product ions simultaneously. Detailed parameter settings were as follows. Waters Xevo G2-XS hybrid quadrupole−time-of-flight (QTOF) mass spectrometer: The following instrument parameter settings were used: sensitivity mode, a mass range of *m/z* 50−1200, the capillary voltage of 3.0 kV for the positive ion mode and −1.5 kV for the negative ion mode, sampling cone of 80 V, source offset of 60 V, source temperature of 120 °C, drying temperature of 500 °C, cone gas flow of 100 L/h, and desolvation gas flow of 1000 L/h. Leucine enkephalin was used as the lock mass for

all MS experiments. MS/MS experiments were performed with the collision energy ramped from 10 to 35 eV.

Waters Triple quadrupole (TQ) mass spectrometer: PLC was performed using a Waters ACQUITY UPLC system (Waters, Milford, MA, USA). Mass spectrometry detection was performed by using a Xevo™ Triple Quadrupole MS (Waters Corp., Milford, MA) equipped with an electrospray ionization (ESI) source operating in positive ionization mode. The desolvation gas flow rate was set to 1000 L/h at a temperature of 550 ℃, the cone gas flow rate was set at 50 L/h and the source temperature was set at 150 ℃. The capillary voltage was set to 3000 V. Agilent 6495C triple quadrupole LC/MS: Chromatography was conducted using a 1290 Infinity II LC system, coupled via Jet Stream interface (ESI) to a 6495C triple-quadrupole mass spectrometer. Electrospray parameters were as follows: gas flow 11 L/min at 200 ℃; nebulizer 15 psi, sheath gas flow 12 L/min at 400 ℃; capillary voltage +3500 V. AB SCIEX QTRAP 6500: a 1200 HPLC (Agilent Technologies). Use the following settings for the QTRAP mass spectrometer: +4,500 V positive; 475 °C; curtain N2 gas set to 40; medium collision energy; ion source gas 1 and 2 set to 55; declustering potential 40; entrance potential 10; collision cell exit potential 10.

5.3.5 Data analysis

For the data acquired by QTOF, chromatographic peaks were manually extracted by using Waters MassLynx V4.2 with mass tolerance $\leqslant$ 20 ppm. Data acquired by TQ was automatically analyzed by built-in software, Waters MassLynx V4.2, AB QTRAP Analyst, or Agilent MassHunter for chromatographic integration and intensity extraction. The compounds, of which light- and heavy-derivatized MRM transitions were all detected with the difference value of retention time < 0.1 min and intensity ratios within 0.5 − 2.0, were defined as potential hydroxyl or carboxyl compounds.

5.3.6 Retention time prediction model development

In our proposed CILMRM workflow, we adopt the implementation of the retention time prediction in session 3.3.3 in chapter 3. We used the experimental retention time of 272 compounds for model development (176 carboxyl compounds and 76 hydroxyl compounds. Chemical descriptors were calculated by the R platform-

based Chemistry Development Kit package(rCDK). We parsed the derivatized SMILES to the prediction function for calculation. The hyperparameters of the prediction model follow the procedures in session 4.3.4 in chapter 4.

5.3.7 Dynamic MRM transition grouping

In monitoring MRM transitions, the dwell time of each transition is an important instrumentation parameter. To gain better peak shape and signal sensitivity, the number of MRM transitions per retention time segment needs optimization to achieve a longer dwell time for each MRM transition. The peak detection algorithm in our workflow required CIL-MRM transitions with different MS1 values per LC run. Therefore, we developed a simple MRM transition grouping function (transitiongroup()) in CILMRM as shown in Fig. S5.1. The four-step procedure includes (1) compound grouping by function groups; (2) compound filtration by MS1 comparison; (3) retention time screening and grouping, and (4) adding light/heavy CIL-MRM transitions parameters for direct implementation to the mass spectrometer. In our demonstration, the transitiongroup() function returned groups with a unique transition. In each group, light/heavy CIL MRM transition pairs had unique MS1 with no overlapping. We set the maximum number of transitions to 50 per minute and we had 25 compounds per minute in the final lists.

5.3.8 CIL-MRM: One Stop Platform of *in silico* derivatization and MS data analysis

CILMRM is an open-source R package for analysing multi-group LC-MRM-MS data with three replicates in each group. It is designed to simplify the analyses procedure by implementing a standardized workflow, which renders reproducible and visualized results in the R environment. By running simple functions in CILMRM, complex results are output in a well-constructed file format, in either tabular or graphical formats. CILMRM workflow is an integration of multiple powerful R packages for LC-MS data analysis, including MS-based packages (XCMS, Msnbase), statistics-based package (tidyverse) and visualization packages (ggplot2, ggpubr, ggrepel). It renders an automatic analysis of MS data with less requirement for handling program languages. CILMRM is free to download on GitHub (“https://github.com/YANGJJ93MS/CILMRM.git”).

CILMRM workflow starts with MS raw data from derivatized samples and corresponding experimental settings. Experimental settings include compound precursor ions mass charge ratio (m/z), product ions m/z, predicted retention time (RT), compound identifier (ID, CAS), compound labeling reaction identifier (ID2), and treatment identifier (control, treatment concentration). Experimental settings, including RT and precursor ions m/z, require compound structures and properties after the labelling reaction. Therefore, we proposed a virtual derivatization python scripted to generate compound structure information, a retention time prediction inside CILMRM for retention time prediction based on derivatized molecular structure, and MRM transitions grouping function in CILMRM, to generate experimental settings in ready-to-use formats.

CILMRM R package workflow is designed to conduct a multi-group CIL-MRM MS data analysis in only 3 steps (Fig. S5.1). We implement the workflow by each step as an R function as the key design concept. This three-step procedure includes (1) environment setup and peaks detection from raw data, (2) compound identification by isotopic peaks matching, and (3) compound-level differential analyses. Before running any steps, users need to convert MS/MS raw data to mzML format using ProteoWizard for data input. The conversion settings of raw MRM data on MSconvert by ProteoWizard vary regarding the mass spectrometer vendors. In our application, we selected the settings based on Agilent raw data. Global environments and R packages were then installed with a simple command line in R for implementation. With raw data input and customized parameters, peaksdetect() function generates a table of signal peaks from each MRM transition. As each compound is targeted by light and heavy isotopic labeling reactions, the compound is identified by peak alignment and matching between light/heavy isotopic signals by peakfine() function. Finally, diffexp() function proceeds with a correspondence analysis across groups of samples and visualizes the results in a tabular format and a graphical format.

5.4 Results and Discussion

5.4.1 User defined dynamic CIL-pesudo-MRM exposome platform

Environmental pollution has a detrimental impact on human health and has been quickly increasing in both number and volume worldwide; this highlights the

need for highly efficient and simultaneous detection analytical methods. Unmet analytical challenges exist for exposome research; As a consequence of the number of environmental chemicals and metabolites, chemical diversity, low abundance, and lack of readily available authentic standards, there is a critical need to overcome these limitations. CIL-pseudo-MRM has many advantages, such as high sensitivity, high coverage, improved peak shape, and predicted MRM transitions. Therefore, in the project, we employed CIL-pseudo-MRM strategy to establish a user-defined exposome platform to gain information on known and unknown exposure biomarkers for human exposome research. A CIL-pMRM exposome database (> 110 k environmental pollutants and their transformation products) was built up to achieve high coverage analysis of environmental pollutants. We also developed a computational pipeline to achieve automatic data processing and constant replenishment of the exposome database. Overall, the user-defined dynamic CIL-ExPMRM platform met the increasing detection needs of the environmental pollutant list to some extent. To the best knowledge, this is the first applicable screening platform at the exposome scale.

5.4.2 Establishment of our CIL exposome database

The chemical database we adopted in this study is the fusion of several human exposome databases, which was completed in our recent study (F. Zhao et al. 2021b). The database is comprised of chemicals with great environmental concern, large production, high human exposure, or toxicity. Though many environmental chemicals themselves cannot be derivatized, it is worth noting that a major portion of environmental pollutants contains hydroxyl or carboxyl groups, and their bio-transformation products were usually polarized with reactive function groups, where usually their urinary exposure biomarkers (Jia et al. 2019). The CIL strategy could produce predicted product ions and greatly improve the sensitivity of hydroxyl or carboxyl compounds. Therefore, we aimed to build up a CIL-pMRM exposome database containing hydroxyl or carboxyl compounds by using the CIL strategy. The derivatization reagents, DnsCl/$^{13}C_2$-DnsCl, have been reported to possess high derivatization efficiency to aromatic and aliphatic hydroxyl compounds in biological samples, while MPEA/d_3-MPEA were designed for aromatic and aliphatic acids (Fig. 5.2A). In this study, we have developed a CIL exposome database of > 110 k

environmental pollutants and their transformation products with hydroxyl or carboxyl groups by CIL strategy. To build up such a database, we transferred the components of compounds in traditional database to those of CIL-pMRM database, including precursor *m/z*, product ions, CE values and RT windows.

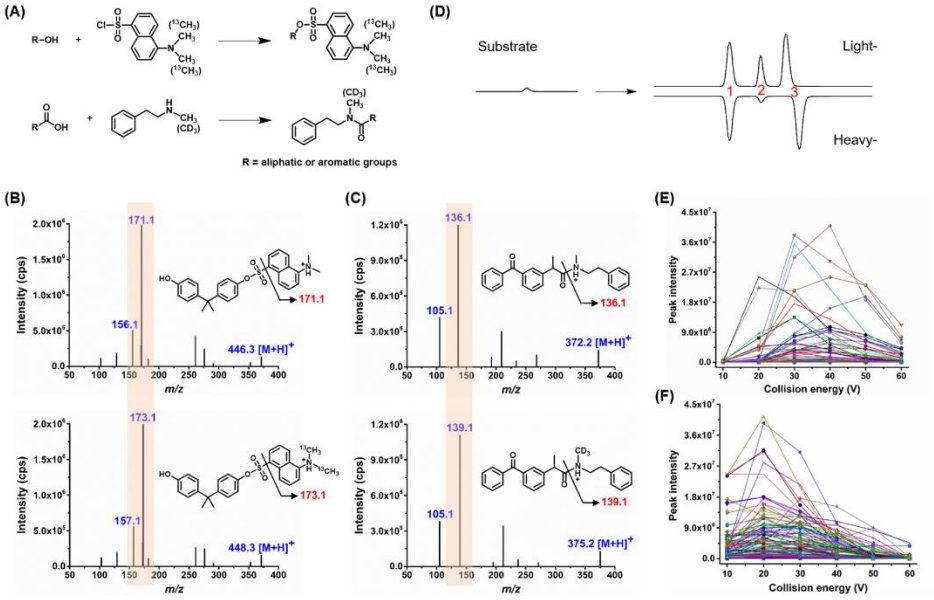


Figure. 5.2. (A) CIL strategy of DnsCl/$^{13}C_2$-DnsCl and MPEA/d_3-MPEA. (B) MS2 spectra of DnsCl/$^{13}C_2$-DnsCl-derivatized bisphenol A. (C) MS2 spectra of MPEA/d_3-MPEA-derivatized ketoprofen. (D) Chromatograms of ketoprofen before and after MPEA/d_3-MPEA derivatization. (E) Optimization of CE value of DnsCl/$^{13}C_2$-DnsCl-derivatized products. (F) Optimization of CE value of MPEA/d_3-MPEA-derivatized products.

5.4.3 Fragmentation behaviours of derivatization products

First, we investigated the feasibility of the derivatization reactions in the field of the exposome. Some standards of FDA drugs, environmental pollutants, and metabolites with carboxyl or hydroxyl groups were used for the performance evaluation. The results showed that upon DnsCl or MPEA derivatizing, all compounds could form stable cation quasi-molecular ions ([M+H$^+$]) with high abundance, indicating that DnsCl and MPEA could be successfully applied in the exposome. Since DnsCl and MPEA themselves possess high mass response, it could enhance the ionization efficiencies of hydroxyl and carboxyl compounds during MS analysis after derivatization. In this respect, here we evaluated the enhancement of the detection sensitivities of hydroxyl and carboxyl compounds by comparing the peak

areas of hydroxyl and carboxyl compounds before and after DnsCl or MPEA derivatization. The results showed that the peak areas of these compounds after derivatization were increased at different levels after derivatization. For DnsCl-OH derivatization products, the detection sensitivity was improved by 2-649 folds; while the detection sensitivity of MPEA-COOH derivatization products was improved by 2-25610 folds, which further indicated that the CIL strategy was necessary and sufficient for the field of the exposome.

Then, we investigated the fragmentation behaviours of derivatization products by ESI-MS/MS. For the hydroxyl part, taking bisphenol A as an example, the most abundant product ion of 171.1 Da, arising from the skeleton of DnsCl, is always present in the MS2 spectra of various DnsCl-OH derivatization products; while its counter-part product ion, 173.1 Da, is present in that of $^{13}C_2$-DnsCl-OH derivatization products (Fig. 5.2B). These characteristic product ions appeared in nearly all MS2 spectra of selected hydroxyl compounds. For the carboxyl part, it is similar to the hydroxyl group. For example, the product ion of 136.1 Da was produced from MPEA-ketoprofen derivatization products. Similarly, the product ion of 139.1 Da was matched with d_3-MPEA-ketoprofen derivatization products (Fig. 5.2C). Therefore, it is easy to predict the product ions of the environmental pollutants and their transformation products in our CIL database according to their functional groups.

In addition to the above, the combination of light- and heavy-reagents in CIL strategy could reduce the false positive rate of screening environmental biomarkers. Since light- and heavy- reagents possess almost the same chemical properties, the corresponding detected light- and heavy-derivatized products would produce similar LC and MS behaviours when the products were mixed equally before injection, including the difference value of RT < 0.1 min, similar MS intensity (0.5 – 2.0). For example, there were three pairs of peaks in the chromatograms of MPEA/d_3-MPEA-ketoprofen derivatization products (Fig. 5.2D). Based on the principles of the CIL strategy, paired peaks 2 and paired peaks 3 were directly excluded due to different MS intensity and RT, respectively. Compared to traditional full scan methods, our strategy greatly decreased false positive rates. In addition, the relative ratio between the two can provide a very effective method to normalize spectral intensity for the compound quantification even with the matrix effect.

5.4.4 Optimization of collision energy values of CIL MRM transitions

With default product ions of DnsCl- or MPEA-derivatized products, we further optimized their optimal collision energy (CE) by setting the values as 10 eV, 20 eV, 30 eV, 40 eV, 50 eV, and 60 eV to obtain relatively high MS response. As indicated in Fig. 2E, when the CE value of 171.1 Da/173.1 was between 30 eV and 40 eV, the MS intensity of most DnsCl/$^{13}C_2$-DnsCl-OH derivatization products reached its plateau; thus, CE at 35 eV was selected as the optimal value for the compounds in CIL exposome database. For MPEA/d_3-MPEA-COOH derivatization products, the CE value of product ions of 136.1 Da/139.1 Da was suitable for most products (Fig. 5.2F). It is possible that the introduction of derivatization reagents to the original analytes makes corresponding derivatization products possess similar chemical skeletons. Therefore, the CE values of MRM transitions of these derivatization products were similar, in addition to the same product ions. In the same way, we optimized the CE values in Water TQ, Agilent TQ, and AB QTRAP MS machines for CIL-MRM transitions (data not shown). The optimized results were similar. Therefore, it is possible for us to optimize the MRM transition without the authentic standards after derivatization.

5.4.5 Prediction of RT window of derivatization products

Besides product ions and CE values of the CIL-pMRM exposome database, another essential component of the CIL-pMRM exposome database was the RT window of derivatized environmental pollutants and their transformation products. Due to the lack of standards for the vast number of exposomic chemicals, we adopted one machine learning RT prediction model by using available standards to obtain the RT windows of 110,000compounds in the database. As mentioned in the method, the SMILES structures of > 110,000 derivatized compounds were firstly automatically generated. With the available SMILES structures of > 110,000 and subsequently calculated molecule descriptors together with experimental RT of the selected chemicals, we have successfully predicted the RT of the compounds with high accuracy using the random forest methods (Moriwaki et al. 2018). At the same time, we also got the exact *m/z* of derivatization products by the model; thus, the CIL exposome database could also be used for HRMS detection. Take MPEA-COOH part as an example, 173 carboxyl compounds were derivatized by MPEA and their RT was

confirmed by LC-MS. Then, these 173 carboxyl compounds were divided into two sets, a training set (132) and a testing set (44) according to the previous work. A training set was used to test the RT prediction model. The results indicated that the predicted RT window of MPEA-COOH derivatization products was 0.92 min with r^2 = 0.93 for the total running time of 20 min. Similarly, the RT window of DnsCl-OH derivatization products was 1.30 min with R^2 =0.85. Compared with traditional RT prediction of regular metabolites, the prediction accuracy of derivatized chemicals has been much improved due to the chemistry homogeneity of the compounds and their interaction with the column.

In sum, we have obtained all essential components of MRM transitions, including precursor *m/z*, product ions, CE values, and RT windows using a high-throughput and non-standard method. Therefore, the CIL exposome database of 110 k environmental pollutants and their transformation products was built up. To achieve automation, all the above procedures were integrated by a computational package. If the interested compounds were not included in the original CIL exposome database, the procedures would be performed for all those functions, including identification of functional groups, generation of derivatized SMILES structures, prediction of RT windows, and calculation of exact *m/z* of derivatization products. Overall, the database provided references for both known and unknown exposure biomarkers for further relationships with human diseases.

5.4.6 Generation of dynamic MRM optimization algorithm

After the CIL-pMRM exposome database is established, users could search it for targeted compounds, followed by LC-pMRM-MS analysis. However, when hundreds of compounds in CIL exposome database need to be analysed, multiple MRM methods in different runs were very necessary due to the limited capacity of scanning speed. In addition, users always had different targeted search lists in different studies. Therefore, we generated a dynamic MRM optimization algorithm to ensure efficient detection, including grouping of compounds and dwell time optimization. Briefly, targeted compounds were first ranked according to their *m/z*. Compounds with the same *m/z* were not allowed to be included in the same CIL-pMRM method. Since the difference values of the *m/z* of light- and heavy-derivatized OH or COOH products were 2 Da (hydroxyl group: *m/z* of $^{13}C_2$-DnsCl — *m/z* of

DnsCl), or 3 Da (carboxyl group: *m/z* of d_3-MPEA — *m/z* of MPEA), we grouped the *m/z* of these compounds with difference value at 2 Da or 3 Da into different CIL-pMRM methods. Therefore, the combination of these two grouping standards greatly simplified the subsequent data processing.

After grouping standards were performed, the capacity of one segment (1 min) to hold MRM transitions on LC-MS was another issue that should be taken into consideration. Excess MRM transitions would always lead to non-gaussian distributed chromatographic peak shape resulting from limited detection points. The setting of dwell time value was vital for the capacity of one segment. Thus, we optimized the dwell time of the MRM method from 5 ms to 100 ms on Waters TQ. As shown in Fig. 5.3E, the peak shape is gaussian distributed when the dwell time was $\geqslant$15 ms. When the dwell time value of the method was set as lower than 15 ms, there would be rough chromatographic peaks. Therefore, we set the dwell time value of Waters TQ as 15 ms and one segment could contain almost 50 MRM transitions. For AB QTRAP, the dwell time at 3 ms was enough, and thus one segment could contain almost 250 MRM transitions. Overall, the MRM transitions of targeted compounds were randomly grouped into different CIL-pseudo-MRM methods according to their *m/z* and one segment of LC contained up to 50/250 MRM transitions. The output results in the format of CSV. can be directly imported to the MRM method of TQ.

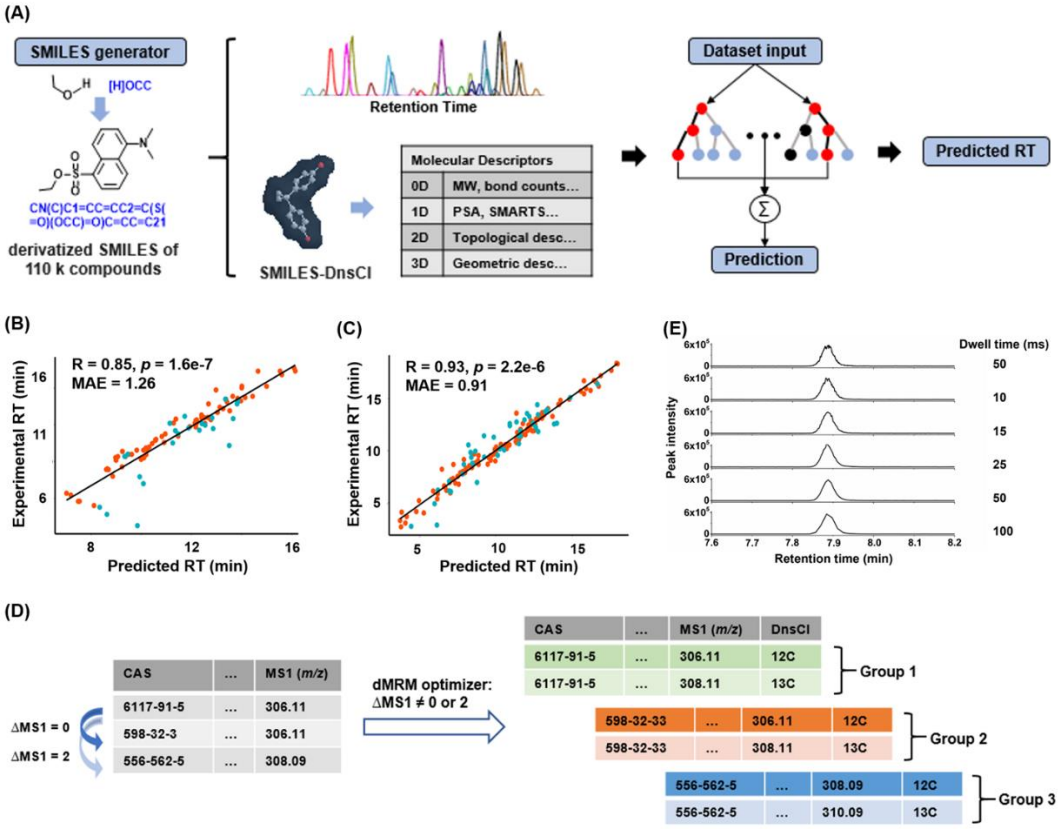


Figure. 5.3. (A) Workflow of RT prediction model for derivatization products. (B) Predicted RT window of MPEA-derivatized carboxyl compounds. (C) Predicted RT window of DnsCl-derivatized hydroxyl compounds. (D) Description of grouping functionality. (E) Optimization of dwell time of MRM transition on Waters TQ.

5.4.7 Peak picking, quantification, and statistical analyses

Peak picking and integration were the fundamental procedure for automatic data analysis. However, the chromatographic retention behaviors of compounds on the column were usually different arising from their various column chemistry, such as hydrogen bonding and ion exchange between hydroxyl, amino, and carboxyl groups and stationary phase. The peak shape of environmental pollutants was usually irregular, especially for those with low abundance and thus manual peak assignment is necessary. However, it seems to be impossible to assign the peak manually for the large dataset of the exposome.

In addition to peak area, peak height can also be used as an evaluation parameter for quantitative analysis, which can often be directly obtained from MS software without manual correction. However, the applicability of intensity in the quantification was based on the condition that a gaussian distributed chromatographic

peak shape exists across different samples. In theory, CIL strategy could greatly improve the peak shapes of compounds due to the similar chemical structure after derivatization. As shown in Fig. 5.4A-B, the peak shape of gemfibrozil was irregular; however, the peak shape of derivatized gemfibrozil and olopatadine tended to be gaussian distributed and the software gave their peak area accurately after MPEA derivatization.

We further explored the possibility of using intensity for the quantification rather than peak area. We fitted the correlation of peak height and peak area of non-derivatized carboxyl compounds, there was no obvious linear relationship resulting from the irregular peak shapes (Fig. 5.4C). In contrast, among the chromatograms of 173 MPEA-derivatized carboxyl compounds, it was found that almost all peaks of the derivatized compound were gaussian distributed and the ratios of peak intensity to the peak area of different compounds were similar. The R^2 of the fitting curve of peak intensity and peak area of 200 pg/mL MPEA-COOH derivatization products was 0.9620, indicating the peak shapes for most derivatization products were nearly normally distributed. Then, we used gradient samples to investigate the correlation of peak height and peak area of MPEA-COOH derivatization products. The results indicated that the peak shape of derivatization products was still symmetrical even under extremely low concentrations (20 pg/mL: $R^2 = 0.9550$, 20 pg/mL: $R^2 = 0.9692$), revealing that the peak shape of carboxyl compounds was similar after MPEA derivatization. As for OH part, we also obtained similar results (Fig. 5.4D). Overall, CIL strategy improved the peak shapes of carboxyl and hydroxyl compounds, so the peak intensity of these compounds could be used for quantitation and high-throughput analysis for the exposome was possible. Subsequently, *t*-test and ANOVA were used for the data statistical difference analysis of the potential environmental biomarkers in biological samples for the pairwise and multiple group analysis; respectively.

