THESE DE DOCTORAT DE

L'ECOLE DES HAUTES ETUDES
EN SANTE PUBLIQUE DE RENNES

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Santé Publique

Par
Jade CHAKER

Methodological developments for the non-targeted characterization of the human internal chemical exposome in epidemiological studies

Optimizing and implementing a high-throughput workflow for the identification of new biomarkers of exposure in blood plasma and serum samples

Thèse présentée et soutenue à Rennes, le 1er Juin 2022
Unité de recherche : Institut de recherche en santé, environnement et travail - Irset - Inserm UMR 1085
Cette thèse a été préparée dans le cadre du Réseau doctoral en santé publique animé par l'Ecole des Hautes Etudes en Santé Publique

NNT : 2022HESP0002

Rapporteurs avant soutenance :

Frédérique COURANT	Maître de conférence – Université de Montpellier (France)
Christophe JUNOT	Directeur de recherche – CEA de Saclay (France)
François LESTREMAU	Ingénieur de recherche – IMT Mines Alès (France)

Composition du Jury :

Présidente :	Valérie CAMEL	Professeure – AgroParisTech Paris (France)
Examinateurs :	Michèle BOUCHARD	Professeure – Université de Montréal (Canada)
	Valérie CAMEL	Professeure – AgroParisTech Paris (France)
	Frédérique COURANT	Maître de conférence – Université de Montpellier (France)
	Christophe JUNOT	Directeur de recherche – CEA de Saclay (France)
	François LESTREMAU	Ingénieur de recherche – IMT Mines Alès (France)
Dir. de thèse :	Arthur DAVID	Enseignant-chercheur – EHESP Rennes (France)

In memory of Bernard Jégou (1951-2021),

Bernard was a passionate and committed researcher, curious of everything and genuinely interested in people. He contributed greatly to the field of male reproduction, and started to study the influence of environmental factors on human reproduction in the early nineties. From 1995 to 2004, he led Inserm unit 435 GERM (Study group of male reproduction), then Inserm unit 625 (Study group of human and mammalian reproduction) from 2005 to 2012. He was a major actor in the creation in 2012 of Irset (Inserm unit 1085), dedicated to researching the impact of the environment on health, and was its director from 2012 to 2019. He was also the research director of the French school of public health (EHESP) from 2014 to 2020. He believed in the importance of interdisciplinarity in science and in public health, and played a key role in organizing and consolidating the processes of health research in the organizations he directed. He also greatly participated in the visibility gain of environmental health research by being one of the pioneers of the exposome concept.

Throughout the decades, he helped train and mentor generations of scientists. He leaves behind an exceptional legacy of exploring science, sharing knowledge, and acting for the common good.

Remerciements

Ces travaux de thèse ont été menés au Laboratoire d'Etude et de Recherche en Santé Environnement (LERES). Je souhaiterais remercier Philippe Quénel et Vincent Bessonneau, qui ont successivement dirigé le LERES, pour m'avoir accueillie dans ce laboratoire. Le LERES est également une plateforme technologique rattachée à l'Ecole des Hautes Etudes en Santé Publique (EHESP) et à l'Institut de recherche en santé, environnement et travail (Irset UMR1085). Je souhaiterais donc remercier à nouveau Philippe Quénel ainsi que Luc Multigner, qui m'ont accueillie au sein de l'équipe 9 de l'Irset « Evaluation des expositions et recherche épidémiologique sur l'environnement, la reproduction et le développement » (3ERD) dont ils étaient co-responsables. Je remercie également Cécile Chevrier et Ronan Garlantézec, qui sont actuellement co-responsables de cette équipe, aujourd'hui baptisée ELIXIR (Épidémiologie et science de l'exposition en santé-environnement).

Je remercie Frédérique Courant, Christophe Junot et François Lestremau, pour avoir accepté d'être les rapporteurs de ce travail. Je remercie également Michèle Bouchard et Valérie Camel d'avoir accepté d'être membres de mon jury de thèse.

J'aimerais remercier les membres de mon comité de suivi, Jean-Philippe Antignac et Laurent Debrauwer, pour les discussions toujours bienveillantes et constructives que nous avons eues au cours de ma thèse.

C'est avec beaucoup d'émotion que j'aimerais remercier mon premier directeur de thèse, Bernard Jégou. C'était un scientifique brillant, passionné et engagé, mais aussi une personne d'une humanité et d'une gentillesse exceptionnelles. Il avait une capacité exceptionnelle à tisser des liens forts avec les gens qu'il rencontrait, et malgré le peu de temps que j'ai eu la chance de le connaître, il a su me transmettre sa passion pour la communication, l'interdisciplinarité, et la curiosité pour toutes choses au-delà de la science. Bernard, merci d'avoir souvent imprimé des articles en pensant à nous, merci pour tes explications sur la reproduction du requin et des vautours, et merci pour toutes tes recommandations de conférences et d'expositions. Comme tu le disais toujours… kénavo !

Un très grand merci à Arthur David, qui a d'abord été mon co-encadrant puis mon directeur de thèse. Merci de m'avoir offert cette superbe opportunité, d'avoir encadré cette thèse pendant trois ans (et demi !), et d'avoir su trouver cet équilibre entre m'accompagner et me laisser la liberté d'explorer ce domaine qui continue de me passionner. Tes conseils avisés, ta disponibilité et ta bienveillance en toutes circonstances m'ont permis de vivre cette thèse avec beaucoup de sérénité et d'enthousiasme. Si je dis à tous les nouveaux que la chose la plus

importante dans une thèse est la relation avec son équipe de direction, parce que tout est surmontable quand on est bien encadré, c'est en immense partie grâce à toi !

Je souhaiterais remercier chaleureusement les membres de l'équipe 9 de l'Irset. Un immense merci à Cécile Chevrier, Christine Monfort et Charline Warembourg pour votre aide sur Pélagie 12, qu'elle concerne la disponibilité des échantillons, ou les discussions motivantes sur les innombrables possibilités d'exploitation des résultats. Un grand merci également à Nathalie Costet qui, au travers de nombreuses présentations et discussions, m'a aidée à mieux comprendre les enjeux statistiques du traitement de données comme les nôtres. Enfin, merci beaucoup à Anne-Claire, Noriane, Zohra, Hélène, Maximilien, Yara, Alan, et tous les autres doctorants, internes en santé publique, stagiaires et post-doctorants de l'équipe, avec qui j'ai adoré discuter au fil des années, et qui m'ont appris à parler (un peu) cette langue étrangère qu'est « l'épidémiologiste » !

J'aimerais aussi remercier toute l'équipe du LERES. Merci aux « habitants » du deuxième étage de m'avoir accueillie dans les labos, et de m'avoir aidée à chaque fois que j'avais un pépin. Un merci tout particulier à Aude Dimeglio, pour m'avoir formée à l'utilisation de SMILE, pour m'avoir aidée tellement de fois à comprendre pourquoi le QTOF était (encore) en erreur, et surtout, pour toutes ces discussions rafraîchissantes qu'on a eues en attendant de réparer ledit QTOF. Je te remercie sincèrement d'avoir été aussi disponible pour moi dès le début de ma thèse ; je ne m'en serais pas sortie toute seule face à ce bébé capricieux ! Merci également à l'équipe administrative, notamment Fleur Chaumet pour ton aide précieuse sur les budgets, Eve Laigle pour ta livraison soutenue de petits cadeaux (même quand il n'y en a pas pour moi !) et Véronique Jolle pour ta disponibilité et ta réactivité face à toutes nos demandes. Merci également aux chercheurs et au personnel de recherche en général, pour avoir partagé votre savoir, vos travaux et vos conseils avec moi au cours de repas, séminaires, et réunions. Un mot aussi pour Malé : merci beaucoup pour ta bonne humeur, ton hospitalité, tes récitations de poèmes arabes et tes histoires qui me rappellent un peu mon enfance.

Un immense merci aux membres de l'équipe logiciel. Merci à Arthur, d'avoir initié ce projet, en lequel je crois beaucoup et sur lequel je suis très heureuse de pouvoir continuer à travailler. Merci à Thibaut Léger, pour nous avoir offert ton recul, tes idées débordantes, et tes astuces improbables mais toujours efficaces ! Et évidemment, merci à Erwann Gilles, pour avoir accepté (et réussi avec brio) la lourde tâche de coder ce petit bijou de technologie. Merci pour ton investissement sans compter, pour ton SAV ouvert même le week-end, pour tes innombrables optimisations du code, et pour les (très) longues discussions sur les isotopologues et autres considérations pratico-pratiques payées en café bien corsé.

Je voudrais également remercier les stagiaires, doctorants et post-doctorants qui sont passés par le labo au cours de ces années : Gaëlle, Lise, Jade V., Ibrahim, Habiba, Laura ... Eva, Fatima et Shareen, merci beaucoup d'avoir accepté de rejoindre notre équipe ; j'admire vos compétences de chercheuses mais aussi votre bonne humeur et votre gentillesse. Ashna, merci pour ton implication sans faille dans tout ce que tu entreprends, y compris tes amitiés, pour ton côté maman poule et pour ton empathie à toute épreuve. Thomas, le petit dernier de l'équipe, merci pour les déjà nombreux fou-rires sur la capacité de ton estomac, ou encore sur l'équipement idéal pour prélever du placenta. Alexis, merci pour ta patience, pour tes coups de fil, et surtout, un grand, grand merci pour 4M. Kahina et Juliana, un grand merci à toutes les deux pour votre amitié précieuse, pour m'avoir laissé vous embarquer dans les nombreuses soirées tricot et crochet, et pour vos mots gentils dans les moments difficiles. Enfin, un immense merci à Hiago, avec qui j'ai vécu toute cette aventure qu'est la thèse. Merci de m'avoir écoutée si souvent, d'avoir partagé mes frustrations et mes angoisses, mais aussi mes joies et mes victoires.

Merci à mes amis de longue date : Camille, Laurène, Jennifer, Benjamin, Julia, Quentin, Dylan... Merci pour les souvenirs précieux, pour les belles photos depuis qu'on est éparpillés aux quatre coins de la France et du monde, merci pour votre patience quand la thèse a occupé tout mon temps et mon esprit.

A Younès et Romain, mes encadrants de stage de fin d'étude, merci de m'avoir aidée à découvrir mon goût pour la recherche, et d'avoir accepté de me parler des joies du doctorat !

Je voudrais remercier la communauté de doctorants du serveur discord PhD Students. Un grand merci à vous tous, que j'ai appris à connaître à la fin de ma thèse, mais qui avez beaucoup contribué à ma dose de rire quotidienne et à ma sérénité sur la fin de ma période de rédaction.

A mes amis les plus chers, Solenn et Romain (et Pwick), merci infiniment pour tout. Merci pour votre gentillesse, pour les soirées et les nuits jeux de société, et pour m'avoir accueillie chez vous quand j'en avais besoin. Merci de m'avoir encouragée et motivée à être toujours meilleure. Je vous souhaite beaucoup de bonheur dans votre toute nouvelle vie avec Noé !

Enfin, je voudrais dire un grand merci à ma famille. Merci à mon père, pour avoir toujours encouragé ma curiosité scientifique (« Dis, tu me poses une question ? »), et à ma mère, pour son oreille attentive, son soutien sans faille, et ses mots d'encouragements toujours justes. Merci à mon frère et ma sœur pour les conversations hilarantes, les petits messages de félicitations qui réchauffent le cœur, et nos mots d'amour fraternel toujours recouverts sous une bonne dose de sarcasme. Peut-être loin des yeux, mais jamais loin du cœur !

Table of contents

TABLE OF CONTENTS 2
TABLE OF ILLUSTRATIONS 9
LIST OF ABBREVIATIONS 15
SCIENTIFIC VALORIZATION 18
1. Articles in peer-reviewed journals 19
1.1. As part of the main research project 19
1.1.1. Published 19
1.1.2. In preparation 19
1.2. As part of side projects 19
2. Oral communications 20
2.1. International conferences 20
2.2. National and regional conferences 20
3. Posters 21
4. Large-scale collaboration 21
5. Outreach and science popularization activities 22
6. Supervising and training 22
RESUME DE LA THESE EN FRANÇAIS 23
1. Introduction 25
2. Acquisition de l'empreinte chimique : optimiser l'équilibre entre sensibilité et sélectivité 27
3. Prétraitement des données et développement d'un logiciel de profilage de suspects 30
3.1. Adaptation des logiciels de prétraitement des données aux applications en exposomique . 30
3.2. Développement d'un logiciel pour assister les approches de profilage de suspects 31
4. Application du workflow développé au sein de la cohorte mère-enfant Pélagie 32
5. Conclusions et perspectives 35
GENERAL INTRODUCTION 39
CHAPTER I. APPLICATION OF HRMS-BASED EXPOSOMICS TO COHORT-BASED EPIDEMIOLOGICAL STUDIES: STATE-OF-THE-ART AND CHALLENGES 45
1. Studying the human internal chemical exposome: context, definitions and challenges 46
1.1. The Exposome: from a concept to a call to action 46
1.2. Main conceptual challenges for the non-targeted characterization of the human chemical exposome: study design questionings 50
1.2.1 Direct and indirect measurements: choosing between environmental and biological matrices... 50
1.2.2 Choosing the biological matrix 52
1.2.3 Analytical platform choice 53
2. Implementing NTA to characterize the exposome: constructing a non-targeted and suspect screening workflow 56

2.1. Acquisition of the chemical fingerprint ... 57
2.2. Data processing ... 59
2.3. Interbatch correction ... 61
2.4. Statistical analysis ... 62
2.5. Annotation: non-targeted and suspect screening ... 63
2.6. Semi-quantification ... 66
2.7. Reporting ... 67
2.8. Conclusion ... 67
3. When non-targeted and suspect screening meet epidemiology: first large-scale applications, achievements and remaining challenges ... 68
3.1. Large-scale applications and achievements ... 68
3.1.1. From 2012 to 2017 ... 69
3.1.2. From 2017 to 2020 (extended to 2022) ... 70
3.1.3. From 2020 onwards ... 71
3.2. Remaining limitations ... 73
3.2.1. Statistical power in non-targeted applications ... 73
3.2.2. The incomplete annotation process ... 73
3.2.3. Interpretability of results: toxicology and determinants of exposure ... 74
4. Conclusion ... 75
CHAPTER II. MATERIAL AND METHODS ... 86
1. Instrumental method development and optimization ... 87
1.1. Mix of standards used for the optimizations ... 87
1.2. Quality assurance and quality control procedures ... 88
1.3. LC method optimization ... 89
1.3.1. Column diameter and flow rate optimization ... 89
1.3.2. Reconstitution phase optimization ... 90
1.4. MS optimization ... 92
1.4.1. MS acquisition ... 92
1.4.2. MS2 acquisition ... 92
1.4.2.1. Data dependent acquisition ... 93
2. Sample preparation methods for non-targeted exposomics ... 94
2.1. Protein precipitation ... 95
2.2. Protein and Phospholipid removal ... 96
2.3. Supported Liquid Extraction ... 97
2.4. Solid Phase Extraction ... 97
3. Data processing methods for non-targeted exposomics ... 98
3.1. Data processing tools ... 99
3.2. Peak picking ... 99
3.3. Alignment ... 100

3.4. Gap filling ... 101
3.5. Normalization ... 101
4. Annotation methods and tools ... 102
4.1. Non-targeted screening: statistical analysis ... 102
4.2. Suspect screening tool ... 102
4.2.1. Suspect screening predictors ... 103
4.2.2. Library module: generating suspects data ... 105
4.2.3. Suspect screening module: computing confidence indices ... 106
4.2.4. Manual curation ... 109
5. Biological samples ... 109
CHAPTER III. SYSTEMATIC EVALUATIONS OF BLOOD-DERIVED SAMPLE PREPARATION METHODS FOR HRMS-BASED CHEMICAL EXPOSOMICS ... 113
1. Context and summary ... 114
2. Abstract ... 116
3. Introduction ... 117
4. Experimental section ... 119
4.1. Biological samples ... 119
4.2. Sample preparation methods comparison ... 119
4.2.1. Preselection ... 120
4.2.2. Comparison to PPT at real-life concentrations ... 121
4.2.3. Final comparison ... 121
4.3. Data acquisition and quality assurance procedures ... 122
4.4. Data processing ... 122
4.4.1. Non-targeted data processing ... 122
4.4.2. Targeted data processing ... 122
4.5. Suspect screening and annotation ... 122
4.5.1. Suspect screening tool ... 122
4.5.2. Annotation ... 123
5. Results and discussion ... 123
5.1. Preselection of most suitable SPM ... 123
5.2. Comparison to PPT at real-life concentrations ... 125
5.3. Final comparison with MDL determination and application on cohort samples ... 128
6. Conclusion ... 131
7. Associated content ... 132
7.1. Supporting Information ... 132
8. Author information ... 132
9. Acknowledgements ... 132
10. References ... 132

CHAPTER IV. OPTIMIZING DATA PROCESSING FOR EXPOSOMICS APPLICATIONS: UNCOVERING THE POTENTIAL OF LOW-ABUNDANT PEAKS AND MS1 DATA 135
1. Context and summary 136
2. Abstract 138
3. Introduction 139
4. Experimental section 140
4.1. Spiking experiments and sample preparation 140
4.2. Data acquisition and quality control 141
4.3. Peak picking optimization: data processing tools 141
4.4. Suspect screening predictors 142
4.4.1. Mass-to-charge ratio (m/z) 142
4.4.2. Retention time 142
4.4.3. Isotopic pattern 143
4.5. Suspect screening annotation tools 144
4.6. Data availability 145
5. Results and discussion 145
5.1. Optimization of HRMS data processing tools for exposomics studies 145
5.1.1. XCMS: automated optimization versus manual selection criteria 145
5.1.2. Manual optimization of MZmine 2 and vendor software 147
5.1.3. Comparison of optimized data processing tools to detect low abundant compounds 147
5.2. Modelling suspect screening predictors 149
5.2.1. Retention time prediction models 149
5.2.2. Isotopic pattern 150
5.3. Efficiency of the suspect screening tool and comparison with other annotation tools 152
6. Conclusion 154
7. Associated content 154
7.1. Supporting information 154
8. Author information 154
9. Acknowledgements 155
10. References 155
CHAPTER V. IMPLEMENTING A LARGE-SCALE SUSPECT SCREENING APPROACH TO CHARACTERIZE THE HUMAN CHEMICAL EXPOSOME 158
1. Outgrowing the scale of a batch: quality control 160
2. Implementing a suspect screening approach at a large scale 164
3.1. Comparing the use of MS1 and MS2 predictors for annotation in an exposomics context. 164
3.2. Describing the environmental chemical exposures in the Pélagie cohort 168
3.3. Exploring the potential of dual sample preparation 171
CONCLUSION AND PERSPECTIVES 178

APPENDICES 183
1. Appendix 1. Chapter II 184
1.1. Detailed list of the optimization mix and internal standards 184
1.2. Column diameter and flow rate optimization 188
1.3. Detailed list of the retention time prediction set 193
2. Appendix 2. Supporting information – Chapter III 199
2.1. Table A1 – Standard compounds form and suppliers 199
2.2. Table A2 – Standard compounds physical-chemical characteristics 201
2.3. Table A3 – Preselection: Recovery, repeatability and matrix effect of all sample preparation methods on individual compounds 202
2.4. Table A4a – Comparison to PPT (Serum): Detection, repeatability, S/N and spiking significance of preselected preparation methods on individual compounds 206
2.5. Table A4b – Comparison to PPT (Plasma): Detection, repeatability, S/N and spiking significance of preselected preparation methods on individual compounds 208
2.6. Table A5a – Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification 210
2.7. Table A5b – Application to cohort samples (Plasma-Danish cohort): Annotations and semi-quantification 220
2.8. Table A6 – Phree and PPT methods detection limits on 30 xenobiotics 226
2.9. Appendix S.1 – Solvents and chemicals 227
2.10. Appendix S.2 – Data acquisition 227
2.11. Appendix S.3 – Quality control procedures 228
2.12. Appendix S.4 – Sample preparation procedures 228
2.13. Appendix S.5 – Application of PPT and Phree to cohort samples 230
3. Appendix 3. Supporting information – Chapter IV 232
3.1. Table A1 – Standard compounds form and suppliers 232
3.2. Table A2 – Computer specifications 236
3.3. Table A3 – Sets of compounds used for spiking samples (n=45), training models (n=134) and evaluating them (n=30) 237
3.4. Table A4 – Calibrant sets used in positive and negative mode for the RTI platform 241
3.5. Table A5.1 – Results of data processing workflows on individual compounds in serum 242
3.6. Table A5.2 – Results of data processing workflows on individual compounds in plasma 250
3.7. Table A5.3 – Summary of results of data processing workflows on individual compounds in plasma and serum 258
3.8. Table A6.1 – Results of annotation after manual curation in serum 259
3.9. Table A6.2 – Results of annotation after manual curation in plasma 261
3.10. Appendix S.1 – Chemicals and solvents 263
3.11. Appendix S.2 – Data acquisition 263
3.12. Appendix S.3 – Quality control 263

3.13. Appendix S.4 – In-house annotation workflow 264
3.14. Appendix S.5 – Modelling the retention time predictor 265
3.15. Appendix S.6 – Optimization of individual data processing tools 266
3.16. Appendix S.7 – Application of the in-house software in real-life conditions 267
4. Appendix 4. Chapter V 269
4.1. Effect of total ion current correction on mean feature area and principal component analysis results 269
4.2. Annotations on Pélagie samples 272
4.3. MS2 data for annotated compounds 276
4.4. Confidence levels, detection frequency and toxicological data 280
4.5. Classification of compounds annotated in Pélagie samples 282
4.6. Presence of annotated compounds on shared suspect lists 285

Table of illustrations

Figures

Figure I.1 – Schematized representation of the distinction between endogenous metabolites and exogenous chemicals and related biotransformation products. These small molecules (50–1200 Da) present in human biological matrices can be profiled using High Resolution Mass Spectometry. 48

Figure I.2 - Conceptual visualisation of the impact of overarching methodological choices on the profiled fraction of the exposome by David et al., Env Int., 2021. Specificities and overlaps of the different HRMS platforms are schematically represented. Log Kow=octanol/water partition coefficient; GC=gas chromatography; LC=liquid chromatography; IC=ion chromatography, CE=capillary electrophoresis, ESI=Electrospray ionisation, HRMS=High Resolution Mass Spectrometry 55

Figure I.3 – Main steps of a non-targeted and suspect screening workflow implemented to investigate the chemical exposome in biological matrices 56

Figure I.4 – Main steps of the non-targeted data processing workflow, comprising of peak picking, alignment, gap filling, and normalization. These steps are presented on quality control (QC) samples. Various strategies and algorithms are available for the peak picking step, the alignment step and the normalization step, as detailed in Chapter II. 60

Figure I.5 – Identification confidence levels in high-resolution mass spectrometry proposed by Schymanski et al., ES&T, 2014. 64

Figure II.1 – Overview of the physical-chemical properties of the 50-compound optimization mix, including endogenous compounds (in blue) and exogenous compounds (in orange). The octanol-water partition coefficient (logP) and the monoisotopic mass (Da) are presented. 88

Figure II.2 – Extracted ion chromatogram for Aminobenzimidazole (logP = 0.91) in ESI (+) mode (A) and Arachidonic acid (logP = 6.98) in ESI (−) mode (B) depending on the reconstitution phase composition (generated with a m/z tolerance of 10 ppm). 91

Figure II.3 – Operating principle of protein precipitation. Samples are mixed with an organic solvent (usually methanol or acetonitrile) at a solvent:sample ratio of 1:1 to 4:1. After a prolonged contact, centrifugation allows forming a protein pellet (in orange) and the purified supernatant can be used... 96

Figure II.4 – Operating principle of solid phase extraction. Samples are filtered through a stationary phase that retains phospholipids (in shades of orange) and leaves other compounds (in green) pass through..... 96

Figure II.5 - Operating principle of supported liquid extraction. The sample is loaded onto a sorbent, which retains the entire sample. The analytes are then selectively eluted using an immiscible organic solvent. 97

Figure II.6 - Operating principle of solid phase extraction. The solid phase is conditioned, followed by sample loading. Interferents are washed usually using water, and compounds of interest are eluted using an elution solvent. 98

Figure II.7 – Representation of the peak picking process, which consists of four steps: centroidation, peak detection, creation of chromatogram objects and deconvolution. 100

Figure II.8 – Schematized operating principle of the in-house annotation workflow in four steps: comparing successively m/z, Rt and isotopic fit, then generating a global scoring. 105

Figure III.1 – Graphical abstract of the research paper titled "Comprehensive evaluation of blood plasma and serum sample preparations for HRMS-based chemical exposomics: overlaps and specificities 116

Figure III.2 – Diagram of the methodology used to compare sample preparation methods. Two low-level spiking experiments were conducted to compare various phospholipid and protein removal plates (PLR), solid phase extraction cartridges (SPE), and supported liquid extraction cartridge (SLE) among themselves, and to the classically used protein precipitation (PPT). The best-suited methods were selected using a set of qualitative and quantitative criteria, then applied to plasma and serum cohort samples to observe the impact of the sample preparation method on the visible chemical space. ... 120

Figure III.3 – Comparison of the recovery (A), repeatability (B), and median matrix effect performances (C) of the eleven considered sample preparation methods using a 50-compound mix spiked in serum (n=4). Preparation methods include protein precipitation (PPT), phospholipid removal (PLR) plates, solid phase extraction (SPE) cartridges, and a supported liquid extraction (SLE) cartridge. For the recovery and repeatability criteria, Q1, Q3 and median values are represented with two green lines and one blue line respectively. Median values for two spiking concentrations are presented for the matrix effect criterion. 123

Figure III.4 – Sample preparation methods evaluation for the detection of 50 low-level spiked compounds in (A) serum and (B) plasma samples (n=4 each). Outer edges identify best performances. 126

Figure III.5 – Comparison of annotated xenobiotics' areas in samples prepared with protein precipitation (PPT) and protein removal plate Phree in Pelagie serum samples (A) and Danish plasma samples (B). Logged values of fold changes (i.e. area ratio between Phree and PPT) are presented on the x-axis, where –∞ and +∞ values represent the absence of compounds in samples prepared with Phree and PPT, respectively. Bars on the left of the y-axis represent compounds presenting higher areas in PPT samples and vice-versa. 129

Figure IV.1 – Graphical abstract for the research paper titled "From metabolomics to HRMS-based exposomics: adapting peak-picking and developing scoring for MS1 suspect screening" 138

Figure IV.2 - Data preprocessing flowchart illustrating all tested parameters, including default parameters for the authors' system (*) and optimized parameters (in bold and red) for each data preprocessing software tool..... 142

Figure IV.3 - Data processing (i.e. peak picking, deconvolution, alignment, gap filling) evaluation using XCMS for detection and semi-quantification of low-level spiked compounds in plasma samples (n=4 each). Four sets of parameters were used: Default (blue squares), manual (green rounds), IPO (purple triangles), and Autotuner (orange diamonds) optimization. Outer edges identify best performances. 146

Figure IV.4 - Data processing (i.e. peak picking, deconvolution, alignment, gap filling) evaluation for detection and semi-quantification of low-level spiked compounds in (A) plasma and (B) serum samples (n=4 each). Four optimized software tools were used: MZmine 2 (blue squares), XCMS (green rounds), MarkerView™ (purple triangles), and Progenesis QI (orange diamonds). Outer edges identify best performances..... 148

Figure IV.5 - Construction (A) of two Rt prediction models using simple linear regression models and validation (B) of all usable Rt prediction models. The logP model uses experimental octanol-water partition coefficients as predictors and the RTI model uses Retention Time Indices (RTI) as predictors. PredRet, a fourth Rt prediction tool, was also tested. PredRet predictions are not depicted as the number of responses were significantly lower (n=16), rendering it not statistically comparable..... 149

Figure IV.6 - Linear regression analysis of A_2/A_0 according to P_2/P_0. Prediction bands placed at 3 RMSE (99%) are depicted in dotted lines. Compounds are separated into 11 groups based on contents in Br, Cl, and S atoms (and combinations) 151

Figure IV.7 - Comparison of five suspect screening tools: xMSannotator (blue), MS-DIAL (purple), msPurity (green), MZmine2 (yellow) and in-house tool (red). Comparison was made on use of in-house databases, use of predicted or experimental Rt and MS/MS, speed of implementation, scoring and prioritization. Details are available in SI Fig.B6. 153

Figure V.1 – Schematized representation of a dual sample preparation process, where half of the supernatant from protein precipitation is injected as is (after reconstitution), and the other half is used for further protein and phospholipid removal before injection on the UHPLC-ESI-QTOF. In total, 960 samples were injected including QCs and MS2 acquisitions. 160

Figure V.2 – Quality control parameters for the application of two sample preparation methods to cohort samples (n=75 samples) before correction. Outer edges identify best performances..... 161

Figure V.3 – Mean feature raw area (A) and mean feature area after total ion current correction (B), shown on samples (including the composite quality control samples) prepared by protein precipitation (PPT) injected in ESI (+) mode on the UHPLC-ESI-QTOF. Blank samples for each batch are identified by orange squares. 162

Figure V.4 – PCA using raw area (A) and PCA using area after total ion current correction (B), shown on samples prepared by protein precipitation (PPT) injected in ESI (+) mode on the UHPLC-ESI-QTOF. 163

Figure V.5 – Quality control parameters for the application of two sample preparation methods to cohort samples (n=75 samples) after correction. Outer edges identify best performances. 163

Figure V.6 – MS1 predictors supporting the pentachlorophenol (A- isotopic pattern, C- retention time) and the triclosan glucuronide (B- isotopic pattern, D- retention time) annotations. Theoretical and experimental isotopic patterns are compared based on coherence between mass/charge ratios and isotopic area ratios. Experimental retention times are compared to values predicted using RTI (orange), Retip (yellow) or a polarity-based linear regression (logP) (green) and their respective confidence intervals (represented by the color gradients). This data was acquired in negative ionization mode on the UHPLC-ESI-QTOF. 166