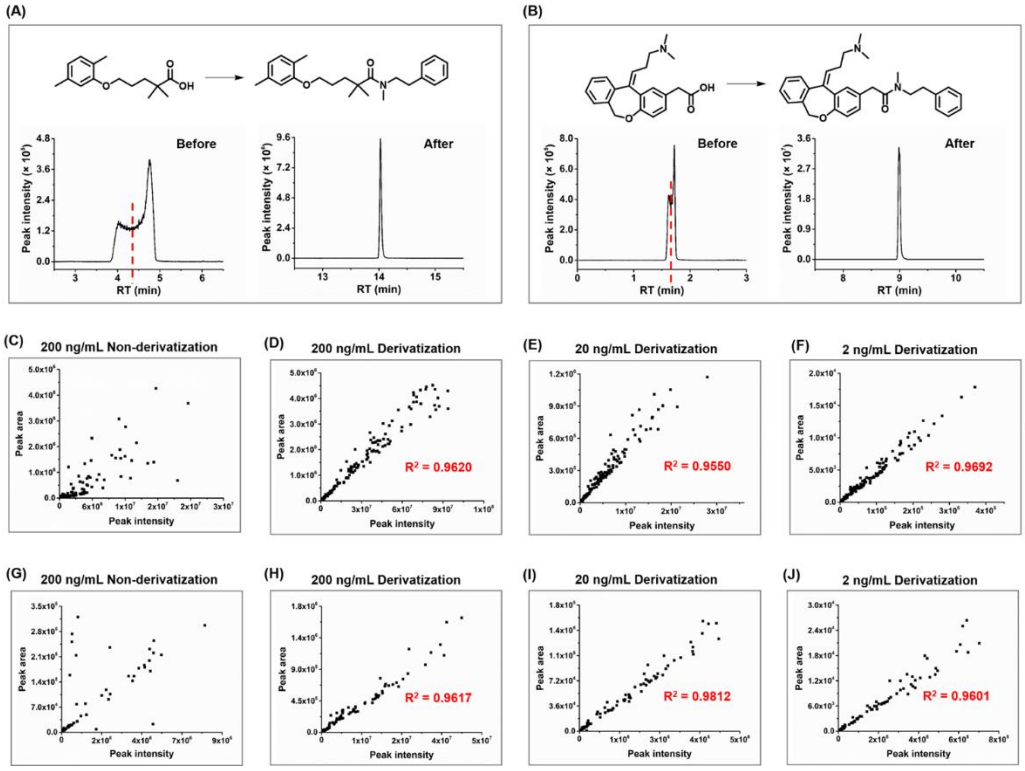


Figure. 5.4. (A) Chromatograms of gemfibrozil before and after MPEA derivatization. (B) Chromatograms of olopatadine before and after MPEA derivatization. (C) 200 ng/mL intact carboxyl compounds. (D) 200 ng/mL MPEA-derivatized carboxyl compounds. (E) 20 ng/mL MPEA-derivatized carboxyl compounds. (F) 2 ng/mL MPEA-derivatized carboxyl compounds. (G) 200 ng/mL intact hydroxyl compounds. (H) 200 ng/mL DnsCl-derivatized hydroxyl compounds. (I) 20 ng/mL DnsCl-derivatized hydroxyl compounds. (J) 2 ng/mL DnsCl-derivatized hydroxyl compounds.

5.4.8 Evaluation performance of the dynamic CIL-ExPMRM platform

After sample pretreatment was fixed, a series of experiments were carried out to test the ability of potential peak screening and statistical analysis in urine samples. We spiked two groups of hydroxyl standards into urine samples, one involved 25 (c1), and the other involved 25 (c1) + 31 (c1). Then, a dynamic MRM optimization algorithm was performed, and CIL-pseudo-MRM methods were generated. Next, an aliquot of urine extracts was derivatized with DnsCl and $^{13}C_2$-DnsCl, respectively, followed by LC-pMRM-MS data acquisition and automatic data analysis. Finally, 25 compounds were searched out in the first group; while 56 compounds were found in the second group by the computational pipeline, indicating that our developed

pipeline could be smoothly carried forward in both individual and parallel urine samples.

To further check if the computational pipeline could be used for environmental biomarker discovery, we set another group of the same 31 hydroxyl standards with group 2 at the concentration of c2. Therefore, 25 compounds with the same concentration (c1) and the other 31 compounds with different concentrations (c1 and c2, c1/c2 = 2.0) were individually distributed in two test groups. Fortunately, 31 compounds with different concentrations were also screened and output as significantly changed compounds (fold change > 1.5 and p value < 0.05) (Figure 5.5D). For example, bisphenol AF (abbr: bpaf) was one of 31 hydroxyl compounds spiked in two urine samples with a concentration of c1 and c2, respectively. It was detected as significantly changed compounds with fold change = 1.7 and p = 5.03E-06. In the view of users, these screened potential environmental pollutants could be further annotated with standards. Our CIL-ExPMRM platform decreased the cost during the preliminary work of biomarker exploration. Therefore, our computational pipeline was certified to achieve environmental biomarker discovery in urine samples.

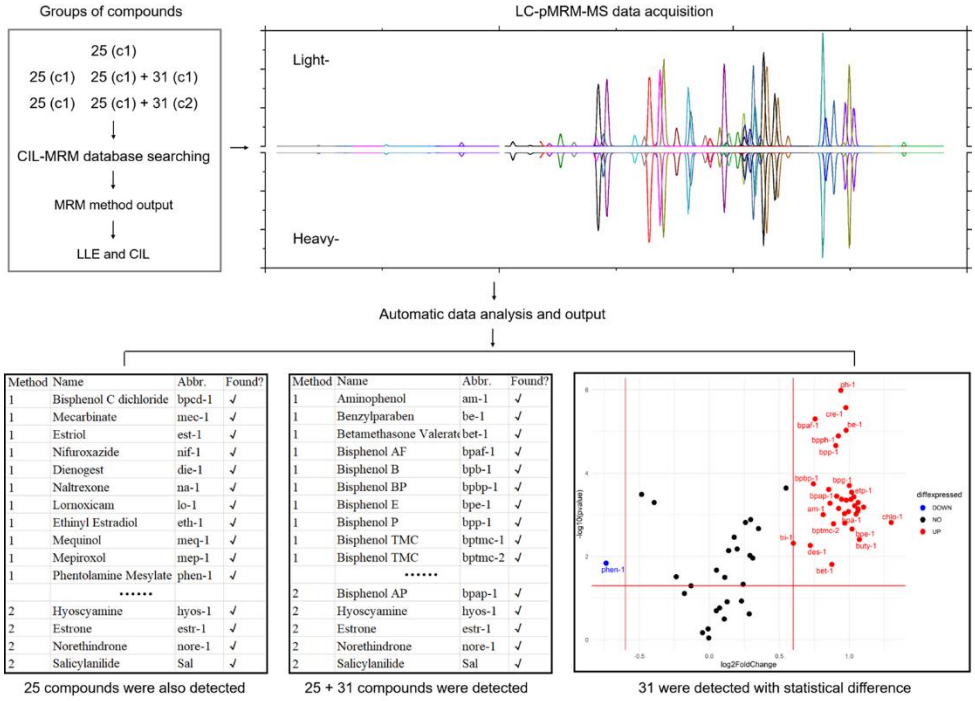


Figure. 5.5. Case study with hydroxyl compounds spiked in urine samples.

5.5 Conclusion

In conclusion, a dynamic pseudo-CIL-MRM exposome platform was developed to address a critical need for public health research. A CIL-pseudo-MRM database containing more than 110,000 environmental pollutants and their transformation products was built up and could be constantly replenished by using our computational pipeline. A series of software packages were used to optimize MRM transition parameters, curate the MRM database, predict accurate retention time, and automize dynamic MRMs. Automatic MRM peak assignment, alignment, and data statistical analysis were also achieved by a computational pipeline. Overall, the platform could be used for discovering environmental biomarkers by different brands of MS instruments. Overall, the whole platform was promising to be widely used.

CHAPTER 6: THE DEVELOPMENT OF A NOVEL RISK-BASED NON-TARGETED ANALYSIS FOR CHARACTERIZING THE HUMAN EXPOSOME

6.1 Summary

In chapter 6, we developed an NTA-based prioritization workflow for the fast prioritization of organic chemicals. We compiled the workflow into an R package application, "NTAprioritization". This R package automates screening and prioritization of identification lists by determining the matching scores between the experimental spectra and the libraries (either experimental or *in silico*), the delta retention time (deltaRT) value between the experimental RT and predicted RT by the random forest model, and the toxicities predicted by EPA TEST and ToxCast. The package allows users to adjust the score thresholds and six types of toxicity endpoints based on their matrix. To test our package for application, we used this workflow to prioritize the identification lists gained by the NTA method from a real sludge water sample spiked with 28 environmental pollutants. The workflow reduced the hit list of over 6,982 candidates to a final list of 1,577 compounds with 6 toxicity endpoints. Further, we classified the lists into 4 groups with their prioritization order. Finally, we found 25 of 28 spiked standards across 4 prioritization lists. Overall, this study shows the added value of an automated prioritization R package for fast screening of known and unknown compounds based on the NTA method.

6.2 Introduction

In this chapter, we hypothesize that the laborious prioritization protocol could be remedied in an automated programming workflow by harvesting computational power for in silico prediction of retention time and toxicity of potentially toxic compounds observed by NTA approaches. Toxicity prediction for toxic compounds is used as potential of exposure for prioritization. Several databases and software programs have been developed to derive toxicity prediction, such as Toxpi, ToxCast, TEST, etc (Martin and Todd 2020a; Reif et al. 2010, 2013). Toxicity data including multiple endpoints carcinogenicity, mutagenicity, genotoxicity, endocrine disruption, and developmental toxicity are comprehensive so that accurate predictions are available for broad ranges of compounds. Both experimental and computational toxicity has been included to meet the demand of various studies.

Overall, there is a demand for additional LC-MS-based filters and a fast prioritization process to reduce the number of features so that environmental exposures posing environmental or human risks can be identified in the first order. In our study, we developed an R package that applies multiple filters in the prioritization of potentially toxic compounds based on NTA acquisition results (Fig. 6.1). The developed workflow can fulfill the following tasks: A) Sample analysis by LC-HR-MS/MS using DIA method; B) Raw data after the acquisition was deconvoluted and candidate lists were generated by MS2 spectra matching; C) Candidates lists were filtered by predicted retention time and MS2 spectra ranking; D) Candidates lists were further filtered by six toxicity endpoints predicted by EPA TEST software and Toxipi score by search ToxCast experimental toxicity database; E) Candidates prioritized in 4 tier levels were generated by toxicity ranking combined with identification ranking. Finally, for demonstration and evaluation, we applied the package to candidate lists gained by LC-HRMS-DIA from a real sludge water sample spiked with 28 environmental pollutant standards with different toxicity potentials. The workflow reduced the hit list of over 6,982 candidates to a final list of 2,779 compounds. We prioritized 21 spiked environmental pollutants in 5 tiers, with 2 compounds highlighted at the first priority level (Tier 1). The package is available online (https://github.com/FangLabNTU/NTAprioritization.git).

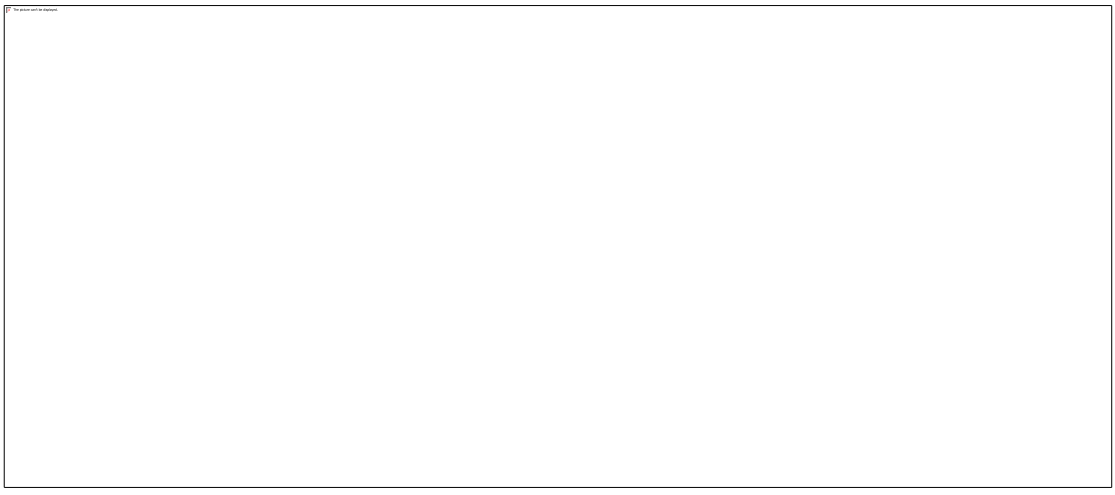


Figure 6.1. Overview of NTA-based prioritization workflow

6.3 Methods and materials

6.3.1 Sample preparation

All solvents including acetonitrile (ACN), methanol (MeOH), and water in LCMS grade were purchased from Sigma (Singapore). 28 pesticides and 146 FDA drugs were obtained from Waters (Singapore). and Sigma (Singapore). For the retention time prediction model development, 146 FDA drugs are prepared in methanol and mobile phase A (1:3). The drug mixtures were kept under -20°C until further analysis. For the validation case of our workflow, 28 chemicals stocked in methanol were spiked into the sewage water sample by 200 ppb to mimic the environmentally relevant concentrations (i.e, low ppt level in the original water). The sewage water sample (from the wastewater treatment plant) was filtered, reconstituted, and concentrated by 500 times in methanol and mobile phase A (1:3). The solvent in samples was used as blank for group comparison. The mixtures were kept under -20°C until further analysis.

6.3.2 LC-HR-MS analysis condition

We used Waters Xevo G2-XS QTOF mass spectrometer with an electro-spray ion source (ESI) interface for data acquisition, coupled with the ACQUITY I-Class ultra-performance liquid chromatography (LC-HR-MS). The ACQUITY HSS T3 column (100 x 2.1 mm, 1.8 µm, Waters) was used for separation at 30 ^{o}C. The mobile phases comprised 0.1% formic acid (FA) in H_2O (A) and 0.1% FA in acetonitrile (B). The mobile phases gradient started with 90% mobile phase A at the first 2 mins, raised to 100% mobile phase B evenly from 2 to 15 mins, hold for 5 mins, decreased to 90% mobile phase A for 1 min, and hold for equilibrium for 9 mins. The injection volume was 2 µL.

The MS conditions were capillary voltage at 3.0 kV in positive and 2.5 kv in negative mode, ion source temperature at 120 ^{o}C, desolvation temperature at 500 ^{o}C, and desolvation gas flow at 1000 L/h. QTOF was operated in MS^E mode with low collision energy and high collision energy at 6 and 10-40 eV, respectively. The mass spectrometer was calibrated using 0.5 mM sodium formate in the solution of 2-propanol and water (90%/10%). The lock mass was set up using 200 pg/µL leucine enkephalin through the analysis to ensure the mass accuracy of the chemicals of

interest. The lock mass for equipment calibration in positive mode and negative mode were 556.2771 and 554.2615 respectively.

6.3.3 Sample analysis and raw data pre-processing by Waters Progenesis QI

For retention time prediction, the mixtures of 28 chemical standards were analysed and their retention time was collected for prediction model development. Detailed chemical and equipment acquisition conditions were described in Table S6.1. For the validation case study, real samples and blanks were analyzed by LC-HR-MS in triplicates. The acquired raw data were pre-processed in six steps (Fig 6.2). (1) Peak picking: chromatograms were identified by peak picking. In Progenesis QI, we performed a peak-picking algorithm using different methods with 3 levels of sensitivity, including automatic, absolute ion intensity, and % base peak. Each sensitivity method filtered noise peaks (signal-to-noise ratio S/N <3). Another parameter of peak picking is the selection of the adduct. $[M+H]^+$, $[M+Na]^+$, $[N+NH_4]^+$, $[M+CH_3OH+H]^+$, $[M+K]^+$, $[M+ACN+H]^+$, $[2M+H]^+$, and $[2M+Na]^+$ were selected as adducts for peak picking in positive mode. $[M-H]^-$, [M-H2O-H]$^-$, $[M+Cl]^-$, $[M+HCOO]^-$, $[M+CH_3COO]^-$ were selected in the negative mode. (2) Peak alignments: In Progenesis QI, alignment is driven by placing landmarks called alignment vectors. Each vector connects the position of a specific compound ion on the reference run with the position of the same ion on the run being aligned. Once the vectors are placed, they are used to calculate a non-linear mapping between the retention times of the reference run and those of the run being aligned. (3) Deconvolution: after peak picking, adducts of the same compound are grouped. Progenesis QI uses our selected adduct information (Fig. 6.2-3 top) for peak picking to determine which groups of compound ions are likely to be different adduct forms of the same compounds. For different compound ions to be considered part of the same compound, they must have the same retention time (Fig. 6.2-3 bottom right) and have very similar mass spectra (Fig. 6.2-3 bottom left). (4) Multivariable analysis: PCA analysis was performed for deconvoluted MS peaks between samples and blanks (Fig. 6.2-4 top). The S-plot was used to find markers with significant differences between the two groups studied (Fig. 6.2-4 bottom).

One-way ANOVA calculation was used to produce a p-value for the markers in these two groups of data. (5) Library searching for MS spectra by progenesis

MetaScope was performed for peak annotation. Fragmentation scores for each peak were generated by spectra matching. For Chemspider database, in-silico predicted MS/MS spectra were used as a reference for spectra matching. For the setting of in-silico fragmentation, the tolerance of both precursor ion and fragmentation ion was set at 5 ppm. The isotope similarity score was set at 95%. The database search generated fragmentation score, mass error, and isotope similarity. The maximum score for each matching criterion is 100. An equal weighting of 20 for each factor is combined to the total score with the maximum value of 60 (Fig. 6.2 bottom left showing the fragmentation score). (6) Peaks identification: The candidate list for each peak by library searching, including compound name, compound ID, neutral mass, m/z, retention time, adduct, formula, mass error, isotope similarity, and fragmentation score, were exported in .csv file for further processing.

Figure 6.2. Candidates list acquisition by data pre-processing in Waters Progenesis QI

6.3.4 Retention time prediction model development

The implementation of the retention time prediction model followed the procedure in session 3.3.3 in chapter 3. Specifically in this chapter, an experimental retention time dataset of 146 compounds was used for modeling. The dataset was

acquired in the LC condition as described in session 6.3.2. All compounds were detected between 2.1- and 16.7-min retention time with a total run time of 30 min. The entire dataset was divided into two subsets for model training and testing. The training and testing splits were made in the ratio of 75:25, respectively, with the caret package in R environment. All data are given in Supporting information Table S3.2.

The correlation R^2 values between observed and predicted retention times were used to indicate linear relationships and for global generalization of the prediction set (Fig 3-A). The hyperparameters of the prediction model follow the procedures in session 4.3.4 in chapter 4. The fine-tuning of the model returned the tree number of 500 and the mtry value of 9. After the model development, both model and the datasets were saved into the packages for direct implementation. The highly efficient random forest regression tree algorithm allows users to train their models and save the models on the local machine with their dataset. In other words, the prediction model can be modified by users' datasets acquired by the local LC system, leading to a lower prediction error. We have developed this workflow as a user-friendly function, RTpredict() inside the NTAprioritization. As more experiment datasets are provided, users can generate models with higher prediction accuracy.

6.3.5. Estimation of six toxicity endpoints and ToxPi score

Candidates' toxicity at six endpoints was estimated using the Toxicity Estimation Software Tool (TEST, version 5.1.1) (Fig 6.4). The prediction is applied to compounds containing only the following element symbols: C, H, O, N, F, Cl, Br, I, S, P, Si, or As. The QSAR-ready SMILES code for each chemical was submitted to the models. The 96-hour fathead minnow LC50 (FMLC50), 48-hour daphnia Magna LC50 (DMLC50), 40-hour Tetrahymena pyriformis IGC50 (TPIGC50), rat oral LD50 (ORLD50), developmental toxicity (DT), and Ames mutagenicity (AM) toxicological properties were estimated to evaluate the chemical toxicities in this study. All prediction models are developed with experimental data sets according to the EPA manual. Prediction accuracies are acceptable for each toxicity endpoint.

As shown in Fig. 6.4A, Developmental toxicity and Ames mutagenicity are binary endpoints. The developmental toxicity endpoint (DT) is the relationship of chemicals with developmental toxicity outcomes in humans and/or animals. The Ames mutagenicity endpoint (AM) is detected by the positive mutation results of test

chemicals in the Ames test, indicating the potential carcinogenicity and teratogenicity. If the calculated score of these two endpoints >= 0.5, the compound activity is positively toxic. The oral rat LD_{50} ($ORLD_{50}$) is the dose of chemical required to kill half the members of a tested population after oral ingestion (mass of chemical in mg per body weight of the rat in kg) ($-Log_{10}$ (mol/kg)), which is one of the most important endpoints of rodent acute toxicity. Chemical toxicity in water organisms can be presented by the 50% growth inhibitory concentration for T.pyriformis after 40 hours (TPIGC in $-Log_{10}$ (mol/L)). Aquatic toxicity of chemicals is also presented by the fathead minnow LC_{50} endpoint (FMLC50) and Daphnia Magna LC50 endpoint (DMLC50). The FMLC50 is the concentration in water that is lethal to half of the exposed fathead minnow in 96 hours (in $-Log_{10}$ (mol/L)). The $DMLC_{50}$ represents the concentration in water that is lethal to half of exposed D.magn in 48 hours (in $-Log_{10}$ (mol/L)). ToxPi score is calculated as a weighted combination of all data sources which comprehensively estimates the toxicity ranking of chemicals based on ToxCast database and provides a transparent visualization of the relative contribution of all information sources to an overall priority ranking.

Besides the 6 predicted toxicity mentioned above, we applied the Toxicological Priority Index (ToxPi) model in 97 ToxCast *in vitro* assay data to further expand our toxicity data as described in our previous study. For the ToxPi modeling, we included the targets of the estrogen, androgen, thyroid pathways, the glucocorticoid receptor, peroxisome proliferator-activated receptors (PPARs), and monoamine signaling, and two physicochemical properties, the octanol-water partitioning coefficient (log P) and bioconcentration factor (BCF). We utilized the potency (AC50) and efficacy (Emax) estimates provided by the ToxCast program as well as estimates for the log P and BCF retrieved from the U.S. EPA Chemistry Dashboard to calculate the ToxPi scores of 8,845 chemicals by ToxPi GUI 2.0.

6.3.6 NTAprioritization R package functions

NTAprioritization enables a complete workflow from raw feature lists to a final prioritized identification list with multiple filters. The workflow starts with two datasheets, library matching scores from waters QI progenesis, and predicted toxicity endpoints. The data_comb() function retrieves all toxicity endpoints data from the toxicity datasheet and imports the data into the library matching datasheet. The get_all() function is used to compile all the toxicity endpoints data of each compound

from all libraries into one datasheet for further screening. The RTpredict() function calculates the SMILES-based molecular descriptors and predicts the retention time of each compound. Compoundclassification() function allows users to specify the toxicity endpoints and retention time threshold for the final prioritization process. The functions of the NTAprioritization package are explained in detail in the GitHub repository (https://github.com/YANGJJ93MS/NTAprioritization.git).

6.4 Results and discussion

6.4.1 Candidate's list acquired by data pre-processing in Progenesis QI

In this study, we analyzed samples using LC-QTOFMS with MS^E mode as a DIA acquisition method to cover a larger chemical space and to detect the maximum number of environmental pollutants in the samples. Progenesis QI was used for data processing, including chromatogram alignment, peak picking, deconvolution, and spectra matching with databases. Ten of thousand molecule features were observed by MS acquisitions and deconvoluted. Multivariate analysis performed on the features identified 137 peaks that varied significantly from blank data (Fig. 6.2). The retention time and mass-to-charge ratio (m/z) for these markers varied from 3 to 21 min and 123 to 1085 (Table S6.2). Multiple chemical libraries were selected to cover the large range of m/z in significant peaks in candidate searching. Each selected library encompasses a large repository of registered chemicals (Fig. 6.2).