Figure V.7 – Updated identification confidence levels accounting for new methodological tools, such as prediction models for retention time (Rt) and biotransformation products. MS2 refers to any form of fragmentation. 167

Figure V.8 – Classification of the major source of annotated compounds (n=92), expressed in percentages. Gut microbiota metabolites are shown in yellow, compounds obtained from food in blues, compounds obtained from health and personal hygiene products in greens, and industrial compounds in oranges. 168

Figure V.9 – Detection of suspect compounds per class in each participant (separated by batch) in protein precipitated samples (A) and Phree samples (B). Preservatives and other stabilizers found in processed foods, health and personal hygiene products and industrial compounds were combined in a single category for clarity. 169

Figure V.10 -Comparison of annotated xenobiotics' areas in samples prepared with protein precipitation (PPT) and protein removal plate Phree in Pelagie serum samples. Logged values of fold changes (i.e. area ratio between Phree and PPT) are presented on the x-axis, where –∞ and +∞ values represent the absence of compounds in samples prepared with Phree and PPT, respectively. Bars on the left of the central vertical axis represent compounds presenting higher areas in PPT samples and vice-versa. 173

Equations

Equation II.1 - Expression of Confidence Indices (CI) for all predictors (i= m/z, Rt, or A_n/A_0 ratio, where A_n refers to the area of the nth isotopologue). Δ_i is a confidence interval and is specifically defined for each predictor as the maximal acceptable deviation from the reference value. 106

Equation III.1– Matrix effect formula, where A is the peak area of a given compound X at a given concentration C. 121

Equation IV.1 – Expression of Confidence Indices (CI) for all predictors (i= m/z, Rt, or M_2/M_0 ratio). Δ_i is a confidence interval and is specifically defined for each predictor as the maximal acceptable deviation from the reference value..... 144

Tables

Table I.1 – Research initiatives investigating the links between the chemical exposome and health. NIEHS: National Institute of Environmental Health Sciences. Adapted from David et al., m/s, 2021... 69

Table II.1 – Median area (and area CV) of compounds from the optimization mix injected in four replicate depending on the column diameter and flow rate. 90

Table II.2 – Impact of the reconstitution phase composition on areas of 50 compounds spiked in serum homogenates and injected on UPLC-ESI-QTOF in positive and negative ionization modes. 91

Table II.3 – Results of the Information Dependent Analysis (IDA) method optimization through the selection of adequate maximum precursor ions per scan for the mix injected at 10ng/ml on QTOF in ESI (−) and ESI (+) modes 93

Table II.4 - Example of SWATH windows generated by the vendor SWATH windows calculator on plasma quality control samples in ESI (−) and ESI (+) ionization modes 94

Table III.1– Percentage of features of quality control samples categorized by fold change value (i.e. area ratio of features in Phree and protein precipitation). Values are computed for Pelagie serum samples and Danish plasma samples. 129

Table V.1 – Overview of the data generated by two suspect screening tools, based on either MS1 or MS2 predictors (In-house software and MS-DIAL respectively). Median and cumulated values are determined based on the four sample preparation method × ionization mode possible combinations (i.e. protein precipitation and Phree phospholipid removal plate in positive and negative ionization modes). 164

Table V.2 – Percentage of features of quality control samples injected in positive and negative ionization modes on the UPLC-ESI-QTOF, categorized by fold change (FC) values (i.e. area ratio of features in Phree and protein precipitation). 172

List of abbreviations

ABH: Adaptive Benjamini-Hochberg

ACN: Acetonitrile

ADAP: Automated Data Analysis Pipeline

ATHLETE: Advancing Tools for Human Early Lifecourse Exposome Research and Translation

BDE: Brominated Diphenyl Ether

CCS: Collision cross section

CE: Capillary Electrophoresis

CEC: Chemicals of Emerging Concern

CHEAR: Children's Health Exposure Analysis Resource

CI: Confidence Indices

CV: Coefficient of Variation

CWT: Continuous Wavelet Transform

DDA: Data Dependent Acquisition

DIA: Data Independent Acquisition

DDT: Dichlorodiphenyltrichloroethane

EFS: Etablissement Français du Sang – French Blood Agency

EHEN: European Human Exposome Network

EIRENE: EnvIRonmental Exposure assessmeNt in Europe

ESFRI: European Strategy Forum on Research Infrastructures

ESI: ElectroSpray Ionization

EWAS: Exposome-Wide Association Study

EXPANSE: EXposome Powered tools for healthy living in urbAN Settings

FC: Fold change

FDR: False Discovery Rate

FP7: Seventh Framework Programme

GC: Gas Chromatography

GWAS: Genome-Wide Association Study

HBM: Human BioMonitoring

HGP: Human Genome Project

HHEAR: Human Health Exposure Analysis Resource

HILIC: Hydrophilic Interaction LIquid Chromatography

HRMS: High Resolution Mass Spectrometry

IC: Ion Chromatography

LC: Liquid Chromatography

LLE: Liquid-Liquid Extraction

MTBE: Methyl Tert-Butyl Ether

m/z: Mass-to-charge ratio

NIEHS: National Institute of Environmental Health Sciences

NTA: Non-Targeted Approaches

PARC: Partnership for the Assessment of Risks from Chemicals

PCA: Principal Component Analysis

PCB: Polychlorinated Biphenyls

PLR: PhosphoLipid Removal

PLS: Partial Least Square

PLS-DA: Partial Least Square-Discriminant Analysis

POP: Persistent Organic Pollutants

PPT: Protein PrecipiTation

QC: Quality Control

QTOF: Quadrupole Time-Of-Flight

RANSAC: RANdom SAmple Consensus

RP: Reverse Phase

Rt: Retention time

SLE: Supported Liquid Extraction

S/N: Signal-to-noise ratio

SNP: Single Nucleotide Polymorphisms

SPE: Solid Phase Extraction

SPF: Santé Publique France

SPM: Sample Preparation Method

SPME: Solid Phase MicroExtraction

S/N: Signal-to-noise ratio

SWATH: Sequential Window Acquisition of all THeoretical fragment ion spectra

TOF: Time-Of-Flight

U(H)PLC: Ultra(-High) Performance Liquid Chromatography

Scientific valorization

1. Articles in peer-reviewed journals

1.1. As part of the main research project

1.1.1. Published

- **Chaker, J.**, Kristensen, D. M., Halldorsson, T. I., Olsen, S.F., Monfort, C., Chevrier, C., Jégou, B., David, A.* (2022). Comprehensive Evaluation of Blood Plasma and Serum Sample Preparations for HRMS-Based Chemical Exposomics: Overlaps and Specificities. *Anal Chem* **(IF=6.8)**, 94(2), 866–874.
  This article constitutes the third chapter of this PhD work.

- Monteiro Bastos da Silva [‡], J., **Chaker, J.** [‡], Martail, A., Costa Moreira, J., David, A.[‖,*], & Le Bot, B.[‖] (2021). Improving Exposure Assessment Using Non-Targeted and Suspect Screening the ISO/IEC 17025: 2017 Quality Standard as a Guideline. *Journal of Xenobiotics* **(IF=N/A)**, 11(1), 1-15.
  [‡], [‖] Both authors contributed equally.
  This article is mentioned in the Material and methods (second chapter) of this PhD work.

- **Chaker, J.**, Gilles, E., Leger, T., Jegou, B., & David, A.* (2021). From metabolomics to HRMS-based exposomics: Adapting peak picking and developing scoring for MS1 suspect screening. *Anal Chem* **(IF=6.8)**, 93(3), 1792-1800.
  This article constitutes the fourth chapter of this PhD work.

1.1.2. In preparation

- **Chaker, J.**, Gilles, E., …, David, A., A new suspect screening software for the rapid annotation of HRMS-based exposomics datasets. In preparation for submission to Analytical Chemistry

- **Chaker, J.**, Bonvallot, N., Warembourg, C., Chevrier, C., David, A., A suspect screening method for the comprehensive characterization of the chemical exposome in a Breton pre-teen population. In preparation for submission to Environment International.

1.2. As part of side projects

- David A., **Chaker J.**, Multigner, L., Bessonneau, V.*, Exposome chimique et approches « non ciblées » : Un changement de paradigme pour évaluer l'exposition des populations aux contaminants chimiques. Med Sci (Paris) 2021, 37 (10), 895-901.

- 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., Miller, G. W. (2021). Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **(IF=7.6)**, 156, 106630.

- David, A.*, **Chaker, J.**, Leger, T., Al-Salhi, R., Dalgaard, M.D., Styrishave, B., [...], Kristensen, D. M. (2021). Acetaminophen metabolism revisited using non-targeted analyses: Implications for human biomonitoring. *Environ Int* **(IF=7.6)**, 149, 106388.

- Christensen, S. L., Rasmussen, R. H., Ernstsen, C., La Cour, S., David, A., **Chaker, J.**, [...], Kristensen, D. M.* (2021). CGRP-dependent signalling pathways involved in mouse models of GTN- cilostazol- and levcromakalim-induced migraine. *Cephalalgia* **(IF=6.3)**, 41(14), 1413-1426.

- Rehfled, A., Frederiksen, H., Rasmussen, R.H., David, A., **Chaker, J**., Nielsen, B.S., Juul, A., Skaakebæk, N.E., Kristensen, D.M.* (2022), Human sperm cells can form paracetamol metabolite AM404 that directly interferes with sperm calcium signalling and function through a CatSper-dependent mechanism. *Human Reproduction* **(IF=6.9)***,* DOI : 10.1093/humrep/deac042

*corresponding author

2. Oral communications

2.1. International conferences

- **Chaker, J.***, Léger, T., Gilles, E., Jégou, B., Kristensen, D.M., David, A., Optimizing a suspect screening annotation workflow for large-scale application in human cohort, SETAC Europe 30th Annual Meeting, 3-7 May 2020, Online.

2.2. National and regional conferences

- **Chaker, J.***, Léger, T., Gilles, E., Jégou, B., Kristensen, D.M., David, A., Optimizing peak picking and development of a suspect screening workflow for a rapid annotation of the human chemical exposome from LC-HRMS datasets, Annual meeting of the doctoral school “Biologie-Santé”, 10-11 December 2020, Online.

- **Chaker, J.***, Léger, T., Gilles, E., Jégou, B., Kristensen, D.M., David, A., Développements analytiques pour la caractérisation non-ciblée de l’exposome chimique dans le sang

humain : challenges et perspectives, Junior researcher workshop, Société Francophone de Santé et Environnement, 04-05 November 2020, Online.

- **Chaker, J.***, Léger, T., Gilles, E., Jégou, B., Kristensen, D.M., David, A., Analytical developments for the non-targeted characterization of the human chemical exposome, Annual meeting of the public health doctoral network "Réseau Doctoral en Santé Publique", 10-11 June 2021, Rennes.

*presenting author

3. Posters

- **Chaker, J.***, Kristensen, D.M., Jégou, B., David, A., Développements analytiques pour la caractérisation non-ciblée de l'exposome chimique dans des matrices biologiques humaines, 12th Annual meeting of the Réseau Francophone de Métabolomique et Fluxomique, 21-23 May 2019, Clermont-Ferrand, France.

- **Chaker, J.***, Kristensen, D.M., Chevrier, C., Jégou, B., David, A., Assessing sample preparation methods for HRMS-based human chemical exposomics: the case of plasma and serum, 17th Annual Conference of the Metabolomics Society, 22-24 June 2021, Online.

*presenting author

4. Large-scale collaboration

- Participation to the NORMAN Network's first collaborative trial in biota (i.e. freeze-dried whole fish homogenate samples from a contaminated and a reference site). The efficiency of reference and in-house sample preparation methods as well as suspect and non-targeted screening workflows were compared between 16 labs. Two months of this PhD were dedicated to this task. A publication titled "*What's in the fish? Harmonization efforts in sample preparation methods for suspect and non-target screening in biota*" summarizing this collaborative trial's findings is in preparation. The implementation of the workflow developed in this PhD (from sample preparation to annotation) resulted in the most successful identification of spiked compounds for LC-HRMS among 16 participating laboratories, and was deemed of interest for further harmonization efforts.

5. Outreach and science popularization activities

- Scientific animation of the French School of Public Health (EHESP) stand on the chemical exposome at the Festival of Science, 05-06 October 2021, Rennes.

- Scientific animation of the joint French School of Public Health (EHESP) and Irset stand on the chemical exposome at the Festival of Science, 13-15 October 2020, Rennes.

- Scientific conception of the joint French School of Public Health (EHESP) and Irset stand on the chemical exposome at the Festival of Science, July-October 2020, Rennes.

- Scientific animation of the French School of Public Health (EHESP) stand on urban health at the Festival of Science, 6 October 2019, Rennes.

- Interviewed for the Swiss radio RTS's program "CQFD" titled "L'exposome ou comment l'environnement agit sur notre santé" regarding the concept of the exposome and its implementation. Broadcasted 15 May 2019.

6. Supervising and training

- Training of all new team members (5 postdoctoral researchers) to sample preparation, LC-HRMS analysis, data processing and annotation since 2019.

- Supervising of Ibrahim Maras during his master's degree internship, « Applications d'approches non-ciblées par UHPLC-ESI-HRMS pour caractériser l'exposition prénatale aux mélanges de xénobiotiques", March 1, 2020 – August 17, 2020.

- Training of Habiba Selmi during her master's degree apprenticeship, « Identification de métabolites de paracétamol issus du microbiote intestinal par analyses non-ciblées par UHPLC-ESI-HRMS et HRMS/MS » to non-targeted approaches (including sample preparation, LC-HRMS analysis, data processing and annotation), January 1, 2021-August 31, 2021.

- Training of Jaroslav Semerad (postdoctoral fellow from the Czech Academy of Science, Prague) to non-targeted data processing and annotation to characterize water samples. November 8, 2021 – December 3, 2021.

« Si le problème a une solution, il ne sert à rien de s'inquiéter. Mais s'il n'en a pas, alors s'inquiéter ne change rien. »

Proverbe tibétain

Résumé de la thèse en français

1. Introduction

Les maladies chroniques, telles que les cancers, les maladies cardio-vasculaires ou encore les diabètes étaient estimées responsables de 71% de la mortalité mondiale en 2018[1]. L'origine de la survenue de ces événements de santé multifactoriels a d'abord été investiguée au travers du Human Genome Project (HGP), qui a permis de procéder à un séquençage des 3 milliards de paires de bases du génome humain à la suite d'un effort international pendant 13 ans[2]. Ce projet a permis de mener des études d'association pangénomiques afin d'identifier des facteurs génétiques de susceptibilité à certains événements de santé[3]. Bien que plusieurs variants génétiques aient pu être associés à certains états de santé, il a également été constaté que les maladies considérées ne se déclenchaient que pour une partie des individus présentant ces variants[4]. Ce phénomène, appelé pénétrance, dépend de nombreux facteurs, tels que l'importance de la voie métabolique affectée, l'existence de voies métaboliques alternatives, ou encore les interactions avec l'environnement[4]. Dans ce contexte, Christopher Wild définit en 2005 le concept d'exposome, marquant le début d'un intérêt croissant de la communauté scientifique pour la caractérisation des liens existant entre les facteurs environnementaux et la survenue d'événements de santé défavorables, incluant les maladies chroniques[5]. L'exposome est alors définit comme étant l'ensemble des expositions environnementales (incluant des facteurs de style de vie), à partir de la période prénatale. En 2012, il étend cette définition pour prendre en compte les réponses biologiques (i.e. l'exposome interne) à ces facteurs environnementaux[6]. La caractérisation de l'exposome est donc une tâche complexe, puisqu'elle implique de capturer des facteurs de natures très diverses (socioéconomiques, physiques, biologiques, chimiques, etc.) et qui évoluent au cours de la vie. En pratique, il n'existe actuellement pas de moyen dynamique de mesurer l'exposome ; il est donc souvent entrepris de se concentrer sur des périodes particulièrement sensibles, telles que la période prénatale, l'enfance, l'adolescence, ou toute autre période d'intérêt vis-à-vis de l'événement de santé considéré. De plus, la caractérisation de l'exposome est souvent partitionnée en fonction de la nature des facteurs environnementaux considérés. Dans le cadre de cette thèse, c'est l'exposition humaine aux contaminants chimiques (i.e. les molécules exogènes dont des xénobiotiques), ou l'exposome chimique humain interne, qui est considéré, puisque cette exposition est fortement suspectée de contribuer à la survenue d'évènements de santé délétères.