Tremendous chemical library resources are available in the software. To focus on environmental exposure, libraries containing common environmental pollutants, drugs, and their metabolites were preferred in our study. We selected four libraries within Chemspider database via Progenesis QI, including EPA Toxcast library (N = 8546), EPA DSSTox library (N = 1146633), NIST library (N = 166473), and NIST spectra library (N = 195917) External MS2 spectra databases were download and imported into Progenesis QI, including NIST MS2 library (N = 9378), MassBank of North America (MoNA) (N = 227045), and MassBank of Europe MS2 (MBE) (N = 14788). Candidate hits by searching each library varied due to different library sizes, with 51 hits from the EPA Toxcast library, 6981 hits from EPA DSSTox library, 997 from NIST library, 1069 hits from NIST spectra library, 508 hits from NISTMS2, 767 from MoNAMS2 library, and 148 from MBE MS2 library. Overall, the powerful Progenesis QI software enabled a high throughput raw data preprocessing. Over 7000

candidate hits, with total matching scores in the range of 0 to 60, showed enough coverage for environmental pollutant detection in this workflow. The total score from matching results was useful for downstream screening and prioritization.

6.4.2 Retention time prediction model by random forest tree model for fast implementation

Retention time prediction by random forest regression models based on QSRR is widely accepted for non-targeted analysis (Bonini et al. 2020; Cao et al. 2015b; Zdravković et al. 2018). In our previous study, we achieved a comparable prediction accuracy by QSRR-based regression model with features selection by random forest algorithm (Yang et al. 2020). However, the modeling process requires tedious model optimization, including feature selection and linear model parameters optimization based on the selected features. It leads to a time-consuming model development process before real application. In this study, we apply a random forest tree algorithm for direct prediction of retention time. Feature selection and hyperparameter optimization can be operated automatically without manual monitoring. The high efficiency of tree models allows the fast development of prediction models from raw datasets.

Overall, the model achieved linear correlations $R^2 = 0.86$ (p = 2.4e-11) in test data with mean absolute error (MAE) of 1.00 min (Fig. 6.3A). With an experimental retention time of 28 chemical standards for evaluation, the prediction shows good accuracy with a median prediction error of 1.11 min and an average prediction error of 1.14 min (Fig. 6.3B). The maximum prediction error is 3.50 min, and the minimum prediction error is 0.086 min. Detailed data are described in Table S1. Fig.S6.2 shows the deltaRT value for all compounds in the features lists using equation 1. In the total range of 0 to 15.5 min, deltaRT of most features fall within 4 min. To include more possible hits with tolerance, we finally set a larger RT threshold at 4 min as a secondary filter. A conservative prediction error was applied to the process of retention time matching between the retention time of identified peaks and the predicted value of candidates. The matching threshold was set at 2 min to achieve more identifications in the prioritization workflow.

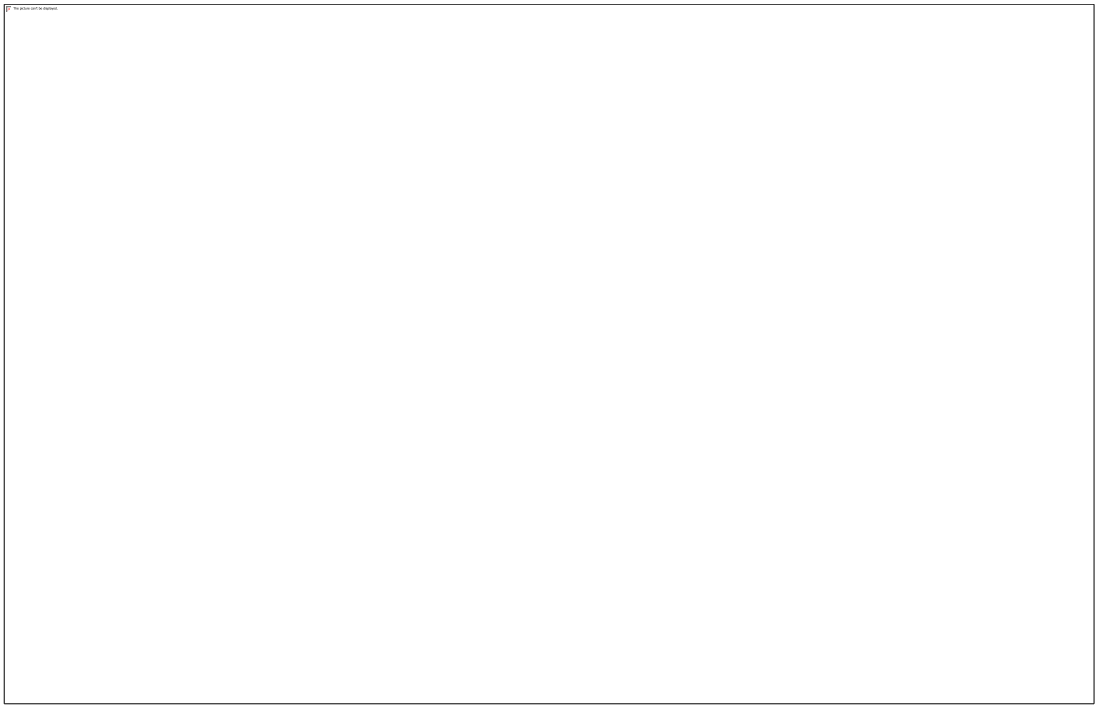


Figure 6.3. A) Retention time prediction by experimental retention time dataset of 146 environmental pollutants. B) Evaluation of predicted retention time of 28 spiked chemical standards. C) Candidates lists prioritized by predicted retention time and matching scores from library matching in waters progenesis QI

6.4.3 Prioritization by Retention time and library searching scores

Prioritization of possible analytes proceeds with retention time matching and library searching scores. Here, deltaRT from retention time matching and maxscore from library searching scores together indicate the confidence level of identification. DeltaRT(ΔRT) is determined by the difference between the experimental retention time observed by MS acquisition and the predicted retention time (equation 1). The maxscore is calculated by selecting the maximum value of all matching scores among 7 selected spectrum databases during the raw data processing step that was processed by water Progenesis QI (equation 2). The total matching score is a combination of mass error, isotopic pattern similarity, and MS2 fragmentation pattern matching. In seven selected chemical libraries, NIST library, NIST spectra library, EPA DSSTox, and EPA Toxcast library were provided by Chemspider database. Experimental MS2 spectra were unavailable in these libraries. In replacement, in silico predicted MS2 spectra for each candidate were provided by Progenesis QI. Candidates with no fragmentation score (fragmentation score = 0) were filtered. And the maximum

maxscore of remaining candidates among all libraries was selected as the final maxscore.

$$\Delta RT = |RTexp - RTpred| \quad \text{equation (1)}$$

$$maxscore = Max\ \{scocres_{NIST}, scores_{MoNA}, scores_{EPA}, \ldots\ldots\} \quad \text{equation (2)}$$

Candidates were classified into 5 levels of priority, according to user-defined deltaRT and maxscore threshold (Fig. 6.3 C). In our application, the prediction error from the prediction model is set to 2 min so that compounds with ΔRT fall into 2 min and are considered as strongly identification. Compounds with ΔRT within 4 mins are considered as less identification. Maxscore threshold is set to 30-40 determined by the overall dataset distribution. For instance, four candidates were acquired by library search for the peak with RT = 9 min. By calculation of delta RT, candidate a was a good match with ΔRT = 0.3 min while candidate d was hardly meet the RT threshold with ΔRT = 5.3 min. By library matching especially in MS2 spectra matching, the m/z with adduct from MS2 fragments of the peak was found in experimental or predicted MS2 spectra of all candidates. However, not all candidates possess the same amount and intensity of secondary or lower fragments as the peak MS2 spectra. Candidate a was classified as RTMS2-Level 1 candidate, due to its narrow ΔRT and its MS2 fragments sharing the highest similarity with the peak MS2 spectra. Candidate d with fragments sharing the worst similarity and with the largest ΔRT was classified as RTMS2-Level 4 candidate. The RTMS2-Level priorition was combined with toxicity prioritization to provide the final priority.

6.4.4 Toxicity prioritization by predicted six toxicity endpoints and ToxPi score

As a pivotal part of this study, we have integrated the computational toxicity from well-established models and experimental toxicity from previous studies into the toxicity database of our platform (Kavlock et al. 2012; Martin and Todd 2020b; Marvel et al. 2018). The users can extract the data from any toxicological endpoints, depending on the major risk concern. Especially, for higher coverage of toxic chemicals space, we introduced 6 key toxicity endpoints and one ToxPi score based on ToxCast database for the application. As shown in Fig. 6.4C, ToxPi and the other six toxicity endpoints were calculated as the toxicity score in our workflow. Candidates were classified as highly toxic compounds (Tox_level 1) and medium toxic compounds (Tox-level 2 and Tox-level 3), regarding the toxicity score. The

selection of different toxicity endpoints based on the model's performance or sample matrix is available in our R application.

Calculated results of the raw candidate hit list of 6982 compounds showed high toxicity at four endpoints (ORLD50, FMLC50, DMLC50, and DT) and low toxicity at two endpoints (TPIGC50, AM) (Fig. S6.2). The ToxPi score showed a broad toxicity range from 0 to 1. For demonstration, we selected ORLD50 as the toxicity endpoint for ranking toxicity and prioritization. After being filtered by retention time, availability of MS2 spectra, and availability of predicted toxicity endpoints, 1477 candidates remained for toxicity prioritization. Among them, 1475 were available for ORLD50 prediction. Specifically, the calculated result of ORLD50 was presented in Fig. 6.4A. Most candidates showed medium toxicity at this endpoint as most points were in medium color, ranging from 1 to 5, with 5 as the highest toxicity.

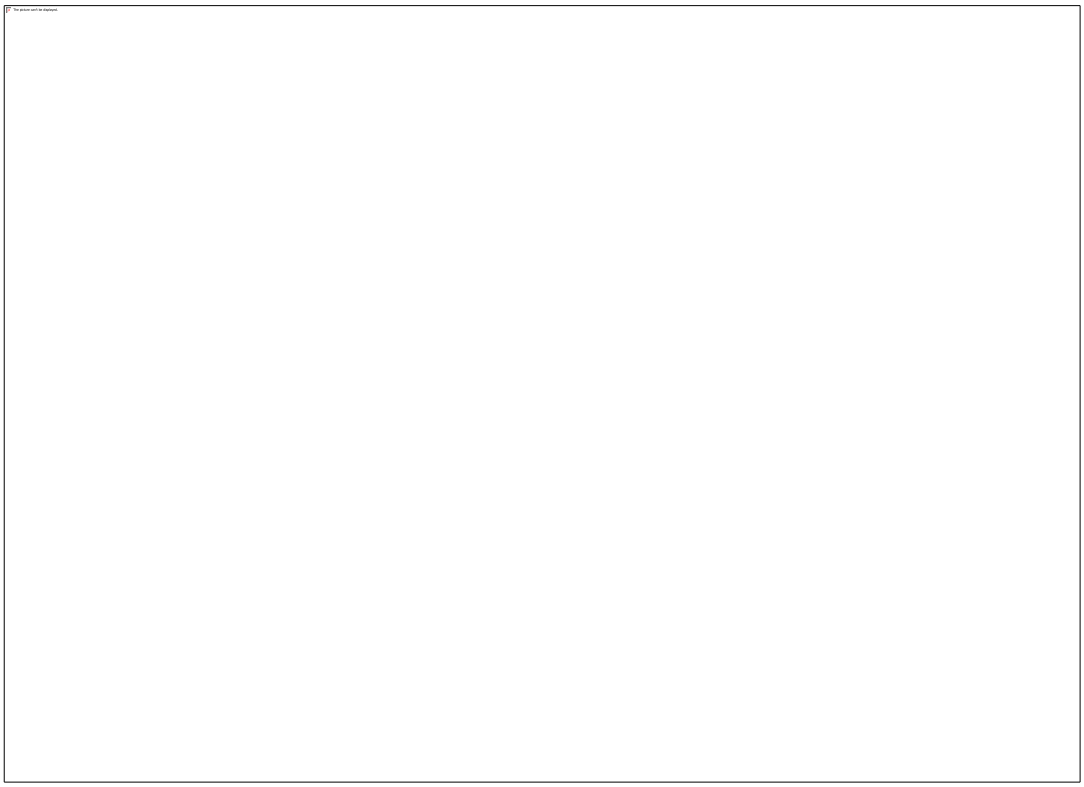


Figure 6.4. A) Predicted toxicity at ORLD50 endpoint. B) The toxicity ranking of 28 spiked chemicals among total candidates. C) Six toxicity endpoints and their prediction accuracy by EPA TEST software and ToxPi.

In our study, the toxicity threshold for each toxicity endpoint were determined in Table S6.4. For ORLD50, compounds were classified by three toxicity levels:

Tox_level 1 with toxicity from 3-5, Tox_level 2 with toxicity from 1.5-3, and Tox_level 3 with toxicity from 0 to 1.5. As a case study for evaluation of our toxicity prioritization workflow, 28 chemical standards including pesticides with different toxicity were spiked and 21 of them were identified. Prioritization of the candidates' list was achieved by ranking 6 toxicity endpoints values. The preliminary predicted ORLD50 values of the total candidates 'list and the 26 spiked standards were described in Fig. 6.4B. Two compounds with high toxicity levels were prioritized in the toxicity ranking curve. The rest of the spiked chemical standards were classified as Tox_level 2 candidates. The ranking curve shows the toxicity-based prioritization process. Compounds with a higher toxicity endpoint value will be prioritized automatically by our application.

6.4.5 Prioritization by combining ΔRT scoring, MS2 spectra scoring, and toxicity levels.

In summary, the prioritization workflow can be summarized in six steps (Fig S6.3): (1) The candidate hit list was imported; (2) the Candidate hit list without MS2 spectral matching records was filtered; (3) Retention time was predicted for candidates by random forest tree models; (4) DeltaRT and the maxscore are calculated by the algorithm according to the equations below (equation 1 & equation 2). Candidates were prioritized using an algorithm based on the deltaRT and maxscore value. (5) Six toxicity endpoints and ToxPi score were calculated by EPA TEST and ToxCast database; (6) Candidates within each class are prioritized by their predicted toxicity based on the selected toxicity endpoints. And finally, a comprehensive prioritization was used to provide a final prioritized list in four tiers of categories, combining retention time, MS2 spectra, and toxicity level.

The 1475 candidate hit list with 26 spiked chemicals prioritized by ΔRT scoring, MS2 spectra scoring, and toxicity levels was presented in Fig. 6.5A-C. The 21 spiked chemicals showed priority in the top 3 RTMS2 levels, with 13 compounds classified as level 1, 3 compounds classified as level 2, and 5 compounds classified as level 3. In each RTMS2 level, candidates were further classified into 3 Tox_levels according to their ORLD50 value. To highlight the exposure posing the most threat to human man, a prioritization ranking pyramid comprised of ΔRT scoring, MS2 spectra scoring, and toxicity levels were proposed in this study. By our definition, final

candidates list were classified into four tiers: tier 1 with RTMS2 level1 and Tox_level 1; tier 2 with RTMS2 level 2 or 3 and Tox_level 1; tier 3 with RTMS level 1 and Tox_level 2 or 3; tier 4 with RTMS2_level 2 or 3 and Tox_level 2 or 3 (Fig. 6.5D).

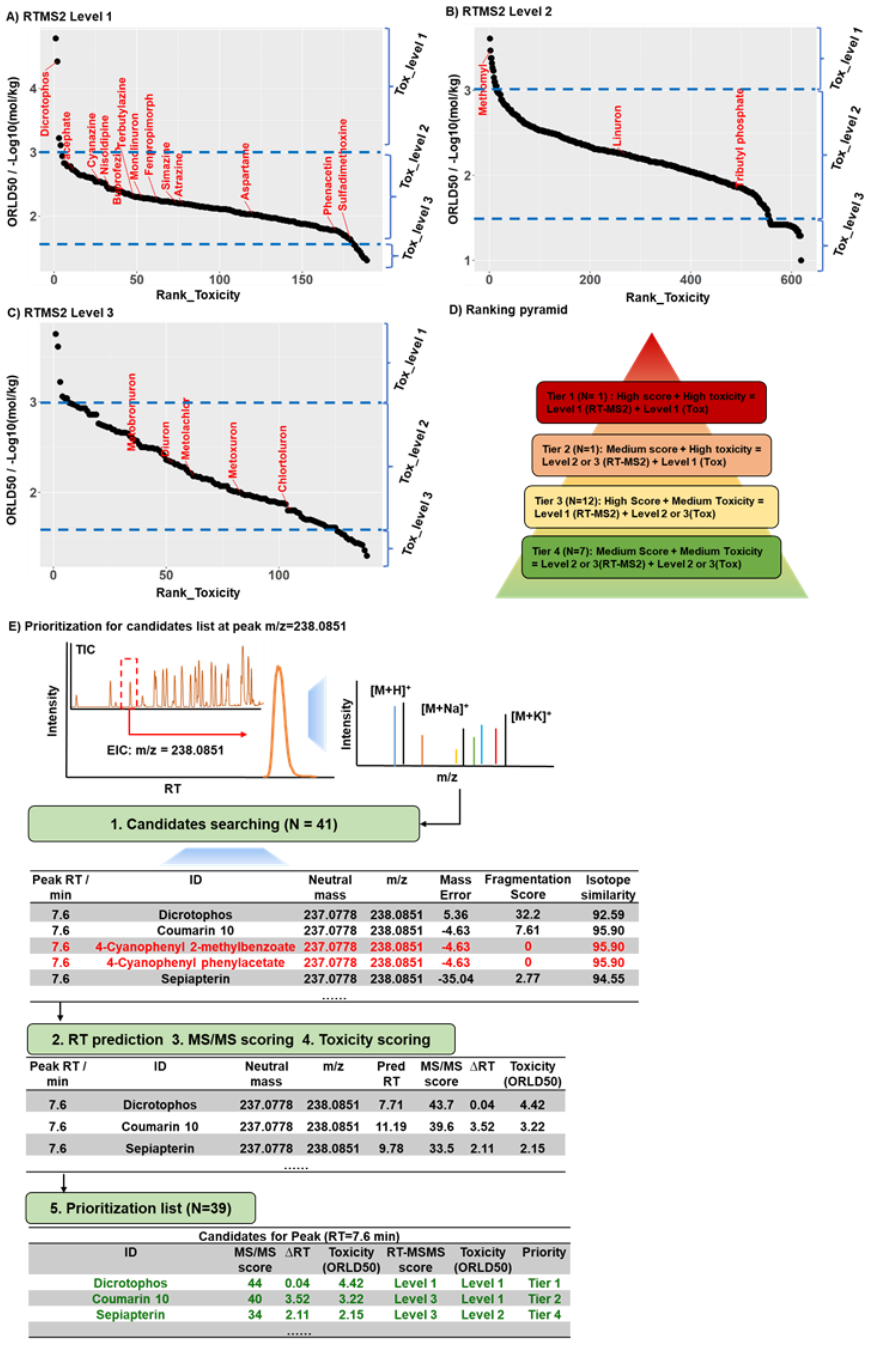


Figure 6.5. A-C) 1475 candidates with 26 spiked chemicals were prioritized by deltaRT, and MS2 spectra matching scoring, and toxicity levels. 26 spiked chemicals were classified in the top 3 RTMS2 levels, with most of them showing toxicity prioritization at Tox_level 2. D) A ranking pyramid of 4 tiers of prioritization combining retention time, MS2 spectra, and toxicity endpoints levels. E) A case study

of the prioritization process for candidate hit list at peak m/z=238.0851 with RT = 7.6 min.

Following the final prioritization pyramid, dicrotophos, as a dialkyl phosphate and an organophosphate insecticide, was classified as Tier 1 compound giving its RTMS2 level 1 and Tox_level 1 priority. Methomyl, as a nematocide, and an insecticide on vegetables was classified as Tier 2 candidate. Other 12 spiked chemicals were classified as tier 3 candidates. And 7 spiked chemicals were classified as tier 4 candidates. The spiked compounds with the highest toxicity were prioritized without losing identification confidence for further identification in this process. Although 12 spiked chemicals were in high RTMS2 level, their priority was lower regarding their lower toxicity to human health. More investigation could be a focus on identifying the toxic candidates.

To demonstrate the prioritization workflow in detail, the prioritization of the candidate hit list for peak m/z = 238.0851 with RT = 7.6 min was described in Fig 3.5E. The raw candidate hit list from Progenesis QI included 41 compounds by library searching. Candidates without MS2 spectra matching scores were filtered (Fragmentation Score = 0). Retention time and ORLD50 toxicity of the remaining 39 candidates were predicted. Maxscore was calculated from library matching. A comprehensive prioritization regarding retention time, maxscore, and toxicity level was conducted on the final candidate list of 39 compounds. Finally, 39 candidates were prioritized in 4 tiers of compounds for further identification.

6.5 Conclusions

This study has provided an NTA-based prioritization workflow for fast prioritization of known and unknown chemicals with high toxicity endpoints to human and environmental health in characterizing human exposome. We developed this workflow into the in-house developed R package, "NTAprioritization.R". This application uses the imported candidate hit list to generate a hit list prioritized by retention time matching, MS2 spectra scoring in library search in waters Progenesis QI, and the toxicity level.

Finally, a real sludge water sample spiked with chemical standards validates this application. The results show that applying the QSPR-based prediction of retention time at the top of spectra matching scores can significantly reduce the number of less confident hits during screening and filtering the candidates by NTA

approaches. Chemical toxicity assessments combined with compound classifications by spectra matching scores and retention time matching scores enable an accurate prioritized compound list. By applying our proposed "NTAprioritization" R-based application, users can achieve a fast prioritization of candidates by user-defined parameters. Further, we can expect efforts on improving the accuracy of the QSRR prediction model and including more toxicity endpoints in the workflow.

CHAPTER 7. CONCLUSION AND RECOMMENDATIONS

7.1 Conclusion

The current phase of exposome research is facing multiple overarching challenges, including the limitation of instrumental method optimization and MRM ion pairs optimization by the availability of chemical standards, low detection coverage for endogenous and exogenous metabolites in the presence of matrix effect, and a lack of convertible instrumental methods for more generalized applications in targeted analysis. On the other hand, there are challenges in characterizing unknown exposure chemicals, such as the time-consuming data deconvolution in non-targeted analysis and the difficulty in the curation of multiple reference databases.

This thesis tackles these challenges and provides several streamlined mass spectrometry-based platforms coupled with gas chromatography and liquid chromatography. This thesis first proposed an optimization equation for the conversion of optimized CE values across different LC-QqQ-MS/MS platforms to analyze small molecules in environmental and biological samples. The platform is built on collective strategies that include method optimizations using the library of MRM transitions and peak matching by the QSRR retention time prediction (Pearson r^2 = 0.63, MAE = 1.35min). The workflow was successfully applied for environmental exposure characterization without chemical standards. Nineteen out of 20 spiked common environmental pollutants were detected in the sludge water sample and urine sample. The xenobiotic transformation product of bisphenol A was well detected in the matrix of cell extract by the workflow.

We further improve the coverage of the targeted method by developing a database of pseudo-single ion monitoring (SIM) and LC-MRM with a powerful algorithm to increase the sensitivity and specificity for over 300,000 exogenous chemicals with available MS/MS fragmentation in the public databases. The pseudo spectral databases cover over 70,295 unique compounds with 1,048,575 pseudo-SIM transitions from NIST EI library and over 10,353 unique compounds with 180,001 pseudo-MRM transitions from MoNA LC-MS/MS library in positive and negative mode. Robust retention time prediction models were developed for GCMS (Pearson r^2 = 0.97, MAE = 67.63 RI) and LCMS (Pearson r^2 = 0.91, MAE = 1.31 min). The pseudo-SIM was applied for the identification of 4 volatile organic compounds spiked in a mixture sample. The results indicated that there were compounds at trace level

out of detection coverage of the current targeted methods. Therefore, we developed a library of chemical isotope labelling-pseudo-MRM for exposure chemicals at trace levels (CIL-ExPMRM). The database covers the pseudo spectra of over 110,000 compounds and their transformation products. The workflow was verified by the characterization of 56 commonly seen exogenous and endogenous compounds in urine samples.

In the final chapter of this thesis, we further proposed one new non-targeted analysis workflow for environmental chemical screening using risk-based chemical prioritization. In this workflow, the identified compounds were prioritized based on spectral matching, retention time prediction, and toxicities prediction by EPA TEST and ToxCast. We compiled the workflow into an R package application, "NTAprioritization". Twenty-one out of 28 compounds spiked were identified and prioritized in five tiers in the sludge water samples.

In summary, the major conclusions and contributions of this thesis study are:

1) Chapter 3 tackled the limitation of MRM method optimization by first proposing a convertible equation for collision energy across different LCMS/MS instruments. The chemical coverage of conventional targeted analysis was broadened by the integration of computational prediction of retention time and HRMS detection into the targeted analysis workflow.
2) Chapter 4 tackled the inherent limitation in detection coverage of targeted analysis by proposing a database of pseudo ion pairs for selected ion monitoring mode (SIM) or multiple reaction monitoring (MRM) for GC/LCMS. Using this one-stop platform, we can complete the suspected screening analyses of tens to hundreds of exposomic chemicals within a short period.
3) Chapter 5 tackled the limited detection coverage of exogenous and endogenous metabolites in human exposomes by proposing an isotope labeling workflow based on the pseudo-MRM databases.
4) Chapter 6 tackled the limitations in non-targeted and semi-targeted analysis in human exposome research, beyond the targeted analysis. The heavy data deconvolutions are simplified by our proposed integrated software platforms.
5) Overall, this thesis contributes to the current exposome research by extending the chemical detection coverage and enhancing the detection sensitivity in

targeted analysis. Beyond the targeted analysis, data deconvolutions in the non-targeted analysis are simplified by our proposed integrated software platforms.

7.2 Recommendation for future exposome study

Future exposome characterization studies need to develop standardized high-throughput measurement platforms. We need to develop biological and chemical methods to detect xenobiotics and their metabolites in low abundance. We can explore existing resources and databases to develop prior information on exposures for accurate and efficient semi-targeted analysis. Integration of all available exposome databases, including MS/MS and risk-based chemical prioritization databases, to improve the accuracy of non-targeted screening. A computational tool such as a retention time prediction model based on quantitative structure-retention relationships (QSPR) can be further developed to assist compound identification in targeted, non-targeted, and semi-targeted analysis. Overall, we summarize our recommendations for future work as follows:

1) In this study, we have only covered the compounds with available MS/MS in the database. The pseudo-SIM and pseudo-MRM need to update regularly in the future with the expansion of the MS/MS library. There will be more data cleanup processes in the future.
2) In the risk-based NTA, we did not consider the abundance of the possible target. In the future, we can predict the possible concentration of compounds by using ionization efficiency prediction. The prioritization of chemicals with both concentration and toxicity will be more reasonable.
3) We haven't applied the method for the large population study. More validation studies can be conducted in the future. In addition, our current algorithm did not consider the normalization of MS signals due to instrumental sensitivity drift, which is essential for large sets of samples.
4) In the retention time alignment, we did not assume the high retention time shift and the current package might not work for this scenario. More refined prediction algorithms are required for future studies.
5) The MS/MS fragment reproducibility might not be consistent between different instruments. The pseudo-SIM and pseudo-MRM transitions from our databases might not be universally applicable on all platforms. Further

validation and improvement of the SIM/MRM *in silico* optimization algorithms are needed in the future.

REFERENCES

Andra, S. S. Austin, C. Patel, D. Dolios, G. Awawda, M. and Arora, M. (2017b). "Trends in the application of high-resolution mass spectrometry for human biomonitoring: An analytical primer to studying the environmental chemical space of the human exposome." Environment International, Vol. 100, pp. 32–61. https://doi.org/10.1016/J.ENVINT.2016.11.026

Anumol, T. and Snyder, S. A. (2015). "Rapid analysis of trace organic compounds in water by automated online solid-phase extraction coupled to liquid chromatography-tandem mass spectrometry." Talanta, Vol. 132, pp. 77–86. https://doi.org/10.1016/j.talanta.2014.08.011

Barupal, D. K. and Fiehn, O. (2019). "Generating the blood exposome database using a comprehensive text mining and database fusion approach." Environmental Health Perspectives, Vol. 127, No. 9. https://doi.org/10.1289/EHP4713

Battal, D. Cok, I. Unlusayin, I. and Tunctan, B. (2014). "Development and validation of an LC-MS/MS method for simultaneous quantitative analysis of free and conjugated bisphenol A in human urine." Biomedical Chromatography, Vol. 28, No. 5, pp. 686–693. https://doi.org/10.1002/bmc.3090

Bendik, J. Kalia, R. Sukumaran, J. Richardot, W. H. Hoh, E. and Kelley, S. T. (2021). "Automated high confidence compound identification of electron ionization mass spectra for nontargeted analysis." Journal of Chromatography A, Vol. 1660,. https://doi.org/10.1016/j.chroma.2021.462656

Beyer, B. A. Fang, M. Sadrian, B. Montenegro-Burke, J. R. Plaisted, W. C. Kok, B. P. C. Saez, E. Kondo, T. Siuzdak, G. and Lairson, L. L. (2018). "Metabolomics-based discovery of a metabolite that enhances oligodendrocyte maturation." Nature Chemical Biology, Vol. 14, No. 1, pp. 22–28.

Boleda, M. R. Galceran, M. T. and Ventura, F. (2013). "Validation and uncertainty estimation of a multiresidue method for pharmaceuticals in surface and treated waters by liquid chromatography-tandem mass spectrometry." Journal of Chromatography A, Vol. 1286, pp. 146–158. https://doi.org/10.1016/j.chroma.2013.02.077

Bonini, P. Kind, T. Tsugawa, H. Kumar Barupal, D. and Fiehn, O. (2020). "Retip: Retention Time Prediction for Compound Annotation in Untargeted

Metabolomics." Analytical Chemistry, Vol. 92, No. 11, pp. 7515–7522. https://doi.org/10.1021/acs.analchem.9b05765

Broecker, S. Herre, S. Wüst, B. Zweigenbaum, J. and Pragst, F. (2011). "Development and practical application of a library of CID accurate mass spectra of more than 2,500 toxic compounds for systematic toxicological analysis by LC–QTOF-MS with data-dependent acquisition." Analytical and Bioanalytical Chemistry, Vol. 400, No. 1, pp. 101–117. https://doi.org/10.1007/s00216-010-4450-9

Cai, S. S. Syage, J. A. Hanold, K. A. and Balogh, M. P. (2009). "Ultra performance liquid chromatography-atmospheric pressure photoionization-tandem mass spectrometry for high-sensitivity and high-throughput analysis of U.S. Environmental Protection Agency 16 priority pollutants polynuclear aromatic hydrocarbons." Analytical Chemistry, Vol. 81, No. 6, pp. 2123–2128. https://doi.org/10.1021/ac802275e

Cao, M. Fraser, K. Huege, J. Featonby, T. Rasmussen, S. and Jones, C. (2015a). "Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics." Metabolomics, Vol. 11, No. 3, pp. 696–706.

Chang, H. Wu, F. Jin, F. Feng, C. Zhao, X. and Liao, H. (2012). "Picogram per liter level determination of hydroxylated polybrominated diphenyl ethers in water by liquid chromatography-electrospray tandem mass spectrometry." Journal of Chromatography A, Vol. 1223, pp. 131–135. https://doi.org/10.1016/j.chroma.2011.12.075

Chen, D. Kannan, K. Tan, H. Zheng, Z. Feng, Y. L. Wu, Y. and Widelka, M. (2016). "Bisphenol Analogues Other Than BPA: Environmental Occurrence, Human Exposure, and Toxicity - A Review." In *Environmental Science and Technology* (Vol. 50, Issue 11, pp. 5438–5453). American Chemical Society. https://doi.org/10.1021/acs.est.5b05387

Chindarkar, N. S. Park, H.-D. Stone, J. A. and Fitzgerald, R. L. (2015). "Comparison of different time of flight-mass spectrometry modes for small molecule quantitative analysis." Journal of Analytical Toxicology, Vol. 39, No. 9, pp. 675–685.

Chung, M. K. Kannan, K. Louis, G. M. and Patel, C. J. (2018). “Toward Capturing the Exposome: Exposure Biomarker Variability and Coexposure Patterns in the Shared Environment.” Environmental Science & Technology, Vol. 52, No. 15, pp. 8801–8810. https://doi.org/10.1021/acs.est.8b01467

Chung, M. K. Rappaport, S. M. Wheelock, C. E. Nguyen, V. K. van der Meer, T. P. Miller, G. W. Vermeulen, R. and Patel, C. J. (2021). “Utilizing a biology-driven approach to map the exposome in health and disease: An essential investment to drive the next generation of environmental discovery.” In *Environmental Health Perspectives* (Vol. 129, Issue 8). Public Health Services, US Dept of Health and Human Services. https://doi.org/10.1289/EHP8327

Cosselman, K. E. Navas-Acien, A. and Kaufman, J. D. (2015). “Environmental factors in cardiovascular disease.” In *Nature Reviews Cardiology* (Vol. 12, Issue 11, pp. 627–642). Nature Publishing Group. https://doi.org/10.1038/nrcardio.2015.152

David, A. Chaker, J. Price, E. J. Bessonneau, V. Chetwynd, A. J. Vitale, C. M. Klánová, J. Walker, D. I. Antignac, J. P. Barouki, R. and Miller, G. W. (2021). “Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives.” Environment International, Vol. 156,. https://doi.org/10.1016/j.envint.2021.106630

Dennis, K. K. Marder, E. Balshaw, D. M. Cui, Y. Lynes, M. A. Patti, G. J. Rappaport, S. M. Shaughnessy, D. T. Vrijheid, M. and Barr, D. B. (2017). “Biomonitoring in the era of the exposome.” Environmental Health Perspectives, Vol. 125, No. 4, pp. 502–510. https://doi.org/10.1289/EHP474

Dhungana, S. Heywood, D. Goshawk, J. and Isaac, G. (n.d.). *De novo Discovery of Natural Products Using Progenesis QI and Natural Product Atlas Library*. http://www.npatlas.org/

Dodds, J. N. Alexander, N. L. M. Kirkwood, K. I. Foster, M. R. Hopkins, Z. R. Knappe, D. R. U. and Baker, E. S. (2021). “From Pesticides to Per- And Polyfluoroalkyl Substances: An Evaluation of Recent Targeted and Untargeted Mass Spectrometry Methods for Xenobiotics.” In *Analytical Chemistry* (Vol. 93, Issue 1, pp. 641–656). American Chemical Society. https://doi.org/10.1021/acs.analchem.0c04359

Domingo-Almenara, X. Montenegro-Burke, J. R. Ivanisevic, J. Thomas, A. Sidibé, J. Teav, T. Guijas, C. Aisporna, A. E. Rinehart, D. Hoang, L. Nordström, A. Gómez-Romero, M. Whiley, L. Lewis, M. R. Nicholson, J. K. Benton, H. P. and Siuzdak, G. (2018a). "XCMS-MRM and METLIN-MRM: a cloud library and public resource for targeted analysis of small molecules." Nature Methods, Vol. 15, No. 9, pp. 681–684. https://doi.org/10.1038/s41592-018-0110-3

Dong, M. Lih, T. S. M. Ao, M. Hu, Y. Chen, S. Y. Eguez, R. V. and Zhang, H. (2021). "Data-Independent Acquisition-Based Mass Spectrometry (DIA-MS) for Quantitative Analysis of Intact N-Linked Glycopeptides." Analytical Chemistry, Vol. 93, No. 41, pp. 13774–13782. https://doi.org/10.1021/acs.analchem.1c01659

Dong, T. Zhang, Y. Jia, S. Shang, H. Fang, W. Chen, D. and Fang, M. (2019). "Human Indoor Exposome of Chemicals in Dust and Risk Prioritization Using EPA's ToxCast Database." Environmental Science & Technology, Vol. 53, No. 12, pp. 7045–7054. https://doi.org/10.1021/acs.est.9b00280

Dresen, S. Ferreirós, N. Gnann, H. Zimmermann, R. and Weinmann, W. (2010). "Detection and identification of 700 drugs by multi-target screening with a 3200 Q TRAP® LC-MS/MS system and library searching." Analytical and Bioanalytical Chemistry, Vol. 396, No. 7, pp. 2425–2434. https://doi.org/10.1007/s00216-010-3485-2

Du, B. Li, Q. Pan, Z. Zhang, Y. Xie, R. Luo, D. and Zeng, L. (2021). "Improved LC-MS/MS Method for the Simultaneous Determination of Synthetic Phenol Antioxidants and Relevant Metabolites Making Use of Atmospheric Pressure Chemical Ionization and a Trap Column." Environmental Science and Technology Letters, Vol. 8, No. 3, pp. 256–262. https://doi.org/10.1021/acs.estlett.0c01013

Escher, B. I. Stapleton, H. M. and Schymanski, E. L. (2020). "Tracking complex mixtures of chemicals in our changing environment." Science, Vol. 367, No. 6476, pp. 388–392. https://doi.org/10.1126/science.aay6636

Fang, M. Hu, L. Chen, D. Guo, Y. Liu, J. Lan, C. Gong, J. and Wang, B. (2021a). "Exposome in human health: Utopia or wonderland?" In *The Innovation* (Vol. 2, Issue 4). Cell Press. https://doi.org/10.1016/j.xinn.2021.100172

129 Geens, T. Roosens, L. Neels, H. and Covaci, A. (2009). “Assessment of human
130 exposure to Bisphenol-A, Triclosan and Tetrabromobisphenol-A through indoor
131 dust intake in Belgium.” Chemosphere, Vol. 76, No. 6, pp. 755–760.
132 https://doi.org/10.1016/j.chemosphere.2009.05.024
133 González-Domínguez, R. Jáuregui, O. Queipo-Ortuño, M. I. and Andrés-Lacueva, C.
134 (2020). “Characterization of the Human Exposome by a Comprehensive and
135 Quantitative Large-Scale Multianalyte Metabolomics Platform.” Analytical
136 Chemistry, Vol. 92, No. 20, pp. 13767–13775.
137 https://doi.org/10.1021/acs.analchem.0c02008
138 Granelli, K. Elgerud, C. Lundström, Å. Ohlsson, A. and Sjöberg, P. (2009). “Rapid
139 multi-residue analysis of antibiotics in muscle by liquid chromatography-tandem
140 mass spectrometry.” Analytica Chimica Acta, Vol. 637, No. 1–2, pp. 87–91.
141 https://doi.org/10.1016/j.aca.2008.08.025
142 Grashow, R. Bessonneau, V. Gerona, R. R. Wang, A. Trowbridge, J. Lin, T. Buren, H.
143 Rudel, R. A. and Morello-Frosch, R. (2020). “Integrating Exposure Knowledge
144 and Serum Suspect Screening as a New Approach to Biomonitoring: An
145 Application in Firefighters and Office Workers.” Environmental Science and
146 Technology, Vol. 54, No. 7, pp. 4344–4355.
147 https://doi.org/10.1021/acs.est.9b04579
148 Guha, R. (2007). “Chemical Informatics Functionality in R.” Journal of Statistical
149 Software, Vol. 18, No. 5, pp. 1–16. https://doi.org/10.18637/jss.v018.i05
150 Guo, N. Peng, C. Y. Zhu, Q. F. Yuan, B. F. and Feng, Y. Q. (2017). “Profiling of
151 carbonyl compounds in serum by stable isotope labeling - Double precursor ion
152 scan - Mass spectrometry analysis.” Analytica Chimica Acta, Vol. 967, pp. 42–
153 51. https://doi.org/10.1016/j.aca.2017.03.006
154 Hagiwara, T. Saito, S. Ujiie, Y. Imai, K. Kakuta, M. Kadota, K. Terada, T. Sumikoshi,
155 K. Shimizu, K. and Nishi, T. (2010). “HPLC Retention time prediction for
156 metabolome analysi.” Bioinformation, Vol. 5, No. 6, pp. 255.
157 Helfer, A. G. Michely, J. A. Weber, A. A. Meyer, M. R. and Maurer, H. H. (2015).
158 “Orbitrap technology for comprehensive metabolite-based liquid
159 chromatographic–high resolution-tandem mass spectrometric urine drug
160 screening – Exemplified for cardiovascular drugs.” Analytica Chimica Acta, Vol.
161 891, pp. 221–233. https://doi.org/10.1016/j.aca.2015.08.018

Hernández, F. Ibáñez, M. Gracia-Lor, E. and Sancho, J. v. (2011). “Retrospective LC-QTOF-MS analysis searching for pharmaceutical metabolites in urban wastewater.” Journal of Separation Science, Vol. 34, No. 24, pp. 3517–3526. https://doi.org/10.1002/jssc.201100540

Hernández, F. Portolés, T. Pitarch, E. and López, F. J. (2009). “Searching for anthropogenic contaminants in human breast adipose tissues using gas chromatography-time-of-flight mass spectrometry.” Journal of Mass Spectrometry, Vol. 44, No. 1, pp. 1–11. https://doi.org/10.1002/jms.1538

Hindorff, L. A. Sethupathy, P. Junkins, H. A. Ramos, E. M. Mehta, J. P. Collins, F. S. and Manolio, T. A. (2009). “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.” Proceedings of the National Academy of Sciences of the United States of America, Vol. 106, No. 23, pp. 9362–9367. https://doi.org/10.1073/pnas.0903103106

Hollender, J. Schymanski, E. L. Singer, H. P. and Ferguson, P. L. (2017a). “Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?” Environmental Science and Technology, Vol. 51, No. 20, pp. 11505–11512. https://doi.org/10.1021/acs.est.7b02184

Hong, P. and Mcconville, P. R. (2016). “Dwell Volume and Extra-Column Volume : What Are They and How Do They Impact Method Transfer.” White Paper, *Figure 1*, pp. 1–9. https://doi.org/10.1093/beheco/arn149

Horning, W. B. and Weber, C. I. (1985). *Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms*.

Hu, X. Walker, D. I. Liang, Y. Smith, M. R. Orr, M. L. Juran, B. D. Ma, C. Uppal, K. Koval, M. Martin, G. S. Neujahr, D. C. Marsit, C. J. Go, Y. M. Pennell, K. D. Miller, G. W. Lazaridis, K. N. and Jones, D. P. (2021). “A scalable workflow to characterize the human exposome.” Nature Communications, Vol. 12, No. 1. https://doi.org/10.1038/s41467-021-25840-9

Imma Ferrer, E. M. T. (2008). “EPA Method 1694 : Agilent ’ s 6410A LC / MS / MS Solution for Pharmaceuticals and Personal Care Products in Water , Soil , Sediment , and Biosolids by HPLC / MS / MS Application Note.” Group, pp. 12.

Jakimska, A. Huerta, B. Bargańska, zaneta Kot-Wasik, A. Rodríguez-Mozaz, S. and Barceló, D. (2013). “Development of a liquid chromatography-tandem mass spectrometry procedure for determination of endocrine disrupting compounds in

fish from Mediterranean rivers." Journal of Chromatography A, Vol. 1306, pp. 44–58. https://doi.org/10.1016/j.chroma.2013.07.050

Jia, S. Xu, T. Huan, T. Chong, M. Liu, M. Fang, W. and Fang, M. (2019). "Chemical Isotope Labeling Exposome (CIL-EXPOSOME): One High-Throughput Platform for Human Urinary Global Exposome Characterization." Environmental Science and Technology, Vol. 53, No. 9, pp. 5445–5453. https://doi.org/10.1021/acs.est.9b00285

Jiang, C. Wang, X. Li, X. Inlora, J. Wang, T. Liu, Q. and Snyder, M. (2018). "Dynamic Human Environmental Exposome Revealed by Longitudinal Personal Monitoring." Cell, Vol. 175, No. 1, pp. 277- 291.e31. https://doi.org/10.1016/j.cell.2018.08.060

Jones, D. P. (2016). "Sequencing the exposome: A call to action." Toxicology Reports, Vol. 3, pp. 29–45. https://doi.org/10.1016/j.toxrep.2015.11.009

Kaliszan, R. (2007). "QSRR: Quantitative Structure-(Chromatographic) Retention Relationships." Chemical Reviews, Vol. 107, No. 7, pp. 3212–3246. https://doi.org/10.1021/cr068412z

Kavlock, R. Chandler, K. Houck, K. Hunter, S. Judson, R. Kleinstreuer, N. Knudsen, T. Martin, M. Padilla, S. Reif, D. Richard, A. Rotroff, D. Sipes, N. and Dix, D. (2012). "Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management." Chemical Research in Toxicology, Vol. 25, No. 7, pp. 1287–1302. https://doi.org/10.1021/tx3000939

Kern, S. Fenner, K. Singer, H. P. Schwarzenbach, R. P. and Hollender, J. (2009). "Identification of Transformation Products of Organic Contaminants in Natural Waters by Computer-Aided Prediction and High-Resolution Mass Spectrometry." Environmental Science & Technology, Vol. 43, No. 18, pp. 7039–7046. https://doi.org/10.1021/es901979h

Klein, D. R. Flannelly, D. F. and Schultz, M. M. (2010). "Quantitative determination of triclocarban in wastewater effluent by stir bar sorptive extraction and liquid desorption-liquid chromatography-tandem mass spectrometry." Journal of Chromatography A, Vol. 1217, No. 11, pp. 1742–1747. https://doi.org/10.1016/j.chroma.2010.01.028

Kolpin, D. W. Furlong, E. T. Meyer, M. T. Thurman, E. M. Zaugg, S. D. Barber, L. B. and Buxton, H. T. (2002). "Pharmaceuticals, hormones, and other organic

wastewater contaminants in US streams, 1999− 2000: A national reconnaissance." Environmental Science & Technology, Vol. 36, No. 6, pp. 1202–1211.

Lichtenstein, P. Holm, N. v. Verkasalo, P. K. Iliadou, A. Kaprio, J. Koskenvuo, M. Pukkala, E. Skytthe, A. and Hemminki, K. (2000). "Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland." New England Journal of Medicine, Vol. 343, No. 2, pp. 78–85. https://doi.org/10.1056/NEJM200007133430201

Liu, M. Jia, S. Dong, T. Han, Y. Xue, J. Wanjaya, E. R. and Fang, M. (2019). "The occurrence of bisphenol plasticizers in paired dust and urine samples and its association with oxidative stress." Chemosphere, Vol. 216, pp. 472–478.

Liu, P. Huang, Y. Q. Cai, W. J. Yuan, B. F. and Feng, Y. Q. (2014). "Profiling of thiol-containing compounds by stable isotope labeling double precursor ion scan mass spectrometry." Analytical Chemistry, Vol. 86, No. 19, pp. 9765–9773. https://doi.org/10.1021/ac5023315

Luo, X. An, M. Cuneo, K. C. Lubman, D. M. and Li, L. (2018). "High-Performance Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry for Exosome Metabolomics." Analytical Chemistry, Vol. 90, No. 14, pp. 8314–8319. https://doi.org/10.1021/acs.analchem.8b01726

Maitre, L. Bustamante, M. Hernández-Ferrer, C. Thiel, D. Lau, C.-H. Siskos, A. Vives-Usano, M. Ruiz-Arenas, C. Robinson, O. Mason, D. Wright, J. Cadiou, S. Slama, R. Heude, B. Gallego-Paüls, M. Casas, M. Sunyer, J. Papadopoulou, E. Z. Gutzkow, K. B. … Vrijheid, M. (n.d.). *Multi-omics signatures of the human early life exposome*. https://doi.org/10.1101/2021.05.04.21256605

Manzetti, S. van der Spoel, E. R. and van der Spoel, D. (2014). "Chemical properties, environmental fate, and degradation of seven classes of pollutants." In *Chemical Research in Toxicology* (Vol. 27, Issue 5, pp. 713–737). American Chemical Society. https://doi.org/10.1021/tx500014w

Martin and Todd. (2020a). *User's Guide for T. E. S. T. (Toxicity Estimation Software Tool) Version 5.1 A Java Application to Estimate Toxicities and Physical Properties from Molecular Structure*. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

Marvel, S. W. To, K. Grimm, F. A. Wright, F. A. Rusyn, I. and Reif, D. M. (2018). "ToxPi Graphical User Interface 2.0: Dynamic exploration, visualization, and

sharing of integrated data models." BMC Bioinformatics, Vol. 19, No. 1, pp. 80. https://doi.org/10.1186/s12859-018-2089-2

Meekel, N. Vughs, D. Béen, F. and Brunner, A. M. (2021). "Online Prioritization of Toxic Compounds in Water Samples through Intelligent HRMS Data Acquisition." Analytical Chemistry, Vol. 93, No. 12, pp. 5071–5080. https://doi.org/10.1021/acs.analchem.0c04473

Melnikov, A. D. Tsentalovich, Y. P. and Yanshole, V. v. (2020). "Deep Learning for the Precise Peak Detection in High-Resolution LC-MS Data." Analytical Chemistry, Vol. 92, No. 1, pp. 588–592. https://doi.org/10.1021/acs.analchem.9b04811

Miller, G. W. and Jones, D. P. (2014). "The nature of nurture: Refining the definition of the exposome." Toxicological Sciences, Vol. 137, No. 1, pp. 1–2. https://doi.org/10.1093/toxsci/kft251

Miossec, C. Lanceleur, L. and Monperrus, M. (2019). "Multi-residue analysis of 44 pharmaceutical compounds in environmental water samples by solid-phase extraction coupled to liquid chromatography-tandem mass spectrometry." Journal of Separation Science, Vol. 42, No. 10, pp. 1853–1866. https://doi.org/10.1002/jssc.201801214

Moriwaki, H. Tian, Y. S. Kawashita, N. and Takagi, T. (2018). "Mordred: A molecular descriptor calculator." Journal of Cheminformatics, Vol. 10, No. 1. https://doi.org/10.1186/s13321-018-0258-y

Mueller, C. A. Weinmann, W. Dresen, S. Schreiber, A. and Gergov, M. (2005). "Development of a multi-target screening analysis for 301 drugs using a QTrap liquid chromatography/tandem mass spectrometry system and automated library searching." Rapid Communications in Mass Spectrometry, Vol. 19, No. 10, pp. 1332–1338. https://doi.org/10.1002/rcm.1934

Naegele, E. (2013). "Detection of trace level pharmaceuticals in drinking water by online SPE Enrichment with the Agilent 1200 Infinity Series Online-SPE solution, Agilent Technologies, Application Note." Inc. Application.