La mesure de l'exposition des humains aux xénobiotiques se fait couramment de manière conventionnelle à l'aide d'approches dites « ciblées », qui permettent de générer des données quantitatives sur des listes préétablies de composés d'intérêts. Bien que ces méthodes soient extrêmement utiles pour évaluer l'exposition humaine à des composés supposés ou avérés

toxiques, elles peuvent être complémentées par des méthodes dites « non-ciblées ». Encouragées par le développement de technologies de pointe telles que la spectrométrie de masse à haute résolution, ces méthodes innovantes commencent à voir le jour pour investiguer l'exposition des humains aux xénobiotiques sans *a priori*. Ces nouvelles méthodes appliquées à des matrices biologiques permettent de profiler des milliers de molécules endogènes et exogènes simultanément sans avoir préalablement établi de liste de composés d'intérêts. Elles peuvent être utilisées à des fins exploratoires pour détecter et identifier de nouvelles molécules de synthèse qui arrivent nouvellement dans l'environnement en remplacement de celles considérées toxiques et dont l'usage devient restreint, ou qui ont été sous-investiguées jusqu'à présent[7]. Les méthodes non-ciblées reposant sur la spectrométrie de masse haute résolution impliquent dans la plupart des cas une technique séparative en amont pour décomplexifier les échantillons biologiques, telle que la chromatographie liquide. Ce couplage permet de générer différentes données chimiques caractérisant les signaux détectés, telles que le ratio masse/charge (m/z) auxquels sont associés un temps de rétention (Rt), et une abondance (e.g. aire) qui est propre à chaque échantillon analysé. Ces informations permettent de remonter à une élucidation structurale (i.e. à l'annotation), c'est-à-dire de les relier à une identité chimique par différents éléments de preuve. Depuis le début des années 2010, ces méthodes ont permis d'évaluer la présence de composés dans des matrices environnementales[8] et biologiques[9, 10].

Bien que très prometteuses concernant l'évaluation de l'exposition humaine aux contaminants chimiques, les méthodes non-ciblées sont toujours sujettes à plusieurs verrous méthodologiques et techniques. Tout d'abord, la large diversité de contaminants chimiques auxquels les humains sont potentiellement exposés implique que chaque choix méthodologique (e.g. technique analytique, préparation d'échantillons, etc.) imposera une limitation de l'espace chimique visible, qu'il convient de définir. En effet, il y a actuellement 111 millions de composés référencés dans la base de donnée PubChem[11] ; la diversité de leurs caractéristiques physico-chimiques (e.g. masse, polarité) explique l'impossibilité d'une part de les profiler avec une seule méthode, et d'autre part d'évaluer les performances de recouvrement et de sensibilité pour les composés détectables par la méthode considérée. Par ailleurs, la caractérisation de l'exposome chimique au travers de matrices biologiques est complexe, puisque ces matrices sont constituées de composés dans une large gamme de concentrations (du g/L pour certains composés endogènes au pg/L pour certains contaminants exogènes environnementaux). Or, ces différentiels de concentration peuvent induire des phénomènes tels que la suppression ionique, qui mène au masquage des composés peu abondants par des composés largement abondants. Il est donc impératif de développer des méthodes analytiques adaptées pour la détection de ces molécules exogènes dans les

matrices biologiques avec des approches non-ciblées (i.e. qui permettent d'éliminer suffisamment de composés matriciels en forte abondance). De plus, les outils bioinformatiques utilisés pour traiter les données non-ciblées ont, pour la plupart, été développés pour la métabolomique, qui s'axe sur l'étude des composés endogènes, qui peuvent être jusqu'à 10^{10} fois plus abondants en matrice biologique que les composés environnementaux[12]. Leur application pour l'identification de composés exogènes peu abondants peut donc être limitée. Enfin, le processus d'annotation est fastidieux et incomplet ; il consiste à rassembler des preuves de différentes natures pour valider une identité chimique pour un signal[13]. Ce processus inclut quasi-systématiquement une vérification manuelle pour éliminer les faux positifs, qui sont souvent nombreux. On estime aujourd'hui que moins de 10% des signaux identifiés sont annotés[14]. Ainsi, ces freins méthodologiques et technologiques doivent être surmontés pour obtenir des méthodes non-ciblées robustes, adaptées aux matrices biologiques, et adaptées aux applications à large échelle.

Dans ce contexte, ce travail de doctorat s'inscrit dans une dynamique visant à apporter, à terme, une réponse opérationnelle au concept d'exposome chimique dans le champ de la santé environnementale. L'objectif final est de pouvoir implémenter ces approches non-ciblées au sein d'études épidémiologiques à large échelle pour contribuer à l'identification de nouveaux mélanges ou de substances émergentes associés à certains évènements de santé. Ainsi, deux objectifs principaux ont été fixés pour ce travail : i) développer un workflow robuste de méthodes innovantes de production et de traitement de données analytiques non-ciblées, incluant la préparation d'échantillon, la méthode analytique de chromatographie liquide couplée à la spectrométrie de masse haute résolution, le traitement des données, et l'annotation, et ii) appliquer ces méthodes à plus large échelle sur 125 échantillons de sérum afin de permettre une évaluation de l'exposition chimique de 125 adolescents bretons.

2. Acquisition de l'empreinte chimique : optimiser l'équilibre entre sensibilité et sélectivité

L'acquisition de l'empreinte chimique a tout d'abord été optimisée. En effet, a caractérisation d'échantillons biologiques tels que le plasma ou le sérum dépend en partie du choix de la méthode de préparation d'échantillon. Ce choix est décisif, puisque les composés éliminés à cette étape initiale ne peuvent pas être récupérés par la suite. De plus, bien que les données chimiques générées pourront être ré-analysées à mesure que de nouveaux outils et algorithmes de traitement des données apparaitront, les échantillons biologiques ne sont disponibles qu'en quantités limitées ; leur préparation doit donc être optimisée initialement.

L'un des freins à l'étude de l'exposome chimique est la présence de certains contaminants à très faible dose dans le corps, et donc dans les échantillons biologiques. De ce fait, de hautes performances en sensibilité sont nécessaires pour caractériser plus exhaustivement l'exposome chimique. Or, les matrices biologiques sont complexes car constituées de composés endogènes en abondance, tels que les protéines et les phospholipides par exemple. Ces composés peuvent, de par leur concentration largement supérieure, limiter la détection des composés exogènes à cause de phénomènes tels que la suppression ionique. Ainsi, il est nécessaire de procéder à une purification de l'échantillon pour éliminer ces interférents analytiques, tout en conservant tous les analytes d'intérêt. Cet équilibre entre sensibilité et sélectivité doit donc être pris en compte lors de l'optimisation de la méthode de préparation d'échantillons.

Dans le cadre de cette thèse de doctorat, douze méthodes de préparation d'échantillons ont été évaluées pour la caractérisation de l'exposome chimique par des échantillons de plasma ou de sérum. Cette évaluation a reposé sur l'implémentation de critères complémentaires rarement utilisés pour l'évaluation des méthodes non-ciblées, à savoir des critères quantitatifs (e.g. taux de recouvrement, répétabilité, effet de matrice, etc.) systématiquement utilisés dans le domaine des analyses ciblées multirésidus, et qualitatifs (e.g. annotation, facilité et rapidité d'implémentation, etc.). Ces critères ont été définis dans le but de documenter au mieux le périmètre analytique observable de l'exposome chimique profilé avec chacune de ces méthodes. Cette délimitation des limites de ces méthodes est cruciale pour l'interprétation des jeux de données HRMS (e.g. aide à l'annotation). Ces méthodes reposent sur quatre principes de fonctionnement: l'élimination des phospholipides (sept méthodes), l'extraction en phase solide (trois méthodes), l'extraction liquide sur support (une méthode), et la précipitation de protéines (une méthode), classiquement utilisée en métabolomique. L'évaluation systématique de ces méthodes a été effectuée en utilisant un mélange de cinquante molécules sélectionnées pour leur diversité de caractéristiques physico-chimiques (i.e. masse, polarité), et leur appartenance à différentes classes chimiques susceptibles d'être présentes dans des échantillons dérivés de sang (i.e. composés endogènes, composés issus de l'alimentation, médicaments, pesticides, etc.).

L'évaluation systématique de ces méthodes de préparation a été effectuée en trois étapes. Tout d'abord, le mélange de molécules a été utilisé pour doper des homogénats de sérum à une concentration moyenne dans un contexte d'exposition (40 ng/mL). Le recouvrement, la répétabilité et l'effet de matrice a été évaluée pour les cinquante molécules et les douze méthodes. Ces premiers résultats ont permis de présélectionner la méthode de précipitation de protéines, une méthode d'élimination des phospholipides, ainsi qu'une méthode

d'extraction en phase solide, qui présentaient toutes des performances satisfaisantes sur tous les critères d'évaluation. La deuxième étape de l'évaluation a consisté en un dopage d'homogénats de sérum et de plasma avec le même mélange de molécules à une concentration plus faible (10 ng/mL). La fréquence de détection, le rapport signal/bruit, la répétabilité, la significativité du dopage (i.e. significativité de la différence d'aires entre échantillons dopés et non-dopés), et la facilité d'implémentation ont été évalués pour les trois méthodes évoquées, ainsi que pour une combinaison de la méthode d'extraction en phase solide et la méthode d'élimination des phospholipides. Cette deuxième étape a permis de démontrer que la précipitation de protéines et la méthode d'élimination des phospholipides permettaient toutes deux d'atteindre des performances supérieures aux deux méthodes impliquant l'extraction en phase solide, notamment sur les critères de répétabilité et facilité d'implémentation. Enfin, ces deux méthodes ont été appliquées sur les mêmes échantillons de cohorte (plasma et sérum) afin de les comparer en conditions réelles (i.e. sans dopage). Des composés exogènes ayant des caractéristiques physico-chimiques diverses ont été annotés, soulignant dans un premier temps la pertinence de ces deux méthodes de préparation pour caractériser l'exposome chimique. De plus, cette comparaison a permis d'observer la complémentarité de ces deux méthodes ; dans les deux matrices, plus de 40% des composés annotés n'étaient visibles qu'avec l'une des deux méthodes de préparation.

Cette approche d'évaluation systématique des méthodes de préparation d'échantillons pour la caractérisation de l'exposome chimique dans du plasma et du sérum humain a donc permis de documenter le périmètre de l'espace chimique détecté. Elle a également permis de démontrer la complémentarité de deux méthodes de préparation d'échantillons qui peuvent être utilisées conjointement au sein d'un workflow simple pour élargir l'espace chimique visible (jusqu'à 80% des marqueurs sont spécifiques à une méthode), et qui sera ensuite utilisé pour la suite des travaux de thèse. Après l'optimisation de l'acquisition de cette empreinte chimique, il est nécessaire d'évaluer les solutions de traitement des données disponibles. Un protocole de préparation d'échantillons impliquant ces deux méthodes a été proposé afin d'augmenter l'espace chimique visible.

3. Prétraitement des données et développement d'un logiciel de profilage de suspects

3.1. Adaptation des logiciels de prétraitement des données aux applications en exposomique

Suite à l'acquisition de l'empreinte chimique d'un ou de plusieurs échantillons, l'information chromatographique et spectrale générée doit être transformée en une liste de marqueurs caractérisés par un rapport masse/charge, un temps de rétention, et une aire par échantillon. Bien qu'il existe de nombreux outils de traitement des données non-ciblées, ils ont été, pour la plupart, développés pour des applications en métabolomique. Dans un contexte d'étude en exposomique, les composés d'intérêts sont souvent peu abondants ; il est donc critique de s'assurer que ces outils sont capables de les différencier du bruit. D'autre part, le processus d'annotation, souvent basé sur la liste de marqueurs générés précédemment, doit également être optimisés pour ces signaux peu abondants qui ne déclenchent pas systématiquement un acquisition MS2. L'objectif de ce chapitre est donc de sélectionner et optimiser l'outil adéquat pour améliorer l'efficacité de ce processus de traitement des données, à l'instar de ce qui a été fait pour les applications en métabolomique[15-17], mais qui n'a pour le moment jamais été fait pour des applications exposomiques.

Dans le cadre de ce travail, quatre outils de traitement des données ont été optimisés et comparés pour le traitement de données non-ciblées issues d'une application en exposomique. Deux de ces outils sont des logiciel vendeur (MarkerView de SCIEX et Progenesis QI for metabolomics de Waters), et les deux autres sont des outils open source fréquemment utilisés en métabolomique (MZmine2[18] et XCMS[19]). Ce travail d'optimisation et de comparaison a été effectué en utilisant les données issues du dopage à 10 ng/mL des échantillons de plasma et de sérum préparés par la précipitation de protéines. Chaque outil de traitement des données a tout d'abord été optimisé individuellement, manuellement et automatiquement si possible (i.e. paramétrage automatisée de XCMS par IPO[16] et Autotuner[15]), et les données issues du paramétrage optimisé pour chaque outil ont été comparées entre elles. Cette comparaison a été effectuée sur cinq critères : la fréquence de détection, le temps de calcul, la facilité d'implémentation, la répétabilité de l'intégration automatique, et la significativité de la détection (i.e. résultat du t-test comparant les aires associées aux composés dopants entre les échantillons dopés et non-dopés). Dans un premier temps, il a été démontré que l'utilisation d'outils automatisés de paramétrage développés pour la métabolomique n'était pas adaptée aux applications en exposomique. Ainsi, le paramétrage suggéré par IPO, basé sur les pics jugés « fiables » en fonction de leur

rapport $^{13}C/^{12}C$, a résulté en une largeur de pic trop élevée (30.7 s), menant à une détection de moins de 30% des composés dopés dans les deux matrices. A l'inverse, l'outil Autotuner a suggéré une largeur de pic trop faible (<10 s), qui a mené à une mauvaise performance en répétabilité (< 20% des composés avec une répétabilité satisfaisante) due à une scission excessive des pics détectés. L'optimisation manuelle a donc été préférée dans le cadre de l'application considérée. Il a dans un second temps été constaté que l'optimisation individuelle des outils permettait d'augmenter la fréquence de détection des composés de jusqu'à 60% (XCMS). En effet, certains paramètres comme la largeur de pic et le niveau de bruit généralement proposés par défaut ne sont pas applicables aux applications en exposomique, et doivent être réduits pour correspondre aux pics d'intérêt. De plus, bien que les outils open source permettent d'avoir beaucoup plus de libertés sur le choix des algorithmes et des paramètres utilisés, ils nécessitent une meilleure connaissance technique et présentent des temps de calcul 4 à 16 fois plus long que les logiciels vendeurs. Ainsi, tous les logiciels ont permis d'obtenir des performances satisfaisantes en termes de fréquence de détection, de répétabilité et de significativité de détection. Dans le cadre d'applications à large échelle, il peut être approprié de s'appuyer sur les logiciels vendeurs pour obtenir des résultats fiables plus rapidement. Il demeure cependant nécessaire de continuer à optimiser ces outils, car aucun d'entre eux n'a permis de détecter tous les composés dopés identifiés manuellement dans les chromatogrammes bien que ceux-ci présentaient des aires, profils isotopiques et profils MS2 fiables.