Neveu, V. Moussy, A. Rouaix, H. Wedekind, R. Pon, A. Knox, C. Wishart, D. S. and Scalbert, A. (2017). "Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors." Nucleic Acids Research, Vol. 45, No. D1, pp. D979–D984. https://doi.org/10.1093/nar/gkw980

294 Newton, S. R. McMahen, R. L. Sobus, J. R. Mansouri, K. Williams, A. J. McEachran,
295 A. D. and Strynar, M. J. (2018). “Suspect screening and non-targeted analysis of
296 drinking water using point-of-use filters.” Environmental Pollution, Vol. 234, pp.
297 297–306. https://doi.org/10.1016/J.ENVPOL.2017.11.033
298 Niedzwiecki, M. M. Walker, D. I. Vermeulen, R. Chadeau-Hyam, M. Jones, D. P. and
299 Miller, G. W. (2019). “The Exposome: Molecules to Populations.” Annual
300 Review of Pharmacology and Toxicology, Vol. 59, No. 1, pp. 107–127.
301 https://doi.org/10.1146/annurev-pharmtox-010818-021315
302 Pareja, L. Cesio, V. Heinzen, H. and Fernández-Alba, A. R. (2011). “Evaluation of
303 various QuEChERS based methods for the analysis of herbicides and other
304 commonly used pesticides in polished rice by LC-MS/MS.” Talanta, Vol. 83, No.
305 5, pp. 1613–1622. https://doi.org/10.1016/j.talanta.2010.11.052
306 Peng, J. and Li, L. (2013). “Liquid-liquid extraction combined with differential
307 isotope dimethylaminophenacyl labeling for improved metabolomic profiling of
308 organic acids.” Analytica Chimica Acta, Vol. 803, pp. 97–105.
309 https://doi.org/10.1016/j.aca.2013.07.045
310 Perera, F. P. (2017). “Multiple threats to child health from fossil fuel combustion:
311 Impacts of air pollution and climate change.” In *Environmental Health*
312 *Perspectives* (Vol. 125, Issue 2, pp. 141–148). Public Health Services, US Dept
313 of Health and Human Services. https://doi.org/10.1289/EHP299
314 Plumb, R. S. Johnson, K. A. Rainville, P. Smith, B. W. Wilson, I. D. Castro-Perez, J.
315 M. and Nicholson, J. K. (2006). “UPLC/MSE; a new approach for generating
316 molecular fragment information for biomarker structure elucidation.” Rapid
317 Communications in Mass Spectrometry, Vol. 20, No. 13, pp. 1989–1994.
318 https://doi.org/10.1002/rcm.2550
319 Provencher, G. Bérubé, R. Dumas, P. Bienvenu, J. F. Gaudreau, É. Bélanger, P. and
320 Ayotte, P. (2014a). “Determination of bisphenol A, triclosan and their
321 metabolites in human urine using isotope-dilution liquid chromatography-tandem
322 mass spectrometry.” Journal of Chromatography A, Vol. 1348, pp. 97–104.
323 https://doi.org/10.1016/j.chroma.2014.04.072
324 Qi, B. L. Liu, P. Wang, Q. Y. Cai, W. J. Yuan, B. F. and Feng, Y. Q. (2014).
325 “Derivatization for liquid chromatography-mass spectrometry.” In *TrAC* -

Trends in Analytical Chemistry (Vol. 59, pp. 121–132). Elsevier B.V. https://doi.org/10.1016/j.trac.2014.03.013

Rappaport, S. M. Barupal, D. K. Wishart, D. Vineis, P. and Scalbert, A. (2014). "The blood exposome and its role in discovering causes of disease." In *Environmental Health Perspectives* (Vol. 122, Issue 8, pp. 769–774). Public Health Services, US Dept of Health and Human Services. https://doi.org/10.1289/ehp.1308015

Rappaport, S. M. and Smith, M. T. (2010). "Environment and Disease Risks." Science, Vol. 330, No. 6003, pp. 460–461. https://doi.org/10.1126/science.1192603

Reif, D. M. Martin, M. T. Tan, S. W. Houck, K. A. Judson, R. S. Richard, A. M. Knudsen, T. B. Dix, D. J. and Kavlock, R. J. (2010). "Endocrine profiling and prioritization of environmental chemicals using ToxCast data." Environmental Health Perspectives, Vol. 118, No. 12, pp. 1714–1720. https://doi.org/10.1289/ehp.1002180

Reif, D. M. Sypa, M. Lock, E. F. Wright, F. A. Wilson, A. Cathey, T. Judson, R. R. and Rusyn, I. (2013). "ToxPi GUI: an interactive visualization tool for transparent integration of data from diverse sources of evidence." Bioinformatics (Oxford, England), Vol. 29, No. 3, pp. 402–403. https://doi.org/10.1093/bioinformatics/bts686

Ren, L. Fang, J. Liu, G. Zhang, J. Zhu, Z. Liu, H. Lin, K. Zhang, H. and Lu, S. (2016). "Simultaneous determination of urinary parabens, bisphenol A, triclosan, and 8-hydroxy-2′-deoxyguanosine by liquid chromatography coupled with electrospray ionization tandem mass spectrometry." Analytical and Bioanalytical Chemistry, Vol. 408, No. 10, pp. 2621–2629. https://doi.org/10.1007/s00216-016-9372-8

Robinson, O. Tamayo, I. de Castro, M. Valentin, A. Giorgis-Allemand, L. Krog, N. H. Aasvang, G. M. Ambros, A. Ballester, F. Bird, P. Chatzi, L. Cirach, M. Dėdelė, A. Donaire-Gonzalez, D. Gražuleviciene, R. Iakovidis, M. Ibarluzea, J. Kampouri, M. Lepeule, J. … Basagaña, X. (2018). "The urban exposome during pregnancy and its socioeconomic determinants." Environmental Health Perspectives, Vol. 126, No. 7. https://doi.org/10.1289/EHP2862

Roca, M. Leon, N. Pastor, A. and Yusà, V. (2014). "Comprehensive analytical strategy for biomonitoring of pesticides in urine by liquid chromatography–orbitrap high resolution mass spectrometry." Journal of Chromatography A, Vol. 1374, pp. 66–76. https://doi.org/10.1016/j.chroma.2014.11.010

359 Rodil, R. Quintana, J. B. López-Mahía, P. Muniategui-Lorenzo, S. and Prada-
360 Rodríguez, D. (2009). “Multi-residue analytical method for the determination of
361 emerging pollutants in water by solid-phase extraction and liquid
362 chromatography-tandem mass spectrometry.” Journal of Chromatography A, Vol.
363 1216, No. 14, pp. 2958–2969. https://doi.org/10.1016/j.chroma.2008.09.041
364 Rodil, R. Quintana, J. B. and Reemtsma, T. (2005). “Liquid chromatography− tandem
365 mass spectrometry determination of nonionic organophosphorus flame retardants
366 and plasticizers in wastewater samples.” Analytical Chemistry, Vol. 77, No. 10,
367 pp. 3083–3089.
368 Ruff, M. Mueller, M. S. Loos, M. and Singer, H. P. (2015). “Quantitative target and
369 systematic non-target analysis of polar organic micro-pollutants along the river
370 Rhine using high-resolution mass-spectrometry – Identification of unknown
371 sources and compounds.” Water Research, Vol. 87, pp. 145–154.
372 https://doi.org/10.1016/J.WATRES.2015.09.017
373 Schlittenbauer, L. Seiwert, B. and Reemtsma, T. (2016). “Ultrasound-assisted
374 hydrolysis of conjugated parabens in human urine and their determination by
375 UPLC–MS/MS and UPLC–HRMS.” Analytical and Bioanalytical Chemistry,
376 Vol. 408, No. 6, pp. 1573–1583.
377 Schymanski, E. L. Jeon, J. Gulde, R. Fenner, K. Ruff, M. Singer, H. P. and Hollender,
378 J. (2014). “Identifying small molecules via high resolution mass spectrometry:
379 Communicating confidence.” In *Environmental Science and Technology* (Vol. 48,
380 Issue 4, pp. 2097–2098). https://doi.org/10.1021/es5002105
381 Schymanski, E. L. Singer, H. P. Longrée, P. Loos, M. Ruff, M. Stravs, M. A. Ripollés
382 Vidal, C. and Hollender, J. (2014). “Strategies to characterize polar organic
383 contamination in wastewater: Exploring the capability of high resolution mass
384 spectrometry.” Environmental Science and Technology, Vol. 48, No. 3, pp.
385 1811–1818. https://doi.org/10.1021/es4044374
386 Takahashi, M. Izumi, Y. Iwahashi, F. Nakayama, Y. Iwakoshi, M. Nakao, M. Yamato,
387 S. Fukusaki, E. and Bamba, T. (2018). “Highly Accurate Detection and
388 Identification Methodology of Xenobiotic Metabolites Using Stable Isotope
389 Labeling, Data Mining Techniques, and Time-Dependent Profiling Based on
390 LC/HRMS/MS.” Analytical Chemistry, Vol. 90, No. 15, pp. 9068–9076.
391 https://doi.org/10.1021/acs.analchem.8b01388

392 Tomasetti, C. Li, L. and Vogelstein, B. (2017). *CANCER ETIOLOGY Stem cell*
393 *divisions, somatic mutations, cancer etiology, and cancer prevention*.
394 https://www.science.org
395 Tshala-Katumbay, D. Mwanza, J. C. Rohlman, D. S. Maestre, G. and Oria, R. B.
396 (2015). "A global perspective on the influence of environmental exposures on
397 the nervous system." In *Nature* (Vol. 527, Issue 7578, pp. S187–S192). Nature
398 Publishing Group. https://doi.org/10.1038/nature16034
399 Ulrich, E. M. Sobus, J. R. Grulke, C. M. Richard, A. M. Newton, S. R. Strynar, M. J.
400 Mansouri, K. and Williams, A. J. (2019). "EPA's non-targeted analysis
401 collaborative trial (ENTACT): genesis, design, and initial findings." Analytical
402 and Bioanalytical Chemistry, Vol. 411, No. 4, pp. 853–866.
403 https://doi.org/10.1007/s00216-018-1435-6
404 Vela-Soria, F. Ballesteros, O. Zafra-Gómez, A. Ballesteros, L. and Navalón, A.
405 (2014). "UHPLC-MS/MS method for the determination of bisphenol A and its
406 chlorinated derivatives, bisphenol S, parabens, and benzophenones in human
407 urine samples." Analytical and Bioanalytical Chemistry, Vol. 406, No. 15, pp.
408 3773–3785. https://doi.org/10.1007/s00216-014-7785-9
409 Vermeulen, R. Schymanski, E. L. Barabási, A. L. and Miller, G. W. (2020b). "The
410 exposome and health: Where chemistry meets biology." Science, Vol. 367, No.
411 6476, pp. 392–396. https://doi.org/10.1126/science.aay3164
412 Vineis, P. (2004). "A self-fulfilling prophecy: Are we underestimating the role of the
413 environment in gene-environment interaction research?" In *International Journal*
414 *of Epidemiology* (Vol. 33, Issue 5, pp. 945–946).
415 https://doi.org/10.1093/ije/dyh277
416 Walker, A. v. (2013). "Secondary Ion Mass Spectrometry☆." In *Reference Module in*
417 *Chemistry, Molecular Sciences and Chemical Engineering*. Elsevier.
418 https://doi.org/https://doi.org/10.1016/B978-0-12-409547-2.05228-8
419 Wang, S. Ang, H. M. and Tade, M. O. (2007). "Volatile organic compounds in indoor
420 environment and photocatalytic oxidation: State of the art." Environment
421 International, Vol. 33, No. 5, pp. 694–705.
422 https://doi.org/10.1016/J.ENVINT.2007.02.011
423 Wang, Z. Klipfell, E. Bennett, B. J. Koeth, R. Levison, B. S. Dugar, B. Feldstein, A. E.
424 Britt, E. B. Fu, X. Chung, Y. M. Wu, Y. Schauer, P. Smith, J. D. Allayee, H.

Tang, W. H. W. Didonato, J. A. Lusis, A. J. and Hazen, S. L. (2011). “Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease.” Nature, Vol. 472, No. 7341, pp. 57–65. https://doi.org/10.1038/nature09922

Warth, B. Raffeiner, P. Granados, A. Huan, T. Fang, M. Forsberg, E. M. Benton, H. P. Goetz, L. Johnson, C. H. and Siuzdak, G. (2018). “Metabolomics reveals that dietary xenoestrogens alter cellular metabolism induced by palbociclib/letrozole combination cancer therapy.” Cell Chemical Biology, Vol. 25, No. 3, pp. 291–300.

Wild, C. P. (2005). “Complementing the genome with an ‘exposome’: The outstanding challenge of environmental exposure measurement in molecular epidemiology.” In *Cancer Epidemiology Biomarkers and Prevention* (Vol. 14, Issue 8, pp. 1847–1850). https://doi.org/10.1158/1055-9965.EPI-05-0456

Wild, C. P. (2012). “The exposome: From concept to utility.” In *International Journal of Epidemiology* (Vol. 41, Issue 1, pp. 24–32). https://doi.org/10.1093/ije/dyr236

Willett, W. C. (2002). “Balancing Life-Style and Genomics Research for Disease Prevention.” Science, Vol. 296, No. 5568, pp. 695–698. https://doi.org/10.1126/science.1071055

Williams, A. J. Lambert, J. C. Thayer, K. and Dorne, J. L. C. M. (2021). “Sourcing data on chemical properties and hazard data from the US-EPA CompTox Chemicals Dashboard: A practical guide for human risk assessment.” In *Environment International* (Vol. 154). Elsevier Ltd. https://doi.org/10.1016/j.envint.2021.106566

Winter, S. V. Meier, F. Wichmann, C. Cox, J. Mann, M. and Meissner, F. (2018). “EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification.” Nature Methods, Vol. 15, No. 7, pp. 527–530. https://doi.org/10.1038/s41592-018-0037-8

Woodruff, T. J. Zota, A. R. and Schwartz, J. M. (2011). “Environmental chemicals in pregnant women in the united states: NHANES 2003- 2004.” Environmental Health Perspectives, Vol. 119, No. 6, pp. 878–885. https://doi.org/10.1289/ehp.1002727

Xu, Z. Jiang, T. Xu, Q. Zhai, Y. Li, D. and Xu, W. (2019). “Pseudo-Multiple Reaction Monitoring (Pseudo-MRM) Mode on the ‘brick’ Mass Spectrometer, Using the

Grid-SWIFT Waveform." Analytical Chemistry, Vol. 91, No. 21, pp. 13838–13846. https://doi.org/10.1021/acs.analchem.9b03315

Xue, J. Derks, R. J. E. Webb, B. Billings, E. M. Aisporna, A. Giera, M. and Siuzdak, G. (2021). "Single Quadrupole Multiple Fragment Ion Monitoring Quantitative Mass Spectrometry." Analytical Chemistry, Vol. 93, No. 31, pp. 10879–10889. https://doi.org/10.1021/acs.analchem.1c01246

Xue, J. Lai, Y. Liu, C. W. and Ru, H. (2019). "Towards mass spectrometry-based chemical exposome: Current approaches, challenges, and future directions." In *Toxics* (Vol. 7, Issue 3). MDPI AG. https://doi.org/10.3390/toxics7030041

Yang, J. J. Han, Y. Mah, C. H. Wanjaya, E. Peng, B. Xu, T. F. Liu, M. Huan, T. and Fang, M. L. (2020). "Streamlined MRM method transfer between instruments assisted with HRMS matching and retention-time prediction." Analytica Chimica Acta, Vol. 1100, pp. 88–96. https://doi.org/10.1016/j.aca.2019.12.002

Yap, C. W. (2011). "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints." Journal of Computational Chemistry, Vol. 32, No. 7, pp. 1466–1474.

Zdravković, M. Antović, A. Veselinović, J. B. Sokolović, D. and Veselinović, A. M. (2018). "QSPR in forensic analysis – The prediction of retention time of pesticide residues based on the Monte Carlo method." Talanta, Vol. 178, pp. 656–662. https://doi.org/10.1016/J.TALANTA.2017.09.064

Zhang, J. Shen, H. Xu, W. Xia, Y. Boyd Barr, D. Mu, X. Wang, X. Liu, L. Huang, Q. and Tian, M. (2014). "Urinary Metabolomics Revealed Arsenic Internal Dose-Related Metabolic Alterations: A Proof-of-Concept Study in a Chinese Male Cohort." Environmental Science & Technology, Vol. 48, No. 20, pp. 12265–12274. https://doi.org/10.1021/es503659w

Zhao, F. Li, L. Chen, Y. Huang, Y. Keerthisinghe, T. P. Chow, A. Dong, T. Jia, S. Xing, S. Warth, B. Huan, T. and Fang, M. (2021a). "Risk-based chemical ranking and generating a prioritized human exposome database." Environmental Health Perspectives, Vol. 129, No. 4. https://doi.org/10.1289/EHP7722

Zhao, F. Li, L. Chen, Y. Huang, Y. Keerthisinghe, T. P. Chow, A. Dong, T. Jia, S. Xing, S. Warth, B. Huan, T. and Fang, M. (2021b). "Risk-based chemical ranking and generating a prioritized human exposome database." Environmental Health Perspectives, Vol. 129, No. 4. https://doi.org/10.1289/EHP7722

Zhao, R. S. Wang, X. Sun, J. Hu, C. and Wang, X. K. (2011). "Determination of triclosan and triclocarban in environmental water samples with ionic liquid/ionic liquid dispersive liquid-liquid microextraction prior to HPLC-ESI-MS/MS." Microchimica Acta, Vol. 174, No. 1, pp. 145–151. https://doi.org/10.1007/s00604-011-0607-2

Zheng, F. Zhao, X. Zeng, Z. Wang, L. Lv, W. Wang, Q. and Xu, G. (2020). "Development of a plasma pseudotargeted metabolomics method based on ultra-high-performance liquid chromatography–mass spectrometry." Nature Protocols, Vol. 15, No. 8, pp. 2519–2537. https://doi.org/10.1038/s41596-020-0341-5

Zheng, J. Gong, G. G. Zheng, S. J. Zhang, Y. and Feng, Y. Q. (2021). "High coverage profiling of carboxylated metabolites in HepG2 cells using secondary amine-assisted ultrahigh-performance liquid chromatography coupled to high-resolution mass spectrometry." Analytical Chemistry, Vol. 93, No. 3, pp. 1604–1611. https://doi.org/10.1021/acs.analchem.0c04048

Zisi, C. Sampsonidis, I. Fasoula, S. Papachristos, K. Witting, M. Gika, H. G. Nikitas, P. and Pappa-Louisi, A. (2017). "QSRR modeling for metabolite standards analyzed by two different chromatographic columns using multiple linear regression." Metabolites, Vol. 7, No. 1, pp. 7.

APPENDIX A

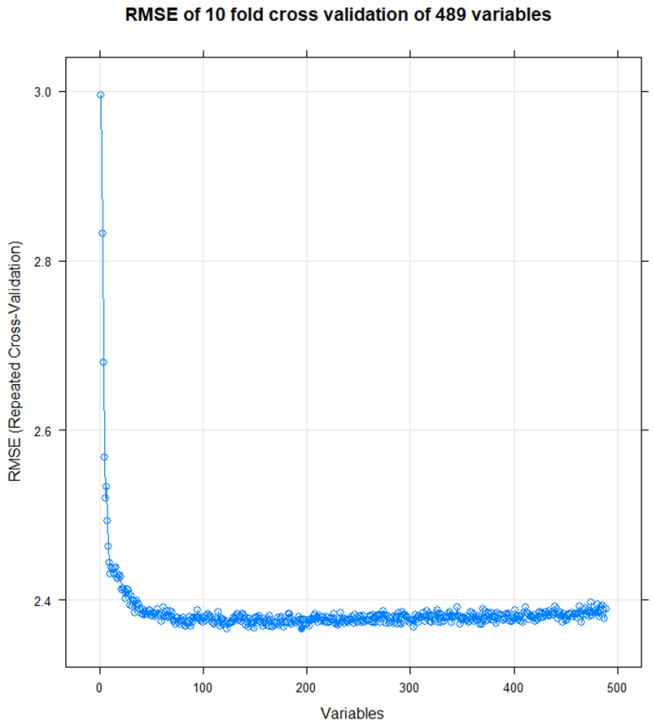


Figure S3.1. RMSE of 10-fold cross-valication by random forest for 489 variables

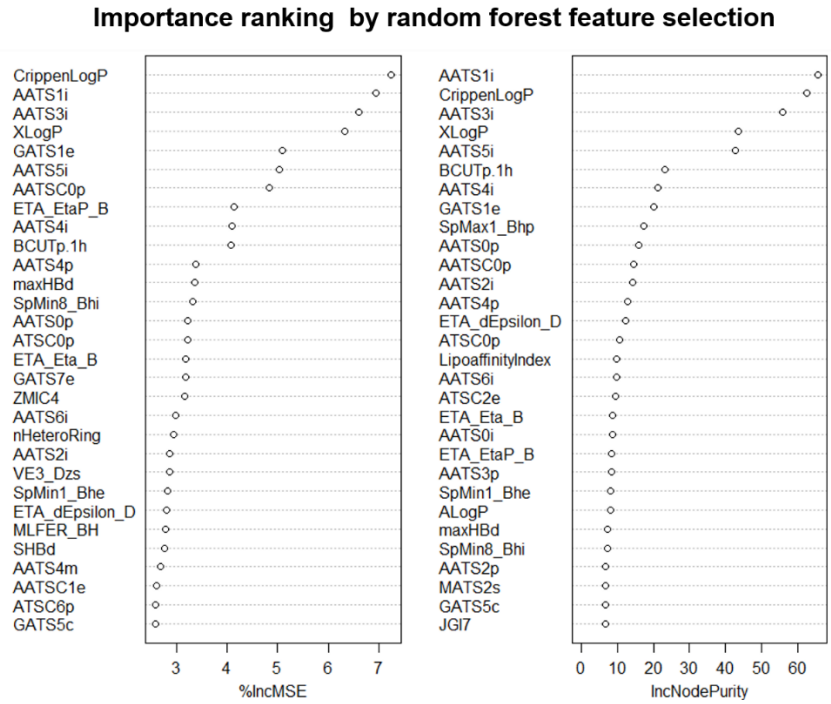
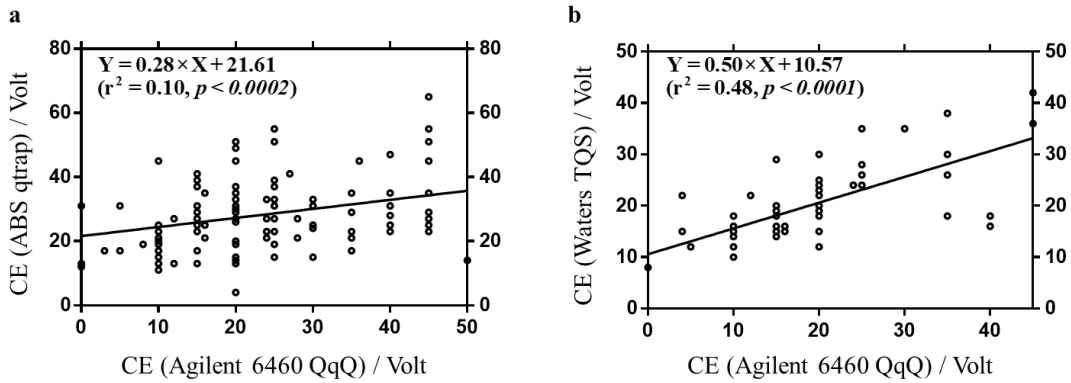
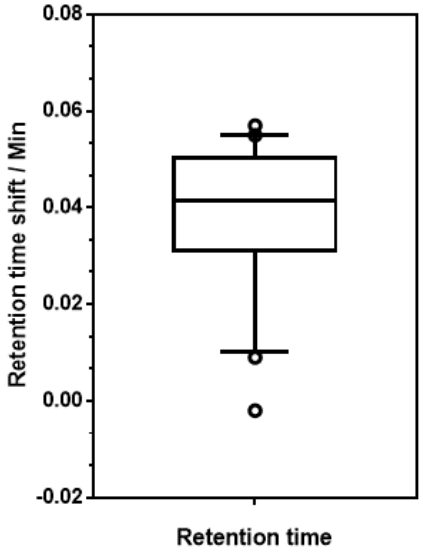


Figure S3.2. Variable's importance ranking by random forest feature selection.



514
515 **Figure S3.3**. Correlation analysis of collision energy (CE) values adopted across
516 various platforms. a) Linear regression analysis of CE performed on Agilent 6460
517 QqQ and API qtrap (Pearson r^2 = 0.31, n = 136); b) Linear regression analysis of CE
518 performed on Agilent 6460 QqQ and Waters TQ-S (Pearson r^2 = 0.69, n = 62).
519



520
521 **Figure S3.4.** Box plot of retention time shift of 20 compounds between LC-MS/MS
522 and HRMS system.

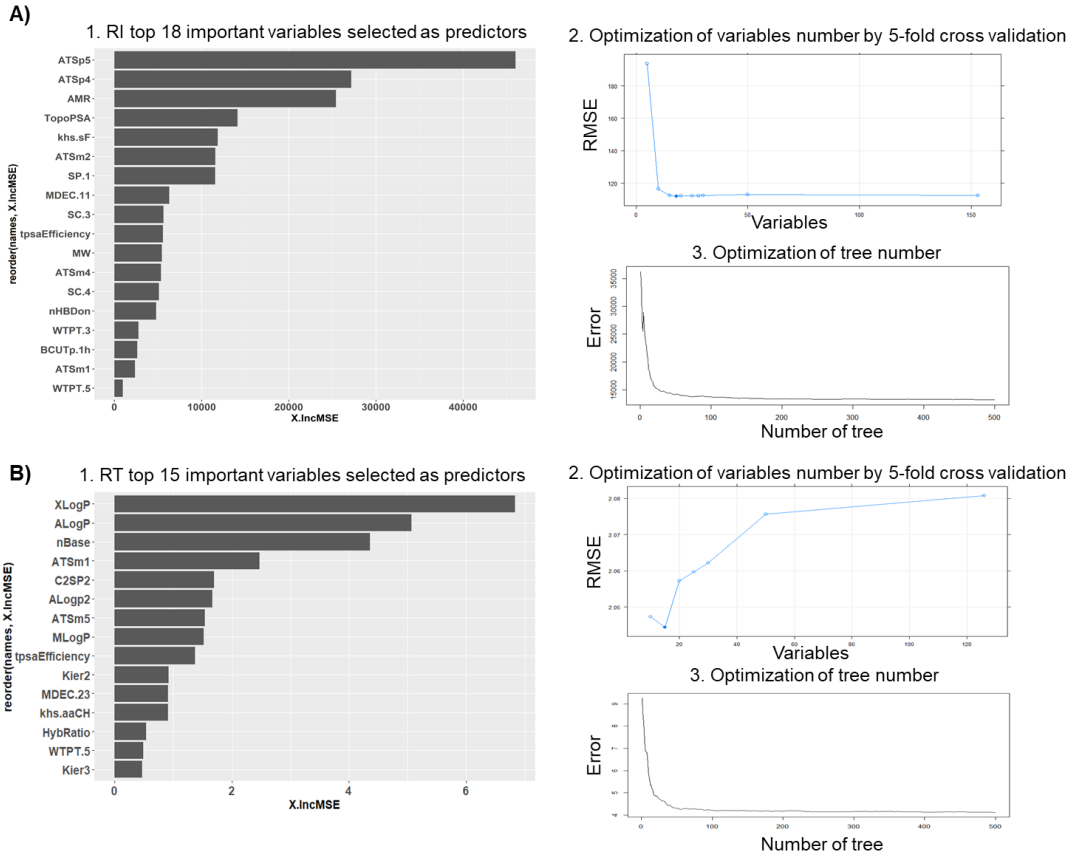
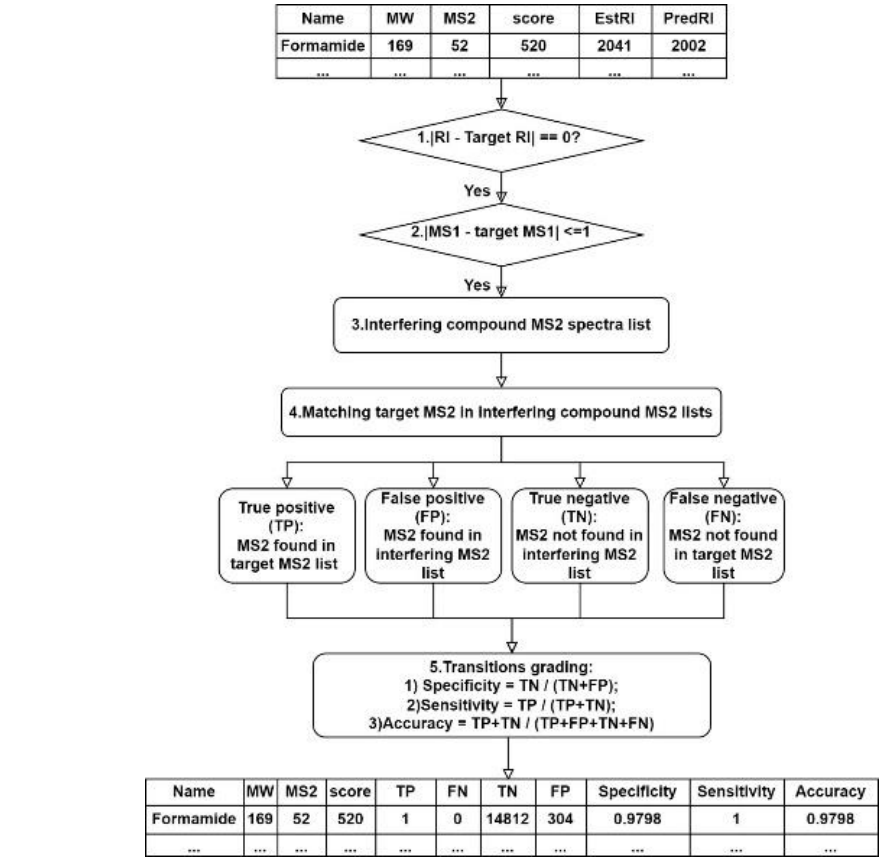
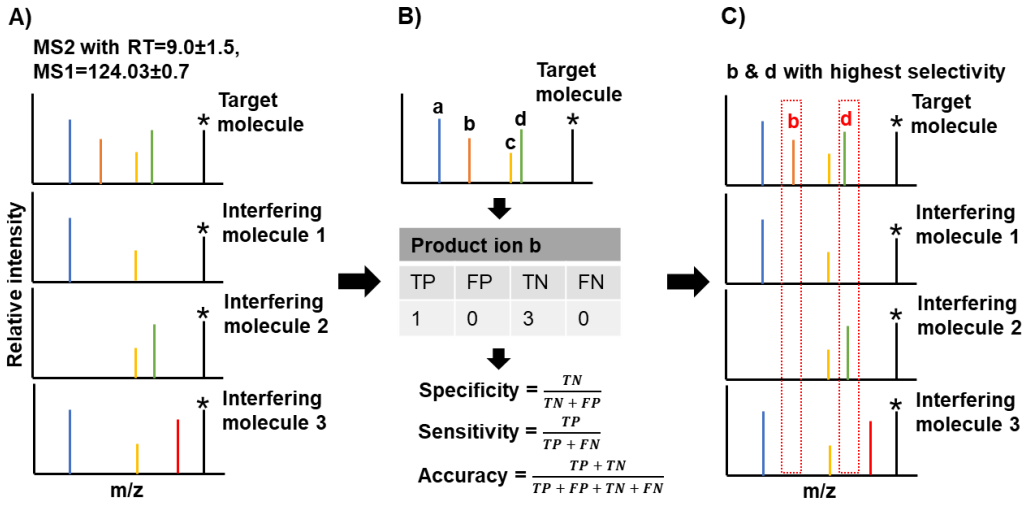


Figure S4.1. parameters optimization for retention prediction model



526

527 **Figure S4.2.** MS2 spectra optimization for pseudo-SIM transitions



528

529 **Figure S4.3.** A graphic demonstration for MS2 spectra optimization of pseudo-MRM
530 transitions

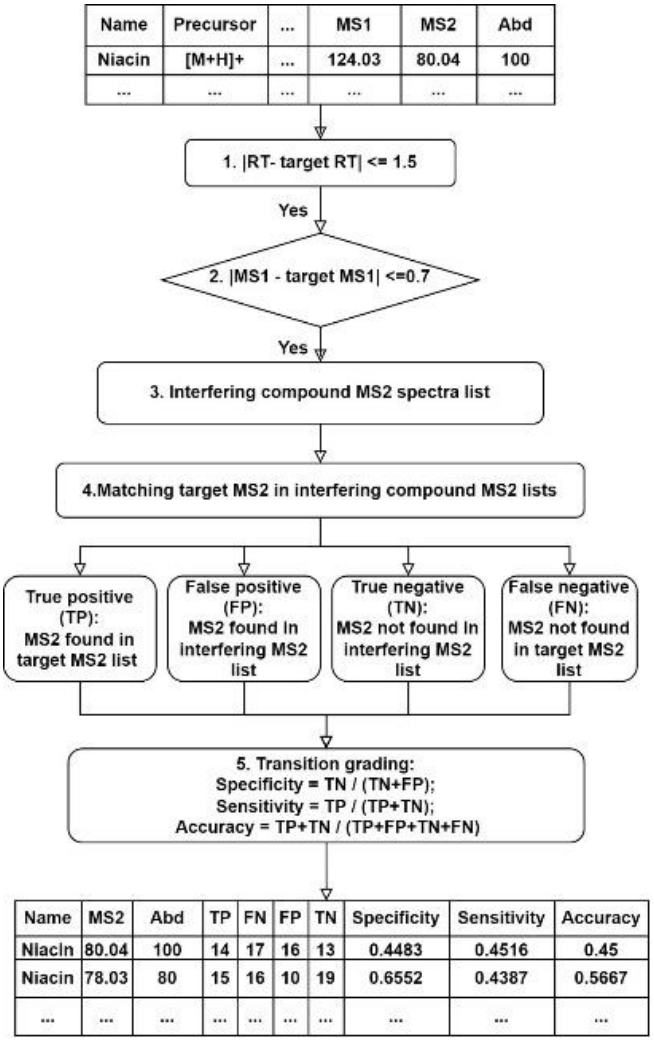


Figure S4.4. MS2 spectra optimization for pseudo-MRM transitions

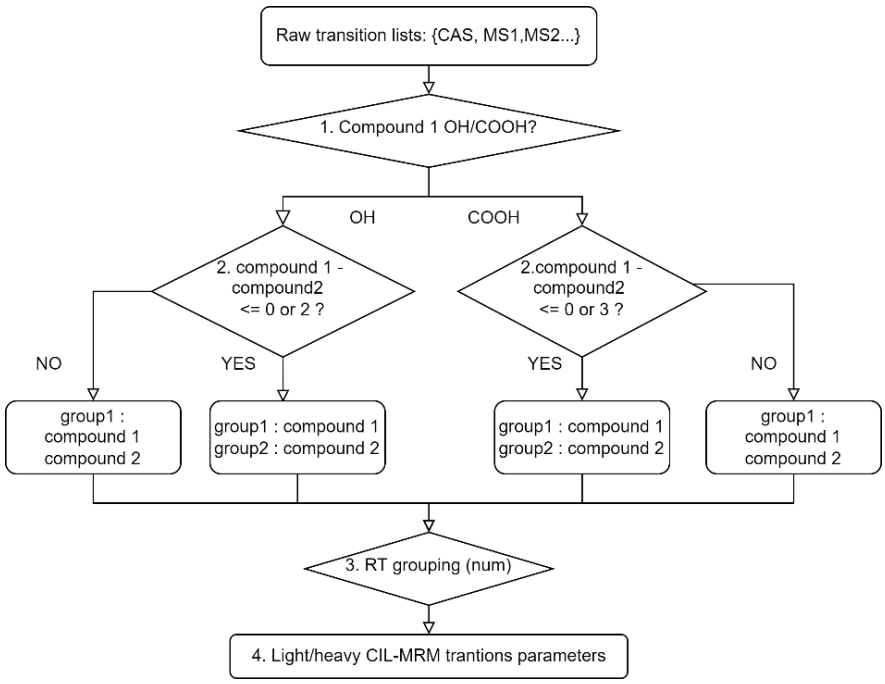


Figure S5.1. Algorithm of MRM transition grouping function in CILMRM

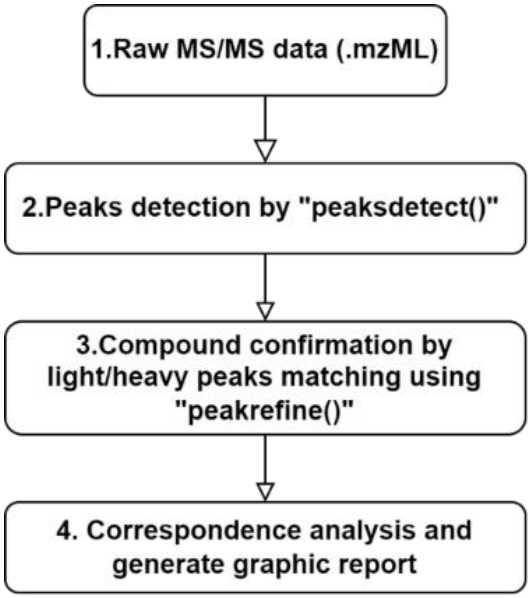


Figure S5.2. The CILMRM R package workflow

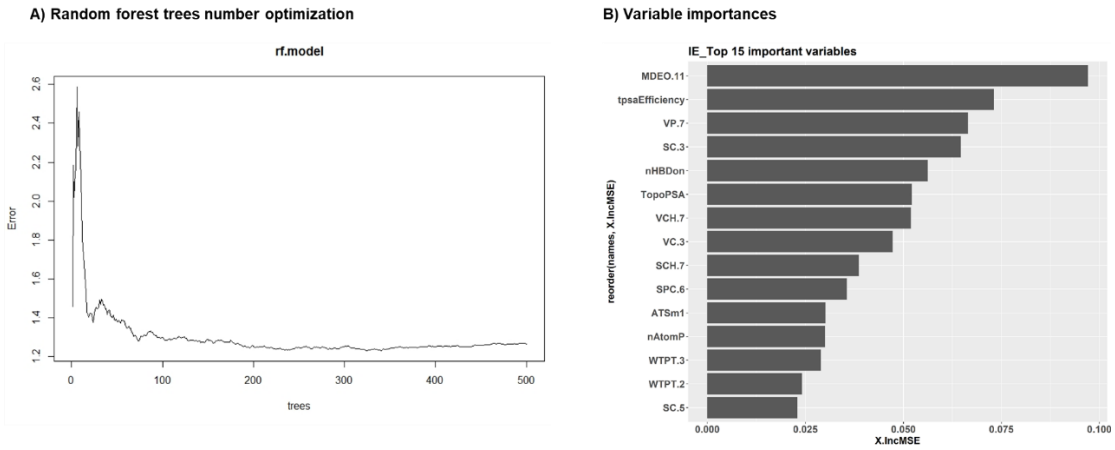


Figure S6.1. Random Forest models tree number optimization for RT and ESI ionization efficiency

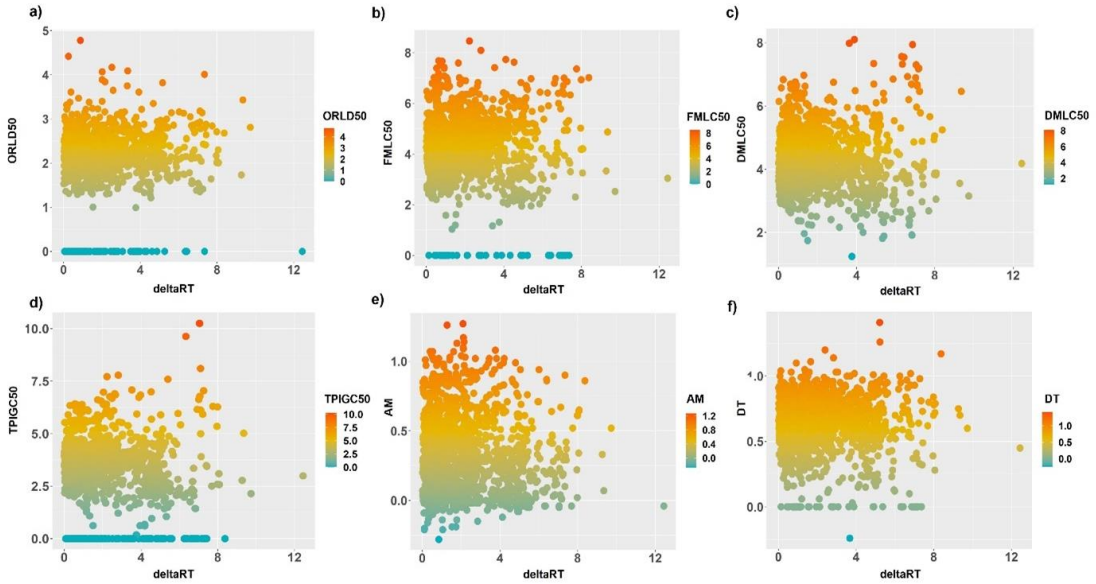


Figure S6.2. Summary plot of six toxicities endpoints versus deltaRT for all candidates against

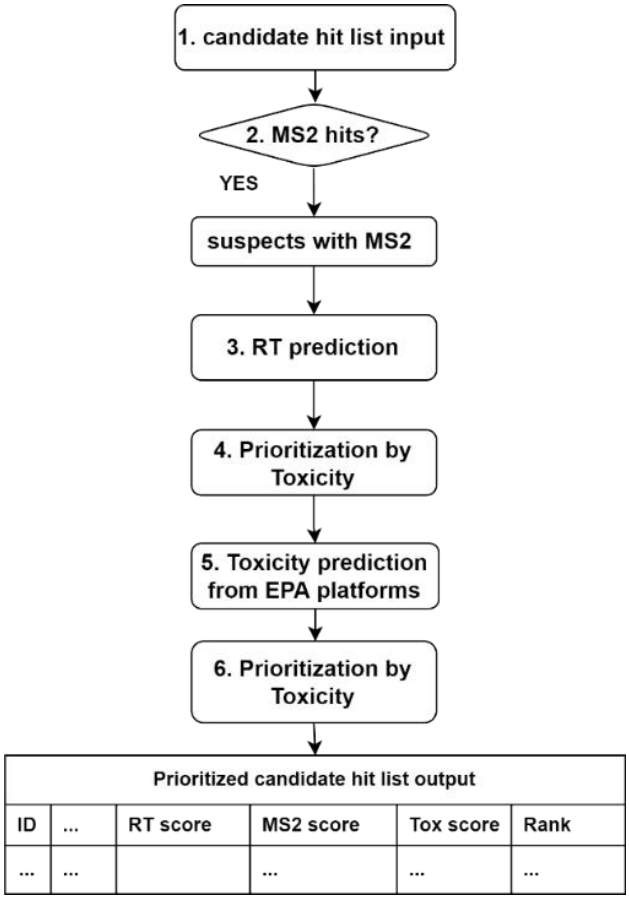


Figure S 6.3. Candidate hit lists prioritization workflow

548 **APPENDIX B**

549 **Appendix Table S2.1** Tandem mass spectral database data by 2017[31,32]

Database	Number of MSMS spectra	Number of compounds	Categories of compounds	Ionization methods	Instrumentation	note
NIST14 MS/MS	193120	9344	Metabolites, drugs, bioactive peptides, lipids, sugars, glycans, pesticides, surfactants, contaminants	CID HCD; 2-5 stepwise of CE changes; 100 different instrument types and ionization types like ESI and APCI, etc.	Majority from orbital ion traps (HCD); Rest from QTOF instruments. No QQQ;	
MassBank	22000	2800	Primary metabolites; flavonoids; gribberellins; saponins; carotenoids; phospholipids; (volatile natural products and synthetic drugs)	ESI for 679 metabolites; EI, CI, Fab for 10286 volatile natural and synthetic compounds.	21 instrument types. LC for ESI; GC for EI or Fab	
METLIN	72268	14034	Metabolites	0, 10, 20, 40 eV in CID; ESI in positive or negative	Agilent LC -QTOF	Silico spectra
LipiBlast	212516	119200	lipids		LC-QTOF, Q/orbitrap	In silico
MoNA	194000	68700				MS sourced from MB,

						lipidblast and in-house spectra
mzCloud	182000	2800	Metabolites and other	Multiple Ionization methods: ESI, NSI, APCI	CID, OR HCD	Developed by thermofisher
MetaboBASE	26000	13000	Human metabolites; Plant metabolites; Drugs;	10-40eV CEs		
GNPS	212230	12694	Plant metabolites; Natural products			Global MS spectra repository
Spektraris	2626	487	Plant metabolites	ESI, APCI; 10,20,40 eV		RT
ReSpect	9000	4000	Plant metabolites			Ionizaiton
MSforID	20000	1200	Drugs, pharmaceutical, pesticides and other small compounds	5-50 eV		
HMDB	5773	3729			GC; LC-MSMS	
MetaMS	150	150		APCI	GC QTOF	
Sumner Library	1734	289	Plant Metabolites: Flavonoids; Isoflavonoids; Phenylpropanoids	ESI; 10-60 eV stepwise CE	LC-Agilent 6430 QqQ	ESI
ChemicalSoft	1619	6476	Drugs and toxic compounds	Dif CEs	QTRAP	RT
UNPD-ISDB	170602	170602	Natural products			In silico spetrum
Designer drugs	10000	750				
LCMSn library	10000	6816	1500 Drugs, poisons and their	ESI	LXQ linear ion	

of Drugs/poisons			metabolites		trap	
MyCompoundID MSMS			Metabolites and predicted metabolites			In silico MSMS based on enzymatic reactions
DTU mycotoxin-Fungal	836	277	Secondary metabolite	Positive and negative; 10,20,40 CE	Agilent 6550 UPLC-QTOF	
WeizMass	3309	3540	Plant-based metabolites	MSn in positive and negative modes	Waters UPLC-QTOF	

550

551 **Appendix Table S3.1** MRM transitions and their instrument parameters of 20 authentic analytical standards from previous publications

Compound	Precursor ion (m/z)	FV/CV(V)*	PI1 (m/z)*	CE (eV)*	PI2 (m/z)*	CE (eV)*	References
Ampicillin	350[M + H]$^+$	70	160	10	106	15	1
	350	23	160	12	106	12	2
	350	25	160	12	106	17	3
Ciprofloxacin	332[M + H]$^+$	110	314	20	245	35	1
	332	40	288	16	245	23	2
	332	46	314	37	231	47	4
	332	51	314	32	288	34	5
	332	3k	314	20			6
Tetracycline	445[M + H]$^+$	110	428	5	410	15	1
	445	25	427	13	410	20	7
	444.16	90	154	28	410.1	16	8
	445	25	154	26	410	20	2

	445	21	427	26			9
	445.3	30	428	15			10
	445	3.5k			410	19	6
Vancomycin	724.8$[M+H]^{2+}$	120	144	12	100.1	40	8
Polymyxin B	602.6$[M+H]^{2+}$	/	101.1	32	241.2	24	11
Genistein	271.2$[M+H]^{+}$	155	215	25			12
	271		215.1				13
Hygromycin B	528.3 $[M + H]^{+}$	170	177.1	25	352	20	8
Malathion	331.1 $[M + H]^{+}$	90	99	20	127	10	14
	331	14	127	12	99	24	15
Streptomycin	582.4 $[M + H]^{+}$	180	263.2	30	245.8	35	8
BADGE*	358 $[M+NH4]^{+}$	20	191	20	135	30	18
Bisphenol S	251.04 $[M + H]^{+}$	2	93.1	24	157	16	17
	350.3	22	97	32	198	20	15
Rhodamine B	443.39 $[M + H]^{+}$	66	399.28	42			18
Tamoxifen	372.2 $[M + H]^{+}$	40	72.2	40			19
Triphenyl phosphate	327 $[M + H]^{+}$	35	215	26	153	30	20
	327	56	77	59	152	53	21
TCEP*	285 $[M + H]^{+}$	25	63	14	161	22	20
	285	26	63	39	99	33	21
	285	95	222.8	10			22
	284.8	11	63	49	98.9	33	23
Triclosan	287 $[M-H]^{-}$	75	35	6			1

	287.1	20	35	10			10
	287	-30	35	-24			21
	286.8	70	35	0			24
	287	-100	35	-40			25
	296.8	-60	34.9	-44			23
Triclocarban	313 $[M-H]^-$	100	166	10	35	25	1
	314.9	90	162	8	126	16	24
	313	100	160	25	126	25	26
	313	110	160	5	126	25	27
Ethyl paraben	165.1 $[M-H]^-$	-38	92	-24	137	-16	17
	165	-110	92	-32			25
	165.1	-28	91.9	-30	137	-30	28
	165	-29	92	-25	136	-22	23
Methyl paraben	151.1 $[M-H]^-$	-38	92	-22	135.9	-14	17
	151	-90	92	-29			25
	151.1	-40	91.9	-25	136.1	-25	28
	151	-30	92	-7	136	-20	23
TBBPA*	542.9 $[M-H]^-$	160	78.8	60			29

FV / CV*: Fragmentor voltage / Cone voltage; PI1*: Product ion 1; PI2*: Product ion 2; CE*: Collision energy
BADGE*: Bisphenol A diglycidyl ether; TCEP*: Tris(2-chloroethyl) phosphate; TBBPA*: Tetrabromo bisphenol A

Appendix Table S3.2 Ingredients of synthetic water

Reagent Added (mg/L)					PH
$NaHCO_3$	$CaSO_4 \cdot 2H_2O$	$MgSO_4$	KCl	Humic acid*	7.4

48.0	30.0	30.0	2.0	10	

*Humic acid was added to provide matrix effect from organic matter that exists commonly in environmental water samples.