3.2. Développement d'un logiciel pour assister les approches de profilage de suspects

Les jeux de données obtenus suite au traitement des données chromatographiques et spectrales sont ensuite utilisés pour l'annotation. L'annotation de données HRMS non-ciblées peut être effectuée par à l'aide de deux stratégies majeures : le profilage non-ciblé, qui repose sur l'annotation de marqueurs priorisés car différenciants entre deux groupes, ou le profilage de suspects, qui repose sur l'annotation de marqueurs priorisés pour leur similitude avec des composés listés dans une librairie/base de données de suspects. Cette deuxième méthodologie est aujourd'hui très prometteuse, en partie car elle a un fort potentiel d'automatisation et permet de prioriser très rapidement des signaux d'intérêt. En effet, la comparaison de marqueurs et de suspects sur des éléments caractéristiques tels que leur rapport masse/charge ou leur profil de fragmentation MS2 peut être effectuée partiellement automatiquement, avant d'être validée manuellement dans la majorité des cas. Cependant, les composés d'intérêt généralement peu abondants en exposomique ne déclenchent pas systématiquement d'acquisition MS2, ce qui limite fortement le niveau de confiance de

l'annotation effectuée[13]. Dans ce contexte, un outil de profilage de suspects adapté aux données MS1 a été développé, et comparé aux outils de profilage de suspect existants (i.e. xMSannotator[20], MS-DIAL[21], msPurity[22] et MZmine2[18]). Ce nouvel outil repose sur la comparaison du rapport masse/charge, du profil isotopique, et de temps de rétentions expérimentaux ou prédits entre marqueurs et suspects, ce dernier prédicteur n'étant implémenté dans aucun autre outil. Ce logiciel permet aussi d'afficher un score de proximité appelé indice de confiance entre le marqueur et le suspect pour ces trois prédicteurs, ainsi qu'un indice de confiance global qui permet d'évaluer efficacement la plausibilité de l'annotation. Bien que la comparaison de ces outils ait été compliquée par la grande diversité de leur principe de fonctionnement, l'implémentation de l'utilisation de temps de rétention expérimentaux et prédits, ainsi que l'affichage des indices de confiance ont permis à notre logiciel de se démarquer des autres outils notamment en l'absence de données MS2. Une comparaison plus poussée avec MS-DIAL est proposée dans le chapitre application à large échelle. Ainsi, cet outil permet de prioriser efficacement les pré-annotations, qui doivent ensuite être validées manuellement. Cette priorisation permet d'effectuer un gain de temps considérable, qui pourrait contribuer à la plus large annotation des jeux de données non-ciblées existants. La pertinence de cet outil a été mise en avant lors de l'essai collaboratif NORMAN (meilleure fréquence de détection des composés dopés en matrice par ce logiciel) qui regroupait 16 laboratoires différents.

4. Application du workflow développé au sein de la cohorte mère-enfant Pélagie

L'intérêt croissant pour l'étude des liens entre expositions environnementales et santé a mené au développement et à l'optimisation de méthodes non-ciblées et de profilage de suspects pour caractériser l'exposome chimique interne humain. Les optimisations de méthodes effectuées dans le cadre de cette thèse ont ainsi permis d'améliorer leurs capacités de sensibilité; leur robustesse a également été vérifiée lors d'une application à plus large échelle. Ainsi, 125 échantillons de sérum sanguins issus de pré-adolescents (12 ans) bretons ont été analysés après leur préparation par deux méthodes de préparation d'échantillon, et dans les deux modes d'ionisation (positif et négatif), représentant ainsi 500 échantillons analysés (960 analyses au total en incluant les échantillons composites de contrôle qualité et les acquisitions MS2). Ces adolescents font partie de la cohorte Pélagie, qui a inclus environ 3500 femmes enceintes entre 2002 et 2005, toujours suivies avec leur enfant à l'heure actuelle. L'un des suivis a été effectué aux 12 ans des enfants, au cours duquel des paramètres cliniques tels que la croissance ou l'adiposité ont été vérifiés. Des échantillons sanguins ont été collectés pour, entre autres, évaluer l'exposition de ces adolescents aux contaminants organiques.

Quatre objectifs majeurs ont été établis pour ce chapitre : tout d'abord, évaluer la robustesse des méthodes analytiques et bioinformatiques optimisées dans le cadre de cette thèse. Ensuite, l'utilisation de prédicteurs MS1 (logiciel développé au laboratoire) et MS2 (MS-DIAL) pour l'annotation de xénobiotiques en matrice complexe a été comparée. Les expositions chimiques des pré-adolescents de la cohorte Pélagie ont subséquemment été caractérisées (n=92 annotations). Enfin, la complémentarité des deux méthodes de préparation d'échantillon utilisées conjointement comme recommandé dans un chapitre précédent a été étudiée à plus large échelle.

Lors de cette application à large échelle, des contrôles qualité (i.e. même échantillon composite injecté plusieurs fois intra- (n=11, dont 5 initiaux pour équilibrer le système) et interbacth (n=110 par méthode de préparation d'échantillons)) basés sur l'aire des marqueurs détectés dans les échantillons composites injectés à répétition au cours des séquences, et sur leur temps de rétention ont été mis en place afin de veiller à la comparabilité des échantillons. De même, la stabilité de l'aire et du temps de rétention des 22 standards internes dopés dans tous les échantillons (n=125 par méthode de préparation d'échantillon) et les échantillons composites injectés entre les échantillons ont été vérifiées, soit dans 310 échantillons au total. Ces vérifications ont permis de constater la nécessité de procéder à une normalisation de l'aire des marqueurs par le courant ionique total, qui présentait une variation batch-dépendante. Cette normalisation a notamment permis de baisser le coefficient de variation calculé sur les aires des marqueurs communs à 80% des marqueurs composites d'environ 35% par rapport à sa valeur brute pour les deux méthodes de préparation des échantillons, démontrant ainsi sa pertinence pour cette application.

Dans un second temps, les données obtenues ont été annotées par une approche de profilage de suspects à l'aide du logiciel développé, qui se base sur des prédicteurs MS1, et MS-DIAL, basé majoritairement sur des prédicteurs MS2. L'utilisation de ces deux outils a permis de comparer ces deux fonctionnements, et a permis de démontrer que l'utilisation de prédicteurs MS1 était pertinente et complémentaire à une approche basée sur la MS2 dans une application exposomique, où les données MS2 ne sont pas toujours de bonne qualité, voire inexistantes. Cependant, la curation manuelle nécessaire pour confirmer ces pré-annotations est plus importante, puisqu'elle implique de rechercher et comparer les motifs de fragmentation manuellement. Ainsi, certains composés n'ayant pas été fragmentés lors de l'acquisition MS2, tels que le pentachlorophenol ou le triclosan glucuronide, n'ont pas été annotés par MS-DIAL. Cependant, ces composés présentent des schémas isotopiques discriminants, ainsi que des valeurs de Rt prédits cohérentes avec les valeurs de Rt expérimentales (indices de confiance sur le Rt supérieurs à 0.84). Dans le cas du triclosan glucuronide, une indication

supplémentaire étayant l'annotation porte sur l'annotation d'un autre métabolite (i.e. triclosan sulfate) provenant du même composé parent (i.e. triclosan). Cette étape a donc également mené à la proposition d'une nouvelle version de la classification des niveaux de confiance des annotations proposée par Schymanski et al. (2014)[13]. Cette nouvelle version de la classification prend en compte les développements méthodologiques qui ont été effectués lors de cette thèse, tels que la vérification des ratios d'isotopologues, et ces dernières années, tels que les modèles de prédiction du temps de rétention[23-26], ou de prédiction de la fragmentation MS2[27, 28], qui permettent de générer des indices forts appuyant ou écartant l'annotation effectuée. Au total, 92 annotations ont été effectuées.

Les composés annotés se répartissent en quatre grandes classes : les métabolites de la flore intestinale (7%), les composés issus de l'alimentation (45%), les composés utilisés pour la santé et l'hygiène (18%, incluant 11% de principes actifs pharmaceutiques) et les composés industriels (30%, incluant 8% de pesticides et 8% de plastifiants). Ces composés présentent des caractéristiques physico-chimiques variées ($-2.7 \leq \log P \leq 16$, et $100.0754 \leq [M+H]^+ \leq 811.4913$), et des sources diverses, ce qui démontre qu'il est possible d'observer un large espace chimique avec les méthodes développées au cours de cette thèse. La détection de ces composés dans chaque échantillon a été évaluée. Il a été établi que les proportions de métabolites intestinaux et de composés naturels issus de l'alimentation étaient très peu variables entre les participants (coefficients de variation CV calculés sur les proportions sous 15% pour chaque classe et chaque méthode de préparation d'échantillons). A l'inverse, les expositions aux retardateurs de flammes organophosphorés (CV de 165% et 210% dans les échantillons PPT et Phree respectivement), aux intermédiaires de synthèse (CV de 115% et 27% dans les échantillons PPT et Phree respectivement) et aux pesticides (CV de 65% et 9% dans les échantillons PPT et Phree respectivement) sont hautement variables d'un individu à un autre. Ces observations sont cohérentes avec une exposition ubiquitaire aux métabolites intestinaux et aux composés naturels de l'alimentation, mais dépendante du style de vie (urbain ou rural, habitudes alimentaires, etc.) en ce qui concerne les polluants environnementaux. Certains pesticides (et métabolites) annotés, tels que le bromoxynil ou le fipronil sulfone, avaient déjà été détectés en population générale à de faibles niveaux (i.e. état de trace à 140 ng/mL)[29, 30]. Ces faibles niveaux documentés constituent une première indication (à confirmer avec des essais ciblés quantitatifs) sur les performances de sensibilité des approches développées au cours de cette thèse. Un métabolite du pesticide bromoxynil très largement détecté dans cette étude et auparavant jamais décrit dans des études de biosurveillance a été annoté. Ce métabolite est plus détecté que bromoxynil (97% contre 61%) et les aires observées dans les échantillons sont 3 à 8 fois plus élevées que celles du bromoxynil. Ces observations confirment ainsi la faisabilité d'utiliser des approches de

profilage de suspects pour identifier de nouveaux biomarqueurs d'exposition de composés d'intérêt.

Enfin, les deux méthodes de préparation d'échantillon utilisées dans le cadre de cette application à large échelle ont été comparées. Les rapports d'aires des composés annotés ainsi que de l'ensemble des marqueurs ont été calculés, et ont permis de déterminer que plus de 80% des marqueurs ne sont visibles que par l'une ou l'autre des méthodes de préparation. A l'échelle des composés annotés, plusieurs tendances observées ont permis d'émettre des hypothèses sur les facteurs influant sur la bonne détection des composés avec l'une des deux méthodes de préparation d'échantillons. Tout d'abord, les composés polaires sont généralement mieux détectés avec la précipitation de protéines, ce qui pourrait s'expliquer par le mécanisme d'action des plaques d'élimination des phospholipides, qui serait basé sur la rétention de la tête polaire des phospholipides[31]. A l'inverse, les composés plutôt apolaires sont mieux détectés dans les échantillons préparés par Phree. Cette observation est cohérente avec la présence importante de phospholipides et lysophospholipides dans les échantillons préparés par PPT, qui peut gêner l'ionisation d'autres composés moins abondants ayant un temps de rétention similaire (i.e. suppression ionique). Ensuite, les retardateurs de flamme organophosphorés sont mieux détectés dans les échantillons Phree, ce qui pourrait s'expliquer par le fait que ces plaques ne retiendraient que les groupes phosphates les plus polaires, tels que ceux qui forment la tête des phospholipides. Enfin, les phthalates semblent mieux détectés avec Phree, à l'exception d'un téréphthalate encombré stériquement, qui n'est pas strictement favorisé par une méthode. Cela pourrait s'expliquer par un mauvais recouvrement de composés encombrés stériquement par les plaques Phree.

Ainsi, ce chapitre a permis d'appliquer les méthodes développées au cours de cette thèse à large échelle, sur 125 échantillons de la cohorte bretonne Pélagie. Cette application a mené à l'annotation de 92 composés d'une grande diversité physico-chimique, qui contribue à la documentation du périmètre de l'espace chimique observable en utilisant les méthodes optimisées décrites. Les données obtenues pourront également être utilisées en association avec d'autres données contextuelles, telles que le lieu de vie ou les habitudes alimentaires, afin de prioriser d'autres marqueurs pour l'annotation avec une approche non-ciblée.

5. Conclusions et perspectives

La caractérisation de l'exposome chimique interne humain avec des approches non-ciblées offre de nouvelles promesses pour l'identification de nouveaux facteurs de risque chimique mais présente encore des obstacles technologique et méthodologiques qui doivent être surmontés. Ces limites viennent principalement du fait que les molécules exogènes sont

souvent présentes à l'état de trace dans des matrices biologiques complexes, et que les outils d'annotation automatisés n'existent pas encore. Ainsi, les méthodes classiquement utilisées en métabolomique doivent être adaptées et optimisées pour ces nouvelles applications nécessitant de meilleures performances en termes de sélectivité, sensibilité et robustesse. Ce travail de thèse présente l'optimisation des étapes méthodologiques les plus critiques pour implémenter des approches non-ciblées à large échelle basées sur la spectrométrie de masse à haute résolution. L'efficacité des méthodes optimisées dans le cadre de cette thèse pour caractériser l'exposome chimique interne humain a été démontrée. Ces approches constituent des atouts importants pour mieux comprendre l'effet de notre environnement chimique sur l'origine d'événements de santé. Elles génèrent un intérêt croissant aux échelles européenne et internationale, comme démontré par la création de l'infrastructure EIRENE par exemple. La mise en place de collaborations à cette échelle permettra de générer des données robustes et comparables entre les laboratoires pour décrire plus précisément et exhaustivement l'exposome chimique interne humain.

References

1. Bennett J.E., Stevens G.A.*, et al.*, NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4. *The Lancet* **2018,** *392* (10152), 1072-88.
2. Collins F.S., Morgan M.*, et al.*, The Human Genome Project: Lessons from Large-Scale Biology. *Science* **2003,** *300* (5617), 286-90.
3. Bodmer W.,Bonilla C., Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **2008,** *40* (6), 695-701.
4. Vineis P., Schulte P.*, et al.*, Misconceptions about the use of genetic tests in populations. *The Lancet* **2001,** *357* (9257), 709-12.
5. Wild C.P., Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **2005,** *14* (8), 1847-50.
6. Wild C.P., The exposome: from concept to utility. *Int J Epidemiol* **2012,** *41* (1), 24-32.
7. Pourchet M., Debrauwer L.*, et al.*, Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. *Environ Int* **2020,** *139*, 105545.
8. Nurmi J., Pellinen J.*, et al.*, Critical evaluation of screening techniques for emerging environmental contaminants based on accurate mass measurements with time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **2012,** *47* (3), 303-12.
9. David A., Abdul-Sada A.*, et al.*, A new approach for plasma (xeno)metabolomics based on solid-phase extraction and nanoflow liquid chromatography-nanoelectrospray ionisation mass spectrometry. *J Chromatogr A* **2014,** *1365*, 72-85.
10. Jamin E.L., Bonvallot N.*, et al.*, Untargeted profiling of pesticide metabolites by LC-HRMS: an exposomics tool for human exposure evaluation. *Anal Bioanal Chem* **2014,** *406* (4), 1149-61.
11. Kim S., Chen J.*, et al.*, PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **2019,** *47* (D1), D1102-D9.
12. Rappaport S.M., Barupal D.K.*, et al.*, The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* **2014,** *122* (8), 769-74.
13. Schymanski E.L., Jeon J.*, et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.
14. David A., Chaker J.*, et al.*, Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **2021,** *156*, 106630.
15. McLean C.,Kujawinski E.B., AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. *Anal Chem* **2020,** *92* (8), 5724-32.
16. Libiseller G., Dvorzak M.*, et al.*, IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **2015,** *16*, 118.
17. Alboniga O.E., Gonzalez O.*, et al.*, Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* **2020,** *16* (1), 14.
18. Pluskal T., Castillo S.*, et al.*, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**.
19. Smith C.A., Want E.J.*, et al.*, XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006,** (78), 779-87.
20. Uppal K., Walker D.I.*, et al.*, xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal Chem* **2017,** *89* (2), 1063-7.
21. Tsugawa H., Cajka T.*, et al.*, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.

22. Lawson T.N., Weber R.J.*, et al.*, msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Anal Chem* **2017,** *89* (4), 2432-9.
23. Aalizadeh R., Nika M.C.*, et al.*, Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *J Hazard Mater* **2019,** *363*, 277-85.
24. Cao M., Fraser K.*, et al.*, Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **2015,** *11* (3), 696-706.
25. Stanstrup J., Neumann S.*, et al.*, PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. *Anal Chem* **2015,** *87* (18), 9421-8.
26. Bonini P., Kind T.*, et al.*, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020,** *92* (11), 7515-22.
27. Allen F., Pon A.*, et al.*, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* **2014,** *42* (Web Server issue), W94-9.
28. Ruttkies C., Schymanski E.L.*, et al.*, MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **2016,** *8*, 3.
29. McMahen R.L., Strynar M.J.*, et al.*, Identification of fipronil metabolites by time-of-flight mass spectrometry for application in a human exposure study. *Environ Int* **2015,** *78*, 16-23.
30. Semchuk K., McDuffie H.*, et al.*, Body mass index and bromoxynil exposure in a sample of rural residents during spring herbicide application. *J Toxicol Environ Health A* **2004,** *67* (17), 1321-52.
31. Ahmad S., Kalra H.*, et al.*, HybridSPE: A novel technique to reduce phospholipid-based matrix effect in LC-ESI-MS Bioanalysis. *J Pharm Bioallied Sci* **2012,** *4* (4), 267-75.

General introduction

Chronic diseases are the leading cause of worldwide mortality and morbidity, representing an estimated 71% of all deaths globally in 2018[1]. For decades, the impact of genetic factors on the emergence of these diseases was investigated through major conceptual and technological developments in the genomics field. These developments were notably achieved through an international collaborative effort during the Human Genome Project (HGP) conducted between 1990 and 2003[2]. In reaching its goal of mapping the human genome, the HGP paved the way for the first genome-wide association studies (GWAS), aimed to establish associations between genetic variants (typically single nucleotide polymorphisms, SNP) and various traits. Despite the identification of highly prevalent SNP (i.e. presence in >5% of the population), their often low penetrance limited the applicability of GWAS alone to exhaustively elucidate the etiology of non-communicable diseases. In 2005, director of Leeds Institute of Genetics, Health and Therapeutics Christopher Wild underlined the necessity of considering environmental exposures to understand chronic disease etiology at the population level, thus introducing the concept of exposome to complement the genome[3]. The exposome was therefore defined as the totality of human environmental exposures from conception onwards, and was extended in 2012 to account for the biological effects resulting from these exposures[4]. Investigating the exposome is hence a complex task, as environmental exposures are both extremely variable in nature and through time. Environmental exposures can be classified in three main categories defined by Wild (2012)[4]: the general external exposome (i.e. social capital, stress, urban or rural environment, etc.), the specific external exposome (i.e. radiation, chemical contaminants, lifestyle factors, etc.), and the internal exposome (i.e. metabolism, gut microflora, ageing, etc.) These definitions are still discussed to account for emerging topics of interest, such as the transformation products of environmental chemicals in the body[5]. It is currently unfeasible to exhaustively characterize the exposome, due to the considerable number and diversity of environmental factors. Hence, investigating the exposome is fractioned in various subfields, including the socio-exposome focused on determinants such as socio-economic category and social inequalities[6], the physical exposome focused on factors such as radiation or noise[7] or the chemical exposome, encompassing chemical exposures that can accumulate in humans through food, medication, pesticides, etc.[5]. Exposure to chemical compounds can occur in various circumstances, counting domestic, industrial or agricultural use of these molecules. Investigating the chemical exposome can therefore be studied both through the analysis of environmental (i.e. water, air, dust, food, etc.) and human biological matrices (i.e. blood, urine, tissue, hair, etc.). However, due to the high diversity of compounds constituting our chemical environment (i.e. tens of thousands), there is still a sore lack of data regarding human exposure. Acquiring broader knowledge on the human chemical exposome is therefore a first necessary step in accurately assessing its effect on human health.

Investigating the human chemical exposome in biological matrices has classically been done using targeted methods, which offer quantitative data on a set list of compounds of interest identified prior to the analysis. While these methods are exceptionally useful and robust to generate exposure data for already known or suspected toxicants, they can now be complemented by non-targeted methods, which allow the characterization of samples through collection of qualitative or semi-quantitative data without an *a priori* list of investigated compounds. These non-targeted approaches may be used as an exploratory tool to detect and identify new chemicals that might be of emerging concern, whether because they have newly appeared in the environment as a replacement to regulated substances or because they are newly identified or suspected toxicants[8]. They may also be useful to describe more thoroughly chemical mixtures, which are a well-documented challenge in exposure science[9], and therefore provide relevant data for toxicological tests. Most non-targeted methods rely on the recent technological progress in the field of high-resolution mass spectrometry (HRMS), resulting in the possibility of screening thousands of compounds simultaneously, with a high mass accuracy. Concurrently, significant progress in the bioinformatics field allowed the processing of such complex data. Compounds detected throughout the analysis can thus be isolated, characterized by their mass-to-charge ratio (m/z), their retention time (Rt) and their area, and annotated (i.e. associated to a chemical identity through various elements of proof). During the first half of the 2010s, these approaches have started to be used to assess the presence of contaminants in environmental matrices[10-12], or exogenous compounds in biological matrices, both animal[13-15] and human[16].

Non-targeted approaches, while valuable and increasingly used, are still subject to a number of technological barriers and methodological issues. Firstly, as there are no predefined analytes, method performances regarding recovery and sensitivity cannot be determined for all potentially detectable compounds. Moreover, it appears unreasonable to expect the exhaustive characterization of a sample, even by such methods; it is therefore necessary to delineate the width of what is observable using any particular workflow. Secondly, existing data processing tools were mostly built for metabolomics applications, i.e. the detection of endogenous (and often rather abundant) compounds, and may not be suitable for exposomics applications aimed to detect exogenous chemicals present at trace levels (below ng/ml). Thirdly, annotation is an often tedious and incomplete process, as it is estimated that less than 10% of non-targeted datasets are annotated[5]. This is further exacerbated for exposomics applications due the limited availability of compound libraries including or dedicated to exogenous chemicals. This process is time-consuming largely due to the necessity of manual curation to dismiss the usually high number of false positive annotations. All of these

technological and methodological bottlenecks must be overcome to create robust non-targeted workflows that may be used for high-throughput applications.

The main aim of this PhD work is to develop an HRMS-based non-targeted workflow applicable to human epidemiological studies, in order to provide an operational solution to assess human exposure to complex chemical mixtures at a large scale. Given the above-mentioned considerations, two specific objectives were defined for this PhD. The first objective is to develop innovative methods to generate and process non-targeted data, including sample preparation, analytical HRMS method(s) coupled to liquid chromatography (LC), data processing, and annotation. These methods must answer the need for sensitivity, robustness, and must be relevant in the case of human blood plasma and serum analysis. The second objective of this work is to apply these developed methods for non-targeted approaches on large-scale epidemiological applications to test the robustness and sensitivity and detect new biomarkers of exposure. This application was performed using samples from a promising local cohort. Blood serum samples from 125 12-year-old boys issued from the Breton mother-child cohort Pélagie were used to implement this large-scale application. This cohort, started in 2002, is a longitudinal study implemented to measure exposure to organic pollutants during the pregnancy. It included approximately 3,500 women pregnant between 2002 and 2005 in Brittany. Follow up was carried out at birth, and then at 2, 6, and 9-16 years old, through the collection of biological samples and clinical data, and answering questionnaires. A questionnaire was provided to 12-year-olds and their families to obtain physical growth data and pubertal stage. A clinical evaluation was performed on a subset of 500 12-year-olds, with the assessment of clinical parameters such as growth, adiposity, visual function and oral-dental health. The considered blood samples were collected at this time to evaluate sex hormones and to assess exposure to organic contaminants. This cohort, in its entirety, therefore offers a promising opportunity to study the long-term consequences of early-life exposure to environmental contaminants.

To reach the first objective, each step of the non-targeted workflow was optimized. Indeed, reference protocols for the preparation and high-throughput injection of plasma and serum samples were established and validated using new quantitative and qualitative criteria to define the perimeter of the profiled chemical exposome. Moreover, in-house libraries were constructed to implement suspect screening approaches, consisting of an *a posteriori* screening of suspected xenobiotics in chromatograms. Concomitantly, a software was developed to partly automatize suspect screening approaches through the implementation of confidence indices, scoring proximity between experimental features and suspects.

Reaching the second objective was achieved by using the previously described methods and tools (initially developed at the batch level) in the case of a high-scale application. Additionally, large-scale quality controls and inter-batch correction were implemented to ensure comparability from first to last sample.

Chapter 1 describes the state-of-the-art regarding the application of HRMS-based exposomics to cohort-based epidemiological studies. Reported technological and methodological challenges regarding the application of such approaches are detailed and discussed.

Chapter 2 relates the instrumental method development, the data processing steps, as well as the annotation tools needed for this work. The suspect screening software developed in the context of this PhD is also thoroughly described.

Chapter 3 presents the systematic evaluation and comparison of sample preparation methods for the purpose of detecting low-abundant chemicals in blood plasma and serum samples. The impact of the two best-performing methods on the visible chemical space is described using cohort plasma and serum samples.

Chapter 4 details the optimization of the data processing step to accurately transform LC-ESI-HRMS data to a list of features when compounds of interest are lowly abundant. Suspect screening tools including the in-house software were compared on cohort samples.

Chapter 5 documents the large-scale application of the optimized non-targeted workflow on 125 serum cohort samples. The use of MS1 and MS2 predictors for annotation is compared and discussed. The identification of markers of exposure is described, and results are discussed in light of the use of two sample preparation methods.

The last chapter is dedicated to the conclusion and perspectives of this work.

References

1. Bennett J.E., Stevens G.A.*, et al.*, NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4. *The Lancet* **2018,** *392* (10152), 1072-88.
2. Collins F.S., Morgan M.*, et al.*, The Human Genome Project: Lessons from Large-Scale Biology. *Science* **2003,** *300* (5617), 286-90.
3. Wild C.P., Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **2005,** *14* (8), 1847-50.
4. Wild C.P., The exposome: from concept to utility. *Int J Epidemiol* **2012,** *41* (1), 24-32.
5. David A., Chaker J.*, et al.*, Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **2021,** *156*, 106630.
6. Senier L., Brown P.*, et al.*, The Socio-Exposome: Advancing Exposure Science and Environmental Justice in a Post-Genomic Era. *Environ Sociol* **2017,** *3* (2), 107-21.
7. Van Kamp I., Persson Waye K.*, et al.*, Early environmental quality and life-course mental health effects: The Equal-Life project. *Environ Epidemiol* **2022,** *6* (1), e183.
8. Pourchet M., Debrauwer L.*, et al.*, Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. *Environ Int* **2020,** *139*, 105545.
9. Barouki R., Audouze K.*, et al.*, The exposome and toxicology: a win-win collaboration. *Toxicol Sci* **2021**.
10. Nurmi J., Pellinen J.*, et al.*, Critical evaluation of screening techniques for emerging environmental contaminants based on accurate mass measurements with time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **2012,** *47* (3), 303-12.
11. Rostkowski P., Haglund P.*, et al.*, The strength in numbers: comprehensive characterization of house dust using complementary mass spectrometric techniques. *Anal Bioanal Chem* **2019,** *411* (10), 1957-77.
12. Delaporte G., Cladiere M.*, et al.*, Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: Feasibility study on tea. *Food Chem* **2019,** *277*, 54-62.
13. David A., Abdul-Sada A.*, et al.*, A new approach for plasma (xeno)metabolomics based on solid-phase extraction and nanoflow liquid chromatography-nanoelectrospray ionisation mass spectrometry. *J Chromatogr A* **2014,** *1365*, 72-85.
14. Bonnefille B., Gomez E.*, et al.*, Metabolomics assessment of the effects of diclofenac exposure on Mytilus galloprovincialis: Potential effects on osmoregulation and reproduction. *Sci Total Environ* **2018,** *613-614*, 611-8.
15. Gomez E., Boillot C.*, et al.*, In vivo exposure of marine mussels to venlafaxine: bioconcentration and metabolization. *Environ Sci Pollut Res Int* **2021,** *28* (48), 68862-70.
16. Jamin E.L., Bonvallot N.*, et al.*, Untargeted profiling of pesticide metabolites by LC-HRMS: an exposomics tool for human exposure evaluation. *Anal Bioanal Chem* **2014,** *406* (4), 1149-61.

Chapter I. Application of HRMS-based exposomics to cohort-based epidemiological studies: state-of-the-art and challenges

1. Studying the human internal chemical exposome: context, definitions and challenges

1.1. The Exposome: from a concept to a call to action

In 1985, chancellor of the University of California Robert Sinsheimer first discussed the possibility of sequencing the human genome, which led to the first funding of research dedicated to genome sequencing in 1986. Four years later, the Human Genome Project (HGP) was launched with the objective of sequencing the entirety of the human genome[1]. The mobilization of over 2,800 researchers from the international scientific community and approximately 4 billion euros over thirteen years allowed reaching the set goal of sequencing the 3 billion base pairs of the human genome[2]. In 2001, Francis Collins, director of the National Human Genome Research Institute, declared about this near-exhaustive vision of the human genome: "It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease."[3]. Understandably, such a tremendous advancement in knowledge on human biology held great promises for a better understanding of disease etiology.

This incredible international effort was accompanied by constant methodological and technological progress. Indeed, the appeal of the HGP federated efforts to develop new high-throughput technologies and new computational strategies[1]. This later proved to be a crucial advantage when the knowledge generated by the HGP was used to identify genes affecting susceptibility to specific diseases. Genome-wide association studies (GWAS) were designed to identify variants associated with multifactorial diseases (with frequencies ≥ 5%)[4]. While many variants associated with different diseases have been identified so far, their often-low penetrance (i.e. the fact that only a small proportion of individuals presenting the variant develop the corresponding phenotype) limit the practical applications of GWAS[5]. It should be noted that the investigation of particularly infrequent single nucleotide polymorphisms (<1%) may still uncover valuable results, even though it would require a high number of participants (and therefore important resources) to achieve satisfactory statistical power. Variant penetrance is a complex characteristic that depends on many factors (i.e. interaction with other genes, importance of the affected pathway, existence of alternative pathways substituting for function loss, etc.), one of which is the interaction with the environment[6].

In this context, Christopher Wild introduced in 2005 the concept of exposome to account for the impact of environmental factors on human health through the genetic-environmental interactions[7]. The conceptualization of an exposome to complement the genome helped to

emphasize the need for reliable exposure assessment tools to better understand disease etiology through a more thorough description of the interplays between environmental exposures and genetic susceptibility factors. He defines the exposome as "life-course environmental exposures (including lifestyle factors), from the prenatal period onwards". In his editorial, he underlines the need for the funding and development of reliable exposure assessment tools to "balance the effort going towards characterization of the genome", and for a strong collaboration between scientists of different backgrounds as was done for the HGP[7]. In 2012, the definition of the exposome concept is expanded to take into account the biological responses to environmental exposures[8]. As of today's most widely accepted definition, the exposome is "an entity that encompasses all life-course environmental exposures and the associated biological responses, including during the prenatal period"[7-11]. One significant aspect of the exposome is the chemical exposome, i.e. the exposure to all chemicals, whether from external or internal sources[12].

Characterizing the chemical exposome is an arduous challenge. Indeed, it has been estimated that up to 350,000 chemical compounds and mixtures are registered for production and use worldwide, with up to 120,000 of them being either unknown or ambiguously defined[13]. As of 2020, there were close to 23,000 compounds registered by the European Chemical Agency, more than 2,000 of which are produced over the 1,000t/year limit[14]. The organic compounds most frequently registered are mostly registered as synthesis intermediates (e.g. styrene, ethylbenzene). This diversity of compounds, coupled to the diversity of potential sources for each compound present important hindrances to exhaustively characterize one's chemical exposures. Human exposure to some persistent organic compounds, such as organochlorine insecticides (e.g. DDT and its metabolites), polychlorinated biphenyls (e.g. PCB 153), brominated flame retardants (e.g. BDE 47, BDE 99), or polycyclic aromatic hydrocarbons (e.g. naphthalene and metabolites) have already been well reported in large-scale HBM studies[15-18]. These compounds have historically been studied for their widespread use, their potential or confirmed toxicity, or their persistence in the environment. Non-persistent compounds such as phthalates or bisphenols, however, are more challenging to accurately describe since their half-life in the human body is limited (a few hours to a few days). Moreover, their metabolization and excretion pathways may not be entirely documented, which may affect the ability to detect these compounds in their relevant forms[19]. Overall, the available data on human exposure to chemicals is limited and mostly oriented towards lists of hundreds of "usual suspects" (i.e. priority substances with already known exposure and toxicity data). This partial view of the human exposure to chemicals (few hundreds as opposed to tens of thousands on the market) undoubtedly leads to an underestimation of the chemical risk evaluation.

Exogenously derived chemicals and their biotransformation products accumulating in human will further be referred to as the internal chemical exposome, and will be distinguished from endogenously derived chemicals that constitute the metabolome. Many of these endogenous compounds, while also important to assess the impact of environmental chemical exposures on human health, are usually largely more abundant in biological matrices compared to exogenous compounds[20], and can be studied using differently optimized metabolomics/lipidomics workflows. A schematized representation of the human internal chemical exposome is represented in Figure I.1.

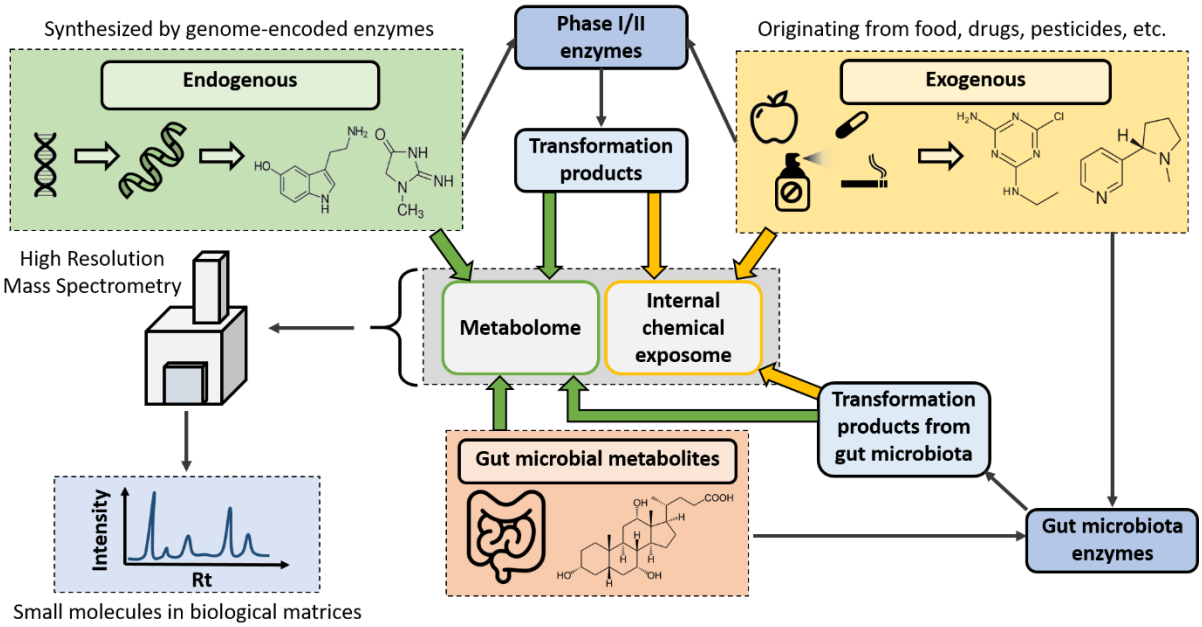

Figure I.1 – Schematized representation of the distinction between endogenous metabolites and exogenous chemicals and related biotransformation products. These small molecules (50–1200 Da) present in human biological matrices can be profiled using High Resolution Mass Spectrometry.

Traditionally, exposure assessments to chemicals have been performed through targeted approaches. These approaches rely on pre-established lists of compounds of interest (for their ubiquity, their high toxicity or both) and developing methods to quantify them in any given matrix of interest. Targeted assays result in highly accurate and robust quantitative data, with limits of detection often being as low as the ng-pg/mL range in complex matrices such as urine[21-23] and blood serum[24, 25]. Together with toxicological and other biological approaches, targeted methods have allowed limiting human exposure to toxic compounds, such as plasticizer bisphenol A[26] or pesticide atrazine[27] through public health measures either limiting or outright banning their use. These approaches were the first to allow the acquisition of HBM data at the massive scale needed to efficiently support policy making, as was started with the

priority lists established through the collaboration of the HBM4EU consortium and a European Union policy board[28].

While exceptionally useful, targeted approaches only allow accounting for already established compounds of interest. Indeed, the inclusion of a given compound in a targeted method must be preceded by the expectation that it is either ubiquitous or toxic enough to warrant medium to large-scale biomonitoring, given that there are an estimated >350,000 compounds currently in use in the human population[29, 30]. This ever-expanding list of diverse chemicals must be prioritized in order to identify chemicals of emerging concern (CEC) and launch the process of toxicological assays and targeted method development. Hence, the technological advancements of the last few years in high resolution mass spectrometry (HRMS)-based analysis has offered new possibilities to tackle the complexity of the chemical exposome. This may be achieved using new non-targeted approaches (NTA), which are complementary to targeted approaches[31] and do not rely on pre-established chemical lists. Through the technological progress achieved notably in HRMS, it is possible to simultaneously screen tens of thousands of small molecules (between 50-1200 Da) in a single analysis. NTA often uses a separative technique prior to HRMS analysis to decomplexify the sample, such as liquid chromatography (LC). These analyses result in lists of signals, called features, each characterized by a mass-to-charge ratio (m/z), a retention time (Rt), and an area. The data acquired during NTA is used to assign chemical identities to the obtained features, allowing potentially identifying new compounds of interest due to high detection frequencies and/or association to a health event. These approaches have already been successfully applied in proof-of-concept studies[23, 32-34], thus demonstrating the relevance and applicability of environment-wide associated studies (EWAS).

Another challenge inherent to the characterization of the exposome, including the chemical exposome, its dynamic nature. Indeed, the temporal variability of chemical exposures constitutes, along with its vast scope, incredibly challenging features of its characterization (for targeted as well as non-targeted approaches). Firstly, the dynamic nature of the chemical exposome entails that its measurement should be dynamic as well, either through an inherently dynamic measurement method or through a series of snapshots at crucial times in an individual's lifetime. This second approach can be applied at key times of life, such as the prenatal period, childhood, puberty and reproductive years, to allow a vision of presumably radically different exposure patterns throughout an individual's life[35].

The prenatal period is a well-known time of vulnerability in one's life. The DOHaD (Developmental Origins of Health and Disease) hypothesis, originally formulated by Barker and Osmond (1986)[36], postulates that nutrition during pregnancy could impact disease

outcome during the lifetime. This concept was expanded to take into account exposure to environmental chemical contaminants during the prenatal period, as evidence of their impact on health endpoints such as obesity arose[37]. Chemical exposures are therefore often investigated during this time to improve knowledge on disease etiology[38-40]. Another period of vulnerability in an individual's life is the transition into adolescence[41]. Indeed, as it is a transitional stage of development (physical, psychological, etc.) implicating significant hormonal activity, the impact of environmental chemicals (and in particular endocrine disruptors) on teenagers' health has been questioned[42].

Despite the many promises held by NTA as an exploratory tool to better understand environmental triggers to chronic diseases, several methodological and technological barriers remain to uncover their full potential. Notably, the still vast scope of the chemical exposome entails the need to determine the impact of matrix and analytical platform choice on the visible chemical space when using NTA.

1.2. Main conceptual challenges for the non-targeted characterization of the human chemical exposome: study design questionings

1.2.1. Direct and indirect measurements: choosing between environmental and biological matrices

Given the complexity of the human chemical exposome, designing a study for its non-targeted characterization raises several questions. Firstly, the human chemical exposome can be characterized through conceptually different approaches. Indeed, direct and indirect measurements are available to this end. Direct measurements consist in screening for chemicals directly in the considered individuals, as for example through biomonitoring[22, 43], while indirect measurements rely on studying the environment, and coupling this data to bioaccessibility studies and/or time of contact data to estimate human exposure[44, 45]. Indirect measurements allow identifying sources and determinants of exposure. They present the advantages of being less invasive, less costly, suitable for passive sampling (thus being more representative on the dynamic aspect of exposure), and using overall less complex matrices than direct measurements. However, they may only approximate the actual human exposure to chemicals. This may be due to the use of mathematical models with inherent uncertainty, or the inexact accounting for a significant source of exposure (whether under- or over-estimated)[46, 47]On the other hand, direct measurements allow evaluating the exposure as a whole, regardless of the sources and routes of exposure. Although the implementing of direct measurements is limited by their more limited cost-effectiveness (usually requiring higher funding and more long-term compliant participants)[48], biomonitoring is widely recognized as a

useful tool for exposure and risk assessment[49, 50]. In this PhD, a biomonitoring approach will be used to contribute to decipher the human internal chemical exposome using HRMS-based methods.

Biomonitoring studies have been widely used as a tool of choice to assess human exposure to environmental chemicals. Unprecedented levels of funding at national and EU levels are currently being implemented to provide novel human exposure data to chemicals through biomonitoring studies. At the national level, the French agency for public health Santé Publique France has led the French HBM program since 2010. This initiative aims to paint a representative image of the French population's exposure to chemical compounds, through the analysis of urine, blood and hair samples. This program consists in two surveys: a subset of the French Elfe cohort (>4100 individuals) as a perinatal component, and the Esteban project, which in general population-based (18-74 years). The data generated by this program is made available to research teams, notably those working on establishing exposure-health associations in the Elfe cohort. Furthermore, this data helps inform the relevant authorities regarding the determined environmental substances[51]. Conjointly with Anses (French Agency for Food, Environmental and Occupational Health & Safety), Santé publique France (SpF) has also launched the PestiRiv project in 2021. This initiative is geared towards the assessment of pesticide exposure for citizens residing in proximity to vineyards. Its main objective is to determine whether the proximity to agricultural land, particularly vineyards, has an effect on pesticide exposure. This may lead to the establishment or modification of public health measures to implement appropriate measures to protect citizen's health. Multiple sources will be accounted for (e.g. air, food, domestic use and profession), and both biological (i.e. urine and hair) and environmental (i.e. air, dust, food) will be collected.

Other sizable HBM studies (detailed in paragraph I.3.1) have gradually been undertaken in the last decade. At the European scale, projects such as HELIX, EXPOsOMICs (both started in 2012), HBM4EU (started in 2017), ATHLETE and EXPANSE (both started in 2020) have used HBM as a key tool to assess individuals' exposure to environmental chemicals. This growing implementation of large-scale HBM studies helps informing researchers and policymakers on the exposure-health relationship.

Overall, direct and indirect approaches are highly complementary and may also be used successively to obtain orthogonal data. For instance, an initial HBM approach may help identify chemicals of interest, and a following indirect measurement approach may allow identifying sources of exposure. Combining the data collected from these approaches may be critical in implementing new relevant public health measures to dampen the health burden of the chemical exposome.

As a part of this large scale collective effort to characterize the human internal chemical exposome[5], this PhD work focuses on implementing direct measures for the non-targeted characterization of the human chemical exposome.

1.2.2. Choosing the biological matrix

When aiming to directly characterize the human chemical exposome, the choice of biological matrix is the second study design element that should be clarified. Many factors can influence the choice of biological matrix: availability, invasiveness and cost of sampling, possible focus on some chemical classes with specific characteristics (e.g. persistence, accumulation in a specific biological compartment, etc.), etc.

One of the most commonly sampled matrices in HBM and epidemiological studies is urine[28, 52-54]. Its sampling is fairly non-invasive and inexpensive, and is easily performed by the participants themselves. Urine is a relevant matrix for exposure assessment as it is the main route of excretion of many non-persistent chemicals, whether in their free form or after phase I and/or phase II metabolization to increase polarity. One of its main drawbacks is fact that only short-term exposure (usually hours or days depending on the chemical's half-life) is visible when using this matrix, with often different forms of the chemical visible at different points in time[19, 55]. The visible window may be widened using pooled repeated measurements, which may be best to capture the dynamic nature of the exposure[11], as was described in the European projects HELIX and EuroMix for the assessment of exposure to phthalates and phenols in urine samples[56, 57]. Another well-known issue when using urine is the need for normalization (often using the creatinine level), as sample volume and chemical concentration may be extremely variable depending on the individual's hydration state[58]. Lastly, this matrix is not the most suitable for the detection of exposures to persistent organic pollutants, which tend to accumulate in other matrices such as blood and hair[55, 59], although the metabolites of these compounds may be found in urine[28].

Blood-derived matrices (i.e. total blood, plasma and/or serum) [53, 60-62] are also commonly sampled in HBM and epidemiological studies, and are therefore frequently available in biobanks. Their sampling is more costly and more invasive than urine, but blood-derived matrices are often considered the golden standard to study chemical exposure. One advantage of blood-derived matrices is that the biologically active parent (i.e. non-metabolized) form of chemicals might be in some cases more readily observable than in urine, which can be an advantage considering the sometimes non-specific nature of metabolites[63]. This was applied in the HBM4EU initiative with, for instance, the biomonitoring of parent halogenated flame retardants in serum, and of four metabolites of organophosphate flame retardants in

urine[64]. However, parent and phase I/II metabolite concentrations in blood are often lower than metabolite concentrations in urine in which they can accumulate over time[65]. Another advantage of blood-derived matrices is that blood circulates in the whole body and is in equilibrium with all tissues, and thus provides a more accurate reflection of internal chemical concentration[58]. In the case of a pregnancy, maternal blood is also in contact with the fetus through the placenta, which is why maternal blood may be relevant to evaluate fetal exposure during the prenatal period[66, 67]. Other matrices such as placenta, cord blood or meconium are also well suited for this purpose[67-69], but their limited quantity and availability is an important hindrance. Maternal hair was also reported to be a suitable matrix to assess prenatal exposure especially for persistent organic pollutants (POPs) [67, 70, 71], although several concerns regarding external pollution and lack of reference data are often put forward[22, 28, 58].

As no matrix will be ideal in every situation depending on target compound class, availability and ease of sampling, it should be understood that its choice will affect the observable internal chemical exposome. In the context of this work, blood-derived samples (i.e. plasma and serum) were used due to the advantages presented by these matrices, as well as for their availability in general in biobanks, and more particularly in the considered epidemiological studies (i.e. Pelagie). This PhD work is one of the first applications of HRMS-based characterizations of the internal chemical exposome in blood[32].

1.2.3. Analytical platform choice

Analyzing biological samples to characterize the human internal chemical exposome can be done using many platforms, most of which rely on chromatography (such as gas chromatography (GC) and liquid chromatography (LC)) coupled to HRMS. The breadth of the chemical exposome due to the ever-expanding number of produced chemical compounds (growth estimated at 3.4% each year until 2030[30]) implies the need to detect compounds with vastly different physical-chemical properties (e.g. polarity). At this time, no single technology allows capturing this diversity; ideally, complementary analytical platforms should be combined to observe the width of the chemical space[5, 72-75]. This is however challenging, since when aiming for large-scale applications such as epidemiological studies, analysis should be as not too expensive and high-throughput as possible to allow analyzing sufficient numbers of sample for statistical power, which is undeniably more difficult to achieve when multiple analytical platforms are involved. The choice of analytical platform(s) will therefore affect the observable chemical exposome, as represented in Figure I.2.

To date, the most commonly used platforms for NTA are equipped with LC, electrospray ionization (ESI) and coupled with time-of-flight (TOF) or Orbitrap analyzers[5, 76]. Hybrid

analyzers such as quadrupole-time-of-flight (QTOF) or quadrupole-Orbitrap (Q-Exactive family) are also frequently used and are important to provide relevant MS2 data[77]. LC-ESI-HRMS platforms are highly versatile and provide a soft ionization[5], which is useful to provide information on the molecular ion and avoid compound fragmentation and obtaining pseudomolecular ion mass[77]. However, the ionization process using ESI sources leads to less reproducible fragmentation patterns, making the construction of reference spectral libraries challenging, and in turn affecting the complexity of compound annotation[78]. LC separations can be performed using a large diversity of stationary and mobile phases, although reverse-phase (RP) columns are often used for their versatility and for easier comparison and harmonization between laboratories. Indeed, RP columns allow the simultaneous detection of compounds with a wide polarity range, such as the polar nicotine metabolite cotinine and the non-polar insecticide chlorpyrifos. Hydrophilic interaction chromatography (HILIC) is also emerging since it offers better performance for highly polar compounds such as pesticide glyphosate and antiviral acyclovir, thus providing orthogonal data to RP chromatography[79]. Two-dimensional chromatography combining HILIC and RP has been used to widen the observable polarity range[80, 81]. Regarding mobile phases, generic methanol/water or acetonitrile/water gradients are commonly used[77, 82] to avoid further limiting the range of observable compounds. The main disadvantages of LC-based platforms are the matrix-related issues such as ion suppression[83].

GC-HRMS platforms have been increasingly used to detect non-polar semi-volatile to volatile compounds such as POPs[28, 84-86], which are not detected using LC-ESI-HRMS. Characterizing the chemical exposure to POPs, which notably include polychlorinated biphenyls and organochlorine pesticides, is particularly relevant, as it has been linked to detrimental health effects such as endocrine disruption, cardiovascular and reproductive diseases, and cancer, in part linked to their bio-accumulative, toxic potential and non-degradable nature[87]. These characteristics also explain their presence in biological and environmental matrices several years or decades after banning. GC-HRMS platforms predominantly use hard ionization sources (i.e. Electron ionization), which often lead to the fragmentation of the molecular ion and the need for large spectral libraries for compound annotation[78]. The choice of stationary and mobile phases is far more limited in GC-based platforms, with a widespread use of nonpolar capillary column with 5% phenyl methylpolysiloxane and helium as carrier gas. While nitrogen and hydrogen can also be used as mobile phases since they are less expensive than helium, they are usually set aside due to efficiency and safety reasons respectively[76]. While GC-based platforms suffer less matrix effect than LC-based platforms, additional sample preparation steps such as derivatization are often required to improve versatility and avoid premature clogging of the column due to non-volatile compounds.

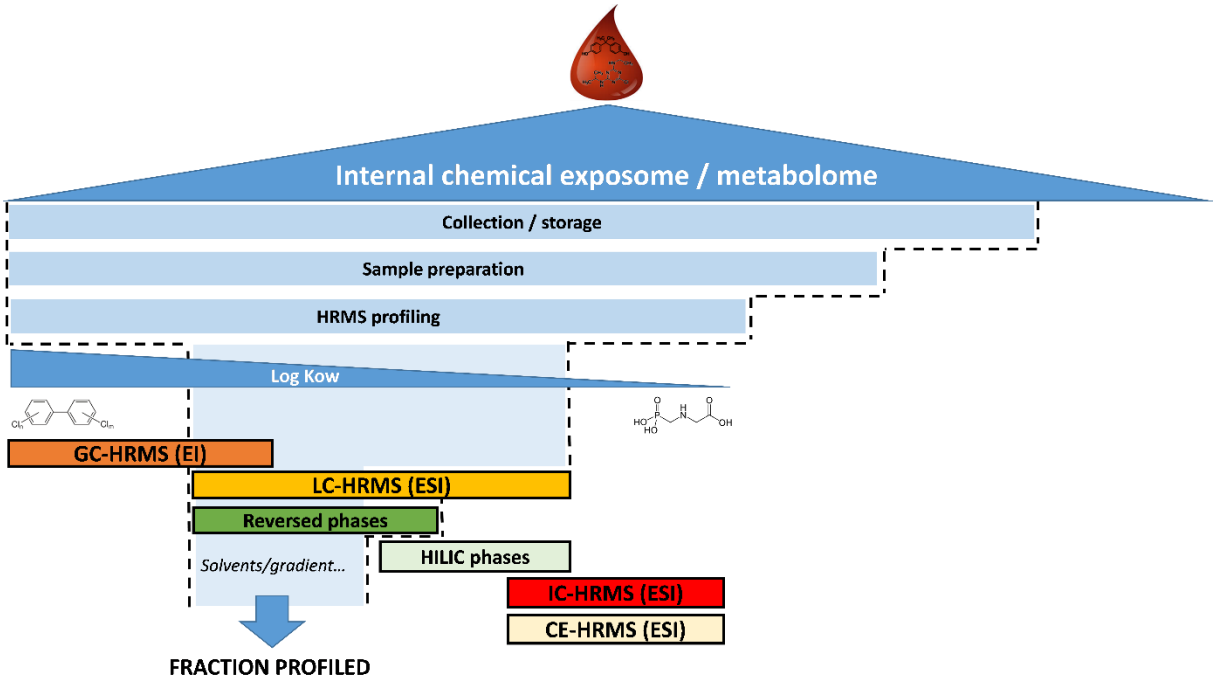

Figure I.2 - Conceptual visualisation of the impact of overarching methodological choices on the profiled fraction of the exposome by David et al., Env Int., 2021. Specificities and overlaps of the different HRMS platforms are schematically represented. Log Kow=octanol/water partition coefficient; GC=gas chromatography; LC=liquid chromatography; IC=ion chromatography, CE=capillary electrophoresis, ESI=Electrospray ionisation, HRMS=High Resolution Mass Spectrometry

Other analytical platforms such as ion chromatography (IC) and capillary electrophoresis (CE) coupled to HRMS can be used to improve coverage of highly ionic and/or polar compounds[77] such as haloacetic acids and antibiotics sulfonamides respectively, although they are not as widespread as LC and GC-based platforms.

The choice of analytical platform is therefore, in itself, a constraint on the observable chemical space of the exposome. Together with the choice of direct or indirect measure and biological matrix, it conditions the structure of the non-target and suspect screening workflow that should be implemented and optimized to characterize the exposome. In the context of this PhD, LC-based approaches were favored for their versatility and their relevance to detect pollutants of emerging concern, which are often non-persistent as opposed to historical contaminants (e.g. POPs). Moreover, the visibility of the pseudomolecular ion due to the soft ionization process, and the substantial availability of MS2 reference data are two important advantages to carry out the annotation process.

2. Implementing NTA to characterize the exposome: constructing a non-targeted and suspect screening workflow

Once the overarching conceptual and methodological choices are made for the generation of the chemical fingerprints, a non-targeted and/or suspect screening workflow including many steps has to be implemented and optimized to correctly process UHPLC-ESI-HRMS raw data. These steps include the implementation of bioinformatics tools to extract chemical features, statistics to prioritize relevant features, and the annotation step to assign a chemical identity to features of interest. To date, there is no comprehensive tool to perform raw data interpretation from data processing to annotation, although some online infrastructures such as Workflow4metabolomics built upon the Galaxy web-based platform tend towards it[88]. Due to the wide variety of available approaches to perform non-targeted and suspect screening, there are also no guidelines to orient the choice of data processing tools, or their parametrization[89]. This is reportedly one of the major bottlenecks of NTA[89, 90].

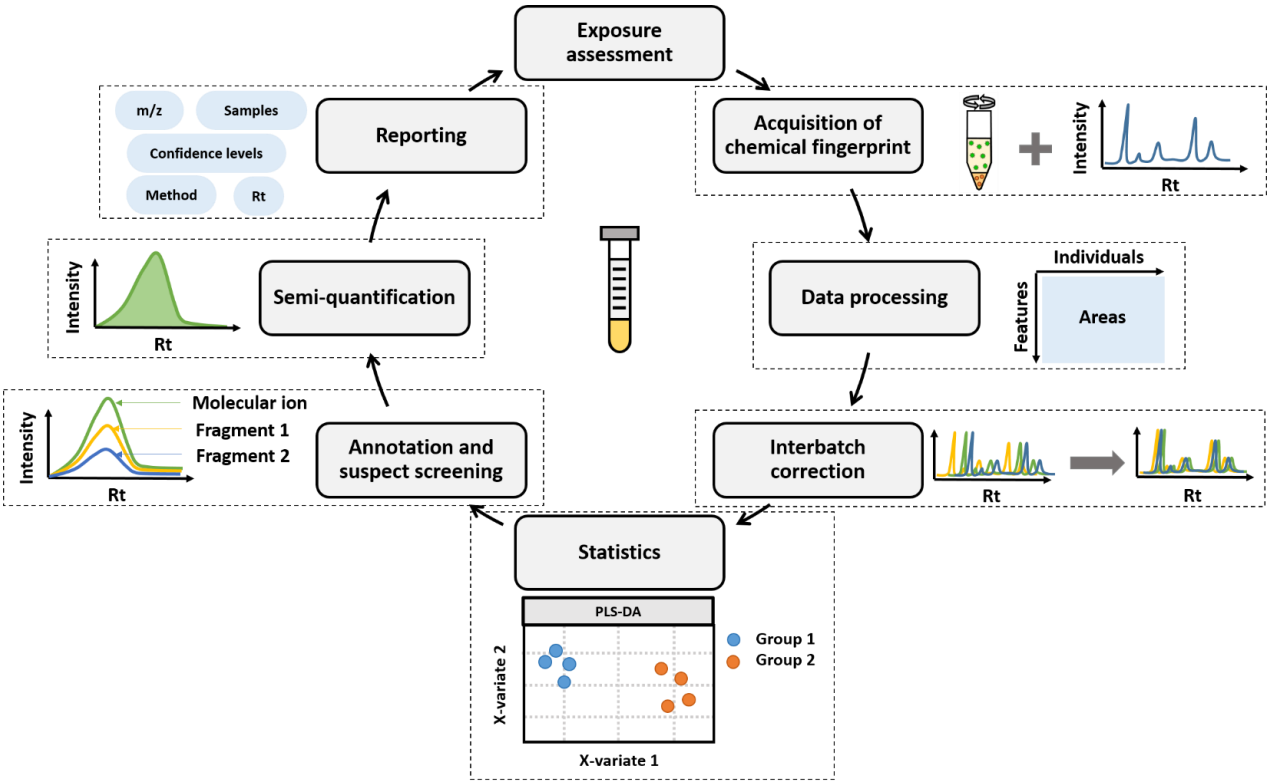

Figure I.3 – Main steps of a non-targeted and suspect screening workflow implemented to investigate the chemical exposome in biological matrices

The main steps of a workflow to characterize the chemical exposome in blood-derived biological matrices using LC-ESI