Appendix Table 3.3 Ten molecular descriptors in the QSRR model

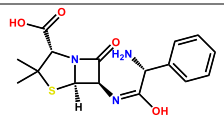
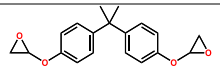
Molecular descriptors	Description	Coefficients
XLogP	Constitutional descriptors-describe hydrophobic/hydrophilic properties	0.2
CrippenLogP	Constitutional descriptors-describe hydrophobic/hydrophilic properties; Atom-based calculation of LogP	0.26
AATSC0p	Auto correlation descriptor/ Average centered Broto-Moreau autocorrelation - lag 0 / weighted by polarizabilities	5.56
ETA_EtaP_B	Extended Topochemical Atom descriptor/Branching index EtaB relative to molecular size	126.96
BCUTp.1h	BCUT descriptor/nlow highest polarizability weighted BCUTS	-0.21
AATS1i	Auto correlation descriptor/Average Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential	-0.19
AATS5i	Auto correlation descriptor/Average Broto-Moreau autocorrelation - lag 5 / weighted by first ionization potential	-0.01
AATS4i	Auto correlation descriptor/Average Broto-Moreau autocorrelation - lag 4 / weighted by first ionization potential	-0.1
AATS3i	Auto correlation descriptor/Average Broto-Moreau autocorrelation - lag 3 / weighted by first ionization potential	0.05

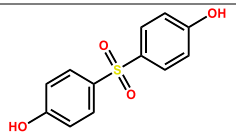
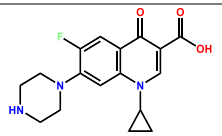
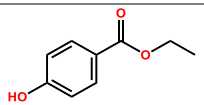
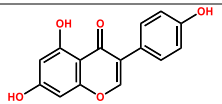
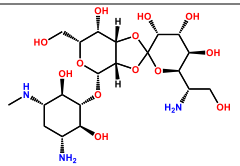
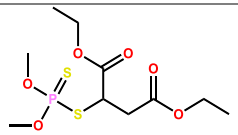
GATS1e	Auto correlation descriptor/Geary autocorrelation - lag 1 / weighted by Sanderson electronegativities	-1.94

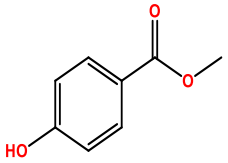
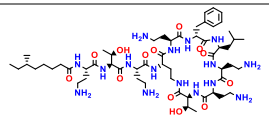
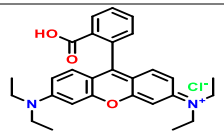
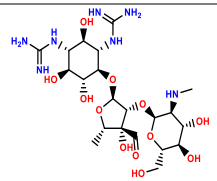
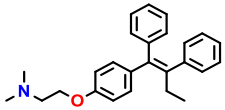
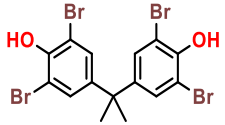
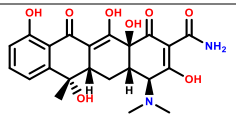
Appendix Table S3.4 Dewell volume differences between Agilent 1290 Infinity II system to Agilent 1290 Infinity I system

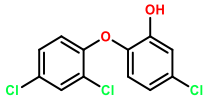
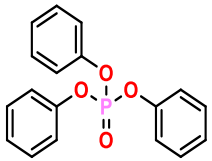
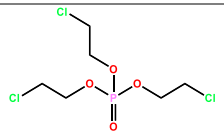
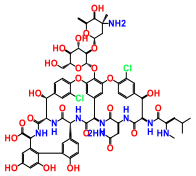
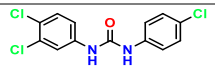
Instrument components	Volume in Agilent 1290 Infinity II system (µL)	Volume in Agilent 1290 Infinity I system (µL)	Volume differences (µL)
Tubing (Pump to the injector)	9	11	-3
Injector	180	125	55
Tubing (Injector to the column)	16	4	12
Total	205	140	64

Appendix Table S3.5 Retention time of 20 selected compounds predicted by in-house QSRR/MLR models

Compound ID	Structure	Formula	Molcular mass (Mw)	RT predicted by model (min)
Ampicillin		C16H19N3O4S	349.41	8.4
BADGE		C21H24O4	340.42	10.8

Bisphenol S		C12H10O4S	250.27	11.9
Ciprofloxacin		C17H18FN3O3	331.35	8.9
Ethyl paraben		C9H10O3	166.18	9.3
Genistein		C15H10O5	270.24	10.1
Hygromycin B		C20H37N3O13	527.52	3.1
Malathion		C10H19O6PS2	330.35	12.7

Methyl paraben		C8H8O3	152.15	8.6
Polymyxin B		C56H98N16O13	1203.50	5.9
Rhodamine B		C28H31ClN2O3	479.02	11.6
Streptomycin		C42H84N14O36S3	1457.38	2.6
Tamoxifen		C26H29NO	371.52	12.3
TBBPA		C15H12Br4O2	543.88	12.9
Tetracycline		C22H24N2O8	444.44	9.5

Triclosan		C12H7Cl3O2	289.54	13.8
Triphenyl phosphate		C18H15O4P	326.29	12.6
Tris(2-chloroethyl)phosphate		C6H12Cl3O4P	285.48	10.6
Vancomycin		C66H75Cl2N9O24	1449.27	8.2
Triclocarban		C13H9Cl3N2O	315.58	11.4

Appendix Table S3.6 Three types of scenarios in sludge water and urine sample from LC-MS/MS and HRMS of 20 compounds.

Compound	**Formula**	**Polarity**	**RT_{HRMS} (min)**	**$RT_{LC-MS/MS}$ (min)**	**$RT_{Predicted}$ (min) ±2.0 min**
Sludge water sample					
Scenario I					
Vancomycin	C66H75Cl2N9O24	Positive	6.5	6.5	6.2-10.2

Streptomycin	C21H39N7O12	Positive	1.2	1.1	0.6-4.6
Hygromycin B	C20H37N3O13	Positive	1.1	1.1	1.1-5.1
Tetracycline	C22H24N2O8	Positive	7.7	7.8	7.4-11.4
Ciprofloxacin	C17H18FN3O3	Positive	7.5	7.6	6.9-10.9
Malathion	C10H19O6PS2	Positive	13.4	13.4	10.7-14.7
TCEP	C6H12Cl3O4P	Positive	11.1	11.1	8.6-12.6
Triclocarban	C13H9Cl3N2O	Negative	14.4	14.5	9.4-13.4
Bisphenol S	C12H10O4S	Negative	9.5	9.5	9.9-13.9
Ethyl paraben	C9H10O3	Negative	10.6	10.6	6.3-11.3
Methyl paraben	C8H8O3	Negative	9.647	9.7	6.6-10.6
Scenario II					
Ampicillin	C16H19N3O4S	Positive	7.2/8.7	7.3	6.4-10.4
Triphenyl phosphate	C18H15O4P	Positive	14.2	14.2/18.9	10.6-14.6
Genistein	C15H10O5	Positive	10.7	7.6/9.3/10.7	8.1-12.1
Scenario III					
Polymyxin B	C56H98N16O13	Positive	ND*	7.7/16.0/18.5	3.8-7.8
Rhodamine B	C28H31ClN2O3	Positive	ND*	11.3/14.6/16.2/18.1	9.5-13.5

Tamoxifen	C26H29NO	Positive	ND*	11.5	10.3-14.3
TBBPA	C15H12Br4O2	Negative	ND*	1.5/14.6	14.9-18.9
Triclosan	C12H7Cl3O2	Negative	ND*	14.5/9.5	11.8-15.8
Urine sample					
Scenario I					
Vancomycin	C66H75Cl2N9O24	Positive	6.4	6.5	6.2-10.2
Tamoxifen	C26H29NO	Positive	11.4	11.5	10.3-14.3
Malathion	C10H19O6PS2	Positive	13.4	13.4	10.7-14.7
Triphenyl phosphate	C18H15O4P	Positive	14.1	14.2	10.6-14.6
TCEP	C6H12Cl3O4P	Positive	11.0	11.1	8.6-12.6
Genistein	C15H10O5	Positive	10.6	10.7	8.1-12.1
TBBPA	C15H12Br4O2	Negative	14.6	14.6	14.9-18.9
Triclosan	C12H7Cl3O2	Negative	14.4	14.5	11.8-15.8
Bisphenol S	C12H10O4S	Negative	9.5	9.5	9.9-13.9
Scenario II					
Polymyxin B	C56H98N16O13	Positive	7.3	7.4/11.2	3.8-7.8
Streptomycin	C21H39N7O12	Positive	1.2	1.2/10.2/12.4	0.6-4.6

Tetracycline	C22H24N2O8	Positive	1.4/7.7	7.8	7.4-11.4
Rhodamine B	C28H31ClN2O3	Positive	11.2/1.4	11.2	9.5-13.5
Ampicillin	C16H19N3O4S	Positive	7.2	7.2/1.9	6.4-10.4
Ethyl paraben	C9H10O3	Negative	8.6	1.6/7.8/8.9/10.6	6.3-11.3
Scenario III					
Ciprofloxacin	C17H18FN3O3	Positive	1.4/7.5	1.4/7.5	6.9-10.9
Triclocarban	C13H9Cl3N2O	Negative	14.4/16.7	14.4/16.7	9.4-13.4
Methyl paraben	C8H8O3	Negative	1.5/8.4/9.6	1.5/9.7	6.6-10.6
Hygromycin B	C20H37N3O13	Negative	1.1	ND*	1.1-5.1

ND*: Non detection (NA) due to its signal to noise ration lower than 3.

Appendix Table S3.7 Instrumental parameters of BPA and BPA metabolites on Agilent 6460 QqQ

Compound	Precursor ion (m/z)	FV/CV*(V)	PI1* (m/z)	CE* (eV)	PI2* (m/z)	CE (eV)	References
BPA*	227 $[M - H]^-$	110	132.8	21	212.1	13	30
BPA Glucuronide	403.1$[M - H]^-$	160	113.1	24	227	24	30
BPA monosulfate	307.3$[M - H]^-$	160	212.3	30	227.2	24	31

BPA*: Bisphenol A.; FV*: fragmentor voltage in agilent platform; CV*: Cone voltage in waters platform; CE*: Colision energy; PI1, 2*: Product ion 1, 2.

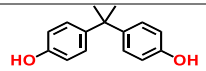
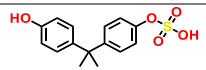
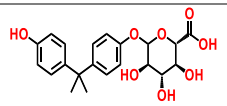
Appendix Table S3.8 CE optimization of BPA, BPAS and BPAG in cell extract

Collison energy (CE)	**$Abundance_{n\times CE}$ / $Abundance_{1\times CE}$**		
	BPA	BPAG	BPAS
	(227.1 > 212.1)	(403.1 > 113.1)	(307.3 > 227.2)

0.2CE	0.84	0.76	0.10
0.4CE	0.77	1.18	0.23
0.6CE	0.90	1.34	0.53
0.8CE	0.93	1.27	0.98
1.0CE	1.00	1.00	1.00
1.5CE	0.92	0.79	0.30
2.0CE	0.84	ND*	0.02
2.5CE	ND*	ND*	ND*

ND*: Non detection (ND) due to its signal to noise ration lower than 3.

Appendix Table 3.9 The predicted retention time by QSRR/MLR model and experimental retention time of BPA, BPA sulfate, and BPA glucuronide from cell extracts and standards.

Compound	Structure	RT_{STD} in MRM; HRMS / min	$RT_{Oberserved}$ in HRMS / min	$RT_{Oberserved}$ in LC-MS/MS / min	RT_{Pred} / min ±2.0 min
Bisphenol A		11.5; 11.5	11.5/19.0	11.5/19.1	9.5-13.5
BPA monosulfate		19.0; 19.0	19.0	19.2	13.1-17.1
BPA Glucuronide		9.7; 9.7	8.3/9.7	1.3/9.8	8.2-12.2

Appendix Table S6.2. Experimental retention time of 146 chemicals for the random forest model

No	Name	CAS	Molecular Formula	RT	Precursor Ion	SMILES
1	Cyromazine	66215-27-	C6H10N6	2.17	167.1034	C1CC1NC2=NC(=NC(=N2)N)N

		8				
2	Formetanate	22259-30-9	C11H15N3O2	2.74	222.1231	CNC(=O)OC1=CC=CC(=C1)N=CN(C)C
3	Fenuron	101-42-8	C9H12N2O	2.76	165.1018	CN(C)C(=O)NC1=CC=CC=C1
4	Acephate	30560-19-1	C4H10NO3PS	4.38	206.0006	CC(=O)NP(=O)(OC)SC
5	Omethoate	1113-02-6	C5H12NO4PS	5.18	236.0109	CNC(=O)CSP(=O)(OC)OC
6	Butocarboxin sulfoxide	34681-24-8	C7H14N2O3S	5.68	229.0605	CC(C(=NOC(=O)NC)C)S(=O)C
7	Dinotefuran	165252-70-0	C7H14N4O3	5.84	225.0952	CNC(=N[N+](=O)[O-])NCC1CCOC1
8	Promecarb	2631-37-0	C12H17NO2	5.88	208.1326	CC1=CC(=CC(=C1)OC(=O)NC)C(C)C
9	Butoxycarboxim	34681-23-7	C7H14N2O4S	6.01	245.0558	CC(C(=NOC(=O)NC)C)S(=O)(=O)C
10	Aldicarb sulfone	1646-88-4	C7H14N2O4S	6.18	245.0558	CC(C)(C=NOC(=O)NC)S(=O)(=O)C
11	Nitenpyram	150824-47-8	C11H15ClN4O2	6.33	271.0946	CCN(CC1=CN=C(C=C1)Cl)C(=C[N+](=O)[O-])NC
12	Oxamyl	23135-22-0	C7H13N3O3S	6.34	242.0561	CNC(=O)ON=C(C(=O)N(C)C)SC
13	Pymetrozin	123312-89-0	C10H11N5O	6.93	218.1031	CC1=NNC(=O)N(C1)N=CC2=CN=CC=C2
14	Flonicamid	158062-	C9H6F3N3O	6.94	230.0529	C1=CN=CC(=C1C(F)(F)F)C(=O)NCC#N

		67-0				
15	Thiamethoxam	153719-23-4	C8H10ClN5O3S	7.1	314.0077	CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl
16	Monocrotophos	6923-22-4	C7H14NO5P	7.44	246.0493	CC(=CC(=O)NC)OP(=O)(OC)OC
17	Dicrotophos	141-66-2	C8H16NO5P	7.79	260.0649	CC(=CC(=O)N(C)C)OP(=O)(OC)OC
18	Aminocarb	2032-59-9	C11H16N2O2	8.1	209.1279	CC1=C(C=CC(=C1)OC(=O)NC)N(C)C
19	Imidacloprid	138261-41-3	C9H10ClN5O2	8.13	278.0407	C1CN(C(=N[N+](=O)[O-])N1)CC2=CN=C(C=C2)Cl
20	Clothianidin	210880-92-5	C6H8ClN5O2S	8.3	271.9971	CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl
21	Fenuron	101-42-8	C9H12N2O	8.59	165.1018	CN(C)C(=O)NC1=CC=CC=C1
22	Vamidothion	2275-23-2	C7H14NO9P	8.7	310.0302	CC(C(=O)NC)SCCSP(=O)(OC)OC
23	Mevinphos I?	26718-65-0	C7H13O6P	8.74	247.0339	CC(=CC(=O)OC)OP(=O)(OC)OC
24	3-Hydroxy carbofuran	16655-82-6	C12H15NO4	8.74	260.0886	CC1(C(C2=C(O1)C(=CC=C2)OC(=O)NC)O)C
25	Acetamiprid	135410-20-7	C10H11ClN4	8.8	223.0739	CC(=NC#N)N(C)CC1=CN=C(C=C1)Cl
26	Dimethoate	60-51-5	C5H12NO3PS2	8.81	251.988	CNC(=O)CSP(=S)(OC)OC
27	Cymoxanil	57966-95-7	C7H10N4O3	9.36	221.0639	CCNC(=O)NC(=O)C(=NOC)C#N

28	Thiacloprid	111988-49-9	C10H9ClN4S	9.47	253.0301	C1CSC(=NC#N)N1CC2=CN=C(C=C2)Cl
29	Mevinphos II	26718-65-0	C7H13O6P	9.6	247.0339	CC(=CC(=O)OC)OP(=O)(OC)OC
30	Tricyclazole	41814-78-2	C9H7N3S	9.88	190.043	CC1=C2C(=CC=C1)SC3=NN=CN23
31	Thiabendazole	148-79-8	C10H7N3S	9.93	202.0429	C1=CC=C2C(=C1)NC(=N2)C3=CSC=N3
32	Aldicarb	116-06-3	C7H14N2O2S	9.96	213.0664	CC(C)(C=NOC(=O)NC)SC
33	Butocarboxim	34681-10-2	C7H14N2O2S	10.08	213.0664	CC(C(=NOC(=O)NC)C)SC
34	Fuberidazole	3878-19-1	C11H8N2O	10.24	185.0707	C1=CC=C2C(=C1)NC(=N2)C3=CC=CO3
35	Oxadixyl	77732-09-3	C14H18N2O4	10.3	301.1153	CC1=C(C(=CC=C1)C)N(C(=O)COC)N2CCOC2=O
36	Carbetamide	16118-49-3	C12H16N2O3	10.57	259.1044	CCNC(=O)C(C)OC(=O)NC1=CC=CC=C1
37	Propoxur	114-26-1	C11H15NO3	10.93	232.0938	CC(C)OC1=CC=CC=C1OC(=O)NC
38	Bendiocarb	22781-23-3	C11H13NO4	10.94	246.0729	CC1(OC2=C(O1)C(=CC=C2)OC(=O)NC)C
39	Carbofuran	1563-66-2	C12H15NO3	11.01	244.0937	CC1(CC2=C(O1)C(=CC=C2)OC(=O)NC)C
40	Metribuzin	21087-64-9	C8H14N4OS	11.08	215.0956	CC(C)(C)C1=NN=C(N(C1=O)N)SC
41	Tridiazuron	51707-55-	C9H8N4OS	11.12	221.0485	C1=CC=C(C=C1)NC(=O)NC2=CN=NS2

		2				
42	Pyracarbolid	24691-76-7	C13H15NO2	11.29	218.1171	CC1=C(CCCO1)C(=O)NC2=CC=CC=C2
43	Sulfentrazone	139-07-1	C11H10Cl2N4O3S	11.3	408.9697	CCCCCCCCCCCC[N+](C)(C)CC1=CC=CC=C1
44	Tebuthiuron	34014-18-1	C9H16N4OS	11.32	229.1111	CC(C)(C)C1=NN=C(S1)N(C)C(=O)NC
45	Carbaryl	63-25-2	C12H11NO2	11.38	224.0676	CNC(=O)OC1=CC=CC2=CC=CC=C21
46	Carboxin	5234-68-4	C12H13NO2S	11.43	258.0551	CC1=C(SCCO1)C(=O)NC2=CC=CC=C2
47	Spiroxamine	118134-30-8	C18H35NO2	11.59	298.2734	CCCN(CC)CC1COC2(O1)CCC(CC2)C(C)(C)C
48	Ethirimol	23947-60-6	C11H19N3O	11.67	210.1596	CCCCC1=C(N=C(NC1=O)NCC)C
49	Methiocarb	2032-65-7	C11H15NO2S	11.68	248.0708	CC1=CC(=CC(=C1SC)C)OC(=O)NC
50	Enilconazole (Imazalil)	35554-44-0	C14H14Cl2N2O	11.72	297.0547	C=CCOC(CN1C=CN=C1)C2=C(C=C(C=C2)Cl)Cl
51	Fluometuron	2164-17-2	C10H11F3N2O	11.75	233.0889	CN(C)C(=O)NC1=CC=CC(=C1)C(F)(F)F
52	Pirimicarb	23103-98-2	C11H18N4O2	11.79	239.1495	CC1=C(N=C(N=C1OC(=O)N(C)C)N(C)C)C
53	Thiofanox	39196-18-4	C9H18N2O2S	11.86	241.0974	CC(C)(C)C(=NOC(=O)NC)CSC
54	Chlortoluron	15545-48-	C10H13ClN2O	11.94	213.0785	CC1=C(C=C(C=C1)NC(=O)N(C)C)Cl

		9				
55	Flutriafol	76674-21-0	C16H13F2N3O	11.98	302.1093	C1=CC=C(C(=C1)C(CN2C=NC=N2)(C3=CC=C(C=C3)F)O)F
56	Simetryn	1014-70-6	C8H15N5S	12.03	214.1116	CCNC1=NC(=NC(=N1)SC)NCC
57	Methabenzthiazuron	18691-97-9	C10H11N3OS	12.13	222.0689	CNC(=O)N(C)C1=NC2=CC=CC=C2S1
58	Isoproturon	34123-59-6	C12H18N2O	12.23	207.1486	CC(C)C1=CC=C(C=C1)NC(=O)N(C)C
59	Isocarbophos	24353-61-5	C11H16NO4PS	12.26	312.0421	CC(C)OC(=O)C1=CC=CC=C1OP(=S)(N)OC
60	Metalaxyl	57837-19-1	C15H21NO4	12.31	302.1354	CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)COC
61	Diuron	330-54-1	C9H10Cl2N2O	12.39	233.0236	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl
62	Forchlorfenuron	68157-60-8	C12H10ClN3O	12.41	248.0577	C1=CC=C(C=C1)NC(=O)NC2=CC(=NC=C2)Cl
63	Desmedipham	13684-56-5	C16H16N2O4	12.45	323.0995	CCOC(=O)NC1=CC(=CC=C1)OC(=O)NC2=CC=CC=C2
64	Cycluron	2163-69-1	C11H22N2O	12.48	199.18	CN(C)C(=O)NC1CCCCCCC1
65	Phenmedipham	13684-63-4	C16H16N2O4	12.58	323.0995	CC1=CC(=CC=C1)NC(=O)OC2=CC=CC(=C2)NC(=O)OC
66	Prometon	7287-19-6	C10H19N5O	12.65	226.1657	CC(C)NC1=NC(=NC(=N1)SC)NC(C)C
67	Azoxystrobin	131860-33-8	C22H17N3O5	12.75	426.1051	COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC

68	Mexacarbate	315-18-4	C12H18N2O2	12.82	223.1435	CC1=CC(=CC(=C1N(C)C)C)OC(=O)NC
69	Fenpropimorph	67564-91-4	C20H33NO	12.84	304.2627	CC1CN(CC(O1)C)CC(C)CC2=CC=C(C=C2)C(C)(C)C
70	Furalaxyl	57646-30-7	C17H19NO4	12.88	324.1196	CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)C2=CC=CO2
71	Nuarimol	63284-71-9	C17H12ClFN2O	12.93	315.0686	C1=CC=C(C(=C1)C(C2=CC=C(C=C2)F)(C3=CN=CN=C3)O)Cl
72	Secbumeton	26259-45-0	C10H19N5O	12.93	226.1656	CCC(C)NC1=NC(=NC(=N1)NCC)OC
73	Dimethomorph I	110488-70-5	C21H22ClNO4	12.94	388.1299	COC1=C(C=C(C=C1)C(=CC(=O)N2CCOCC2)C3=CC=C(C=C3)Cl)OC
74	Methoprotryne	841-06-5	C11H21N5OS	12.94	272.1529	CC(C)NC1=NC(=NC(=N1)NCCCOC)SC
75	Ametryn	834-12-8	C9H17N5S	12.97	228.127	CCNC1=NC(=NC(=N1)SC)NC(C)C
76	Fenamidone	161326-34-7	C17H17N3OS	13	312.1157	CC1(C(=O)N(C(=N1)SC)NC2=CC=CC=C2)C3=CC=CC=C3
77	Mandipropamid	374726-62-2	C23H22ClNO4	13.03	434.1116	COC1=C(C=CC(=C1)CCNC(=O)C(C2=CC=C(C=C2)Cl)OCC#C)OCC#C
78	Linuron	330-55-2	C9H10Cl2N2O2	13.04	270.9997	CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC
79	Boscalid	188425-85-6	C18H12Cl2N2O	13.13	343.0388	C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl
80	Siduron	1982-49-6	C14H20N2O	13.13	233.1567	CC1CCCCC1NC(=O)NC2=CC=CC=C2

81	Dimethomorph II	110488-70-5	C21H22ClNO4	13.2	388.1295	COC1=C(C=C(C=C1)C(=CC(=O)N2CCOCC2)C3=CC=C(C=C3)Cl)OC
82	Paclobutrazol	76738-62-0	C15H20ClN3O	13.26	294.1357	CC(C)(C)C(C(CC1=CC=C(C=C1)Cl)N2C=NC=N2)O
83	Siduron	1982-49-6	C14H20N2O	13.28	233.1567	CC1CCCCC1NC(=O)NC2=CC=CC=C2
84	Myclobutanil	88671-89-0	C15H17ClN4	13.35	289.1206	CCCCC(CN1C=NC=N1)(C#N)C2=CC=C(C=C2)Cl
85	Cyproconazole I	94361-06-5	C15H18ClN3O	13.4	292.134	CC(C1CC1)C(CN2C=NC=N2)(C3=CC=C(C=C3)Cl)O
86	Mepronil	55814-41-0	C17H19NO2	13.43	270.148	CC1=CC=CC=C1C(=O)NC2=CC(=CC=C2)OC(C)C
87	Bifenazate	149877-41-8	C17H20N2O3	13.45	323.1358	CC(C)OC(=O)NNC1=C(C=CC(=C1)C2=CC=CC=C2)OC
88	Fluoxastrobin	361377-29-9	C21H16ClFN4O5	13.51	459.0861	CON=C(C1=CC=CC=C1OC2=C(C(=NC=N2)OC3=CC=CC=C3Cl)F)C4=NOCCO4
89	Chloroxuron	1982-47-4	C15H15ClN2O2	13.52	291.0885	CN(C)C(=O)NC1=CC=C(C=C1)OC2=CC=C(C=C2)Cl
90	Mefenacet	73250-68-7	C16H14N2O2S	13.52	321.0659	CN(C1=CC=CC=C1)C(=O)COC2=NC3=CC=CC=C3S2
91	Bromuconazole	116255-48-2	C13H12BrCl3N3O	13.57	375.9602	C1C(COC1(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl)Br
92	Cyproconazole II	94361-06-5	C15H18ClN3O	13.62	292.1203	CC(C1CC1)C(CN2C=NC=N2)(C3=CC=C(C=C3)Cl)O

93	Tetraconazole	112281-77-3	C13H11Cl2F4N3O	13.63	372.0277	C1=CC(=C(C=C1Cl)Cl)C(CN2C=NC=N2)COC(C(F)F)(F)F
94	Spirotetramat	203313-25-1	C21H227NO5	13.66	396.1768	CCOC(=O)OC1=C(C(=O)NC12CCC(CC2)OC)C3=C(C=CC(=C3)C)C
95	Fluquinconazole	136426-54-5	C16H8Cl2FN5O	13.66	376.0188	C1=CC2=C(C=C1F)C(=O)N(C(=N2)N3C=NC=N3)C4=C(C=C(C=C4)Cl)Cl
96	Iprovalicarb	140923-17-7	C18H28N2O3	13.68	343.1982	CC1=CC=C(C=C1)C(C)NC(=O)C(C(C)C)NC(=O)OC(C)C
97	Flufenacet	142459-58-3	C14H13F4N3O2S	13.7	386.0544	CC(C)N(C1=CC=C(C=C1)F)C(=O)COC2=NN=C(S2)C(F)(F)F
98	Fenarimol	60168-88-9	C17H12Cl2N2O	13.7	331.0391	C1=CC=C(C(=C1)C(C2=CC=C(C=C2)Cl)(C3=CN=CN=C3)O)Cl
99	Prometryn	7287-19-6	C10H19N5S	13.71	242.1425	CC(C)NC1=NC(=NC(=N1)SC)NC(C)C
100	Mepanipyrim	110235-47-7	C14H13N3	13.73	224.1176	CC#CC1=NC(=NC(=C1)C)NC2=CC=CC=C2
101	Etaconazole	60207-93-4	C14H15Cl2N3O2	13.8	328.0606	CCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl
102	Terbutryn	886-50-0	C10H19N5S	13.87	242.1426	CCNC1=NC(=NC(=N1)SC)NC(C)(C)C
103	Flusiliazole	85509-19-9	C16H15F2N3Si	13.93	316.1069	C[Si](CN1C=NC=N1)(C2=CC=C(C=C2)F)C3=CC=C(C=C3)F
104	Picoxystrobin	117428-22-5	C18H16F3NO4	13.94	390.0913	COC=C(C1=CC=CC=C1COC2=CC=CC(=N2)C(F)(F)F)C(=O)OC
105	Bupirimate	41483-43-	C13H24N4O3	14.05	317.1636	CCCCC1=C(N=C(N=C1OS(=O)(=O)N(C)C)NCC)C

		6	S			
106	Tebufennozide	112410-23-8	C22H28N2O2	14.06	375.2033	CCC1=CC=C(C=C1)C(=O)NN(C(=O)C2=CC(=CC(=C2)C)C)C(C)(C)C
107	Spinetoram	935545-74-7	C42H69NO10	14.1	748.4994	CCC1CCCC(C(C(=O)C2=CC3C(C2CC(=O)O1)CCC4C3CC(C4)OC5C(C(C(C(O5)C)OC)OCC)OC)C)OC6CCC(C(O6)C)N(C)C
108	Dimoxystrobin	149961-52-4	C19H22N2O3	14.13	349.1514	CC1=CC(=C(C=C1)C)OCC2=CC=CC=C2C(=NOC)C(=O)NC
109	Kresoxin-methyl	143390-89-0	C18H19NO4	14.18	336.1199	CC1=CC=CC=C1OCC2=CC=CC=C2C(=NOC)C(=O)OC
110	Carfentrazone-ethyl	128639-02-1	C15H14Cl2F3N3O3	14.21	434.0244	CCOC(=O)C(CC1=CC(=C(C=C1Cl)F)N2C(=O)N(C(=N2)C)C(F)F)Cl
111	Neburon	555-37-3	C12H16Cl2N2O	14.22	275.0704	CCCCN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl
112	Diclobutrazole	66345-62-8	C15H19Cl2N3O	14.22	328.0969	CC(C)(C)C(C(CC1=C(C=C(C=C1)Cl)Cl)N2C=NC=N2)O
113	Penconazole	66246-88-6	C13H15Cl2N3	14.28	284.0709	CCCC(CN1C=NC=N1)C2=C(C=C(C=C2)Cl)Cl
114	Tebuconazole	107534-96-3	C16H22ClN3O	14.32	308.1518	CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O
115	Propiconazole	60207-90-1	C15H17Cl2N3O2	14.43	342.0762	CCCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl
116	Pyraclostrobin	175013-18-0	C19H18ClN3O4	14.47	410.0867	COC(=O)N(C1=CC=CC=C1COC2=NN(C=C2)C3=CC=C(C=C3)Cl)OC

117	Benalaxyl	98243-83-5	C20H23NO3	14.47	348.1562	CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)CC2=CC=CC=C2
118	Hexaconazole	79983-71-4	C14H17Cl2N3O	14.53	314.0816	CCCCC(CN1C=NC=N1)(C2=C(C=C(C=C2)Cl)Cl)O
119	Zoxamide	156052-68-5	C14H16Cl3NO2	14.54	336.0311	CCC(C)(C(=O)CCl)NC(=O)C1=CC(=C(C(=C1)Cl)C)Cl
120	Metconazole	125116-23-6	C17H22ClN3O	14.58	320.1515	CC1(CCC(C1(CN2C=NC=N2)O)CC3=CC=C(C=C3)Cl)C
121	Fenhexamid	126833-17-8	C14H17Cl2NO2	14.62	340.0261	CC1(CCCCC1)C(=O)NC2=C(C(=C(C=C2)O)Cl)Cl
122	Thiobencarb	28249-77-6	C12H16ClNOS	14.69	258.0704	CCN(CC)C(=O)SCC1=CC=C(C=C1)Cl
123	Pencycuron	66063-05-6	C19H21ClN2O	14.73	329.1408	C1CCC(C1)N(CC2=CC=C(C=C2)Cl)C(=O)NC3=CC=CC=C3
124	Difenoconazole	119446-68-3	C19H17Cl2N3O3	14.75	406.0708	CC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)OC4=CC=C(C=C4)Cl)Cl
125	Trifloxystrobin	141517-21-7	C20H19F3N2O4	14.8	431.1179	CC(=NOCC1=CC=CC=C1C(=NOC)C(=O)OC)C2=CC(=CC=C2)C(F)(F)F
126	Diniconazole	70217-36-6	C15H17Cl2N3O	14.81	326.0816	CC(C)(C)C(C(=CC1=C(C=C(C=C1)Cl)Cl)N2C=NC=N2)O
127	Novaluron	116714-46-6	C17H9ClF8N2O4	14.86	515.0005	C1=CC(=C(C(=C1)F)C(=O)NC(=O)NC2=CC(=C(C=C2)OC(C(OC(F)(F)F)F)(F)F)Cl)F
128	Ipconazole I	125225-	C18H24ClN3O	14.9	334.1671	CC(C)C1CCC(C1(CN2C=NC=N2)O)CC3=CC=C(C=C3)Cl

		28-7				
129	Ipconazole II	125225-28-7	C18H24ClN3O	15.03	334.1672	CC(C)C1CCC(C1(CN2C=NC=N2)O)CC3=CC=C(C=C3)Cl
130	Clethodim	99129-21-2	C17H26ClNO3S	15.17	360.1388	CCC(=NOCC=CCl)C1=C(CC(CC1=O)CC(C)SCC)O
131	Benfuracarb	82560-54-1	C20H30N2O5S	15.24	433.1753	CCOC(=O)CCN(C(C)C)SN(C)C(=O)OC1=CC=CC2=C1OC(C2)(C)C
132	Temephos	3383-96-8	C16H20O6P2S3	15.27	466.9668	COP(=S)(OC)OC1=CC=C(C=C1)SC2=CC=C(C=C2)OP(=S)(OC)OC
133	Tebufenpyrad	119168-77-3	C18H24ClN3O	15.29	334.1673	CCC1=NN(C(=C1Cl)C(=O)NCC2=CC=C(C=C2)C(C)(C)C)C
134	Furathiocarb	65907-30-4	C18H26N2O5S	15.33	405.1445	CCCCOC(=O)N(C)SN(C)C(=O)OC1=CC=CC2=C1OC(C2)(C)C
135	Lufenuron	103055-07-8	C17H8Cl2F8N2O3	15.34	532.9667	C1=CC(=C(C(=C1)F)C(=O)NC(=O)NC2=CC(=C(C=C2Cl)OC(C(C(F)(F)F)F)(F)F)Cl)F
136	Buprofezin	69327-76-0	C16H23N3OS	15.4	306.1628	CC(C)N1C(=NC(C)(C)C)SCN(C1=O)C2=CC=CC=C2
137	Pyriproxyfen	95737-68-1	C20H19NO3	15.51	322.1433	CC(COC1=CC=C(C=C1)OC2=CC=CC=C2)OC3=CC=CC=N3
138	Flufenoxuron	101463-69-8	C21H11ClF6N2O3	15.63	511.0243	C1=CC(=C(C(=C1)F)C(=O)NC(=O)NC2=C(C=C(C=C2)OC3=C(C=C(C=C3)C(F)(F)F)Cl)F)F
139	Hexythiazox	78587-05-0	C17H21ClN2O2S	15.67	375.0894	CC1C(SC(=O)N1C(=O)NC2CCCCC2)C3=CC=C(C=C3)Cl

140	Quinoxyfen	124495-18-7	C15H8Cl2FNO	15.69	308.0034	C1=CC(=CC=C1OC2=C3C(=CC(=CC3=NC=C2)Cl)Cl)F
141	Progargite	2312-35-8	C19H26O4S	15.74	373.1433	CC(C)(C)C1=CC=C(C=C1)OC2CCCCC2OS(=O)OCC#C
142	Spiromesifen	283594-90-1	C23H30O4	15.79	393.2023	CC1=CC(=C(C(=C1)C)C2=C(C3(CCCC3)OC2=O)OC(=O)CC(C)(C)C)C
143	Etoxazole	153233-91-1	C21H23F2NO2	15.8	360.1762	CCOC1=C(C=CC(=C1)C(C)(C)C)C2COC(=N2)C3=C(C=CC=C3F)F
144	Chlorfluazuron	71422-67-8	C20H9Cl3F5N3O3	15.87	561.9518	C1=CC(=C(C(=C1)F)C(=O)NC(=O)NC2=CC(=C(C(=C2)Cl)OC3=C(C=C(C=N3)C(F)(F)F)Cl)Cl)F
145	Pyridaben	96489-71-3	C19H25ClN2OS	16.23	387.1257	CC(C)(C)C1=CC=C(C=C1)CSC2=C(C(=O)N(N=C2)C(C)(C)C)Cl
146	Eprinomectin	133305-88-1	C50H75NO14	16.24	936.5084	CCC(C)C1C(C=CC2(O1)CC3CC(O2)CC=C(C(C(C=CC=C4COC5C4(C(C=C(C5O)C)C(=O)O3)O)C)OC6CC(C(C(O6)C)OC7CC(C(C(O7)C)NC(=O)C)OC)OC)C)C
147	Fenazaquin	120928-09-8	C20H22N2O	16.4	307.1798	CC(C)(C)C1=CC=C(C=C1)CCOC2=NC=NC3=CC=CC=C32
148	Avermectin B1a	65195-55-3	C48H72O14	16.45	895.482	CCC(C)C1C(C=CC2(O1)CC3CC(O2)CC=C(C(C(C=CC=C4COC5C4(C(C=C(C5O)C)C(=O)O3)O)C)OC6CC(C(C(O6)C)OC7CC(C(C(O7)C)O)OC)OC)C)C
149	Moxidectin	113507-06-5	C37H53NO8	16.69	662.366	CC1CC(=CCC2CC(CC3(O2)CC(=NOC)C(C(O3)C(=CC(C)C)C)C)OC(=O)C4C=C(C(C5C4(C(=CC=C1)CO5)O)O)C)C

Appendix Table S6.3. 137 peaks identified as significant markers by PCA and S-plot in sample wise comparison.

Peak ID	m/z	Retention time (min)	Average peak_abundance
13.86_388.1676n	411.1567872	13.86438333	88435.89733
9.19_310.0752n	311.082458	9.19085	96271.72336
16.42_531.3833n	1085.755912	16.41756667	182381.4268
14.32_422.3141n	445.3033487	14.31653333	158299.0467
8.07_192.1370n	215.1262333	8.06855	290893.5281
15.18_357.2547n	737.4986986	15.17601667	322711.7416
17.89_348.1752n	371.1643978	17.88813333	1029515.871
13.92_251.1089n	252.1161554	13.92193333	90816.62647
12.99_212.1518n	447.2929001	12.99028333	763733.7259
16.25_510.8737n	1022.754634	16.24715	290648.2901
10.97_252.1629n	275.1520793	10.96526667	60335.8068
11.07_308.2187n	331.2078825	11.06975	956746.3461
15.30_390.7912n	804.5715636	15.29875	254861.7051
15.45_386.2771n	795.5433677	15.44838333	345285.8586
15.30_305.1578n	306.165074	15.29875	48014.1713
14.89_317.7706n	674.5043535	14.88881667	400239.2186
14.82_266.1682n	289.157414	14.81856667	81425.78187
14.52_288.7507n	616.4646241	14.51893333	393279.7397
12.14_366.2606n	389.2498281	12.14241667	983967.1526
13.87_116.0728n	271.1087029	13.87151667	28709.93684
15.67_415.2983n	853.585904	15.67311667	331548.9282
16.70_568.4862n	607.4493149	16.7001	220944.2099

9.83_250.1788n	273.1680098	9.830183333	924953.6719
6.99_472.2537n	495.2429503	6.985516667	287133.4682
19.45_842.7345n	865.7237431	19.44583333	1753082.145
6.51_428.2274n	451.2165827	6.513166667	266578.9008
7.67_237.0779n	238.0851568	7.665516667	65025.00255
18.74_530.4022n	553.3914655	18.74066667	260443.7947
6.31_264.0693n	265.0766026	6.308266667	81566.31342
21.11_662.4485n	663.4557588	21.10625	2435640.666
20.12_474.4081n	497.3973119	20.11943333	396397.8193
4.19_183.0121n	206.0012837	4.19155	27766.47842
8.26_567.3999n	609.4337236	8.2582	490534.2844
13.93_283.1353n	306.1245443	13.92916667	28928.61309
13.92_253.1018n	254.109068	13.92193333	52556.66138
21.54_544.4836n	577.5171337	21.5401	228485.882
21.38_698.5862n	721.5754011	21.37821667	158562.4988
13.84_370.1553n	371.162612	13.8355	33076.42433
14.85_656.4677n	679.4568951	14.84651667	486734.1536
14.18_540.3718n	563.3610367	14.18016667	673697.9561
21.52_868.6950n	869.7022383	21.51805	535672.6525
17.86_1044.5235n	1067.512679	17.85656667	169717.4486
17.01_348.1755n	371.1647436	17.01091667	94893.38437
20.01_651.3016n	674.290817	20.00725	171009.8408
15.46_375.8097n	790.5825007	15.4562	431149.4309

18.86_616.6150n	617.6223093	18.8554	182988.3177
15.68_423.8094n	848.6260499	15.6812	396321.6962
15.90_444.3186n	911.6263353	15.8954	239868.9139
15.90_452.8312n	906.6697427	15.90246667	417967.3141
16.09_481.8505n	964.7083104	16.09123333	322019.1337
18.53_362.1909n	385.1801445	18.53188333	379654.636
18.22_530.4716n	553.460852	18.21748333	1725014.65
16.42_539.8961n	1080.799434	16.41756667	228801.8607
17.90_616.6139n	639.6031177	17.89603333	87278.30926
14.99_352.2745n	375.2636997	14.98958333	463948.1932
13.65_482.3331n	505.3223492	13.65413333	729591.932
21.70_663.3603n	702.3234214	21.7047	176693.5978
11.94_828.4770n	829.4842372	11.93985	228753.7305
11.85_258.0006n	259.0079017	11.85013333	16972.25579
6.98_233.6450n	490.2791363	6.978666667	247121.5043
7.51_728.3425n	729.3498179	7.512866667	130529.6341
11.50_214.0514n	215.0587149	11.4963	11660.53153
11.47_383.2597n	767.5266544	11.46858333	2253526.595
12.24_232.0175n	233.0247401	12.24401667	25924.94784
11.40_285.2347n	609.4326389	11.39695	897856.8721
11.20_720.4784n	753.5119248	11.20115	196026.1496
11.80_212.0774n	213.0846268	11.8029	33083.34412
10.97_129.0541n	171.087943	10.96526667	27773.45153

4.21_141.9858n	142.9930749	4.214266667	69647.45828
12.87_248.0129n	249.0202252	12.86605	11466.27446
10.49_240.0895n	241.0967543	10.48866667	30620.2179
8.49_692.3852n	715.3744237	8.489966667	154687.2254
10.94_748.4731n	781.5065818	10.94493333	167244.6499
11.63_315.2054n	316.212648	11.6328	117270.4775
7.66_578.3400m/z	578.3399995	7.658666667	211357.3606
18.78_536.1673m/z	536.1672711	18.77916667	75007.65788
7.97_622.3669m/z	622.3669198	7.973966667	248972.4109
8.07_627.3225m/z	627.3225125	8.06855	378876.9409
19.07_578.5216m/z	578.5215877	19.06968333	69843.11236
18.29_811.7562m/z	811.7561775	18.29403333	547874.1724
19.47_822.7530m/z	822.753008	19.46843333	3106561.543
19.91_610.1850m/z	610.1850094	19.91111667	69966.25356
8.22_305.2201m/z	305.2200801	8.223083333	344303.0638
13.09_233.1310m/z	233.1309797	13.09318333	7342.798264
7.33_534.3131m/z	534.3131493	7.3316	210152.0795
21.69_619.5293m/z	619.5292672	21.6897	66913.86913
21.63_810.7534m/z	810.7533634	21.6305	197829.5123
21.61_891.7393m/z	891.7393124	21.60833333	202062.9287
21.59_836.7694m/z	836.7694145	21.5862	148167.9602
21.56_879.7393m/z	879.7393204	21.55676667	387663.1879
21.54_853.7259m/z	853.7258875	21.5401	5373649.58

3.41_123.0555m/z	123.0555001	3.413883333	22470.21308
8.39_263.1084m/z	263.1083611	8.38785	2697.98503
21.50_848.7698m/z	848.7698174	21.49583333	2423435.784
6.74_185.0358m/z	185.0357875	6.740966667	1058.028543
20.99_795.6501m/z	795.6500512	20.99426667	207272.8854
20.80_792.7069m/z	792.7069156	20.79678333	64109.23982
20.64_836.7685m/z	836.7684781	20.6353	641244.0509
20.61_639.6036m/z	639.6036495	20.61323333	292428.253
20.35_828.2329m/z	828.2329088	20.35053333	36132.11893
20.23_827.7083m/z	827.7082866	20.2309	5956997.081
6.53_446.2587m/z	446.2587332	6.526983333	214788.3372
8.39_295.1300m/z	295.1300481	8.38785	27552.91526
9.99_229.0749m/z	229.074865	9.9858	14446.90319
17.30_375.2626m/z	375.2626085	17.30146667	236540.475
13.22_442.2976m/z	442.2976428	13.22418333	255796.0147
13.48_289.2054m/z	289.2053863	13.4804	51229.81818
13.48_289.3035m/z	289.3035135	13.4804	32.5547752
13.49_500.3805m/z	500.3805355	13.48756667	227493.1138
13.49_295.1201m/z	295.1201091	13.49488333	2049.790319
13.09_230.1176m/z	230.1175974	13.08581667	21391.34284
12.90_442.3047m/z	442.3047245	12.89503333	223097.9632
12.84_230.1179m/z	230.1179054	12.8372	45204.18174
13.87_313.1608m/z	313.1607897	13.87151667	45966.13849

21.70_841.6731m/z	841.6730516	21.69723333	355870.3767
12.49_229.0759m/z	229.07592	12.4939	218.5903629
12.45_259.0080m/z	259.0079806	12.44506667	335.3744992
14.06_558.4232m/z	558.4232287	14.05791667	326186.9121
12.03_218.0985m/z	218.0985261	12.03428333	27000.08405
8.39_563.1234m/z	563.1233737	8.38785	369.0619215
14.49_621.4141m/z	621.4140798	14.49031667	439212.4951
14.63_304.4410m/z	304.4410305	14.63221667	31616.19745
14.65_304.2623m/z	304.2623345	14.64621667	236008.2858
14.78_304.3009m/z	304.3008714	14.77603333	168258.7759
15.06_749.5211m/z	749.5211325	15.05555	184913.4615
11.61_298.2021m/z	298.2020704	11.61246667	94289.95156
15.19_732.5365m/z	732.5365075	15.19093333	388009.2639
11.60_338.1949m/z	338.1948703	11.59898333	518273.502
15.71_749.5186m/z	749.5186468	15.7133	154963.8385
10.94_202.0855m/z	202.0854697	10.93808333	47429.49497
13.09_174.0549m/z	174.0549468	13.09318333	25762.51642
9.75_180.1024m/z	180.1024183	9.747083333	9866.678752
9.29_188.0710m/z	188.0709831	9.286033333	22140.63658
8.49_710.4187m/z	710.4186615	8.489966667	162864.9951
17.01_749.5204m/z	749.5203649	17.01091667	154690.6148
12.03_216.1012m/z	216.1011605	12.03428333	56769.32498
3.03_142.0088m/z	142.0088429	3.03075	17941.7868

Appendix Table S6.4. Toxicity scores threshold for different toxicity level

Endpoints	Level 1	Level 2	Level 3
FMLC50 /-Log10(mol/L)	6-9	2-6	0-2
DMLC50 /-Log10(mol/L)	6-9	2-6	0-2
TPIGC50 /-Log10(mol/L)	6-9	2-6	0-2
ORLD50 /-Log10(mol/kg)	3-5	1.5-3	0-1.5
DT	>=0.5	<0.5	-
AM	>=0.5	<0.5	-
ToxPi	1-0.6	0.3-0.6	0-0.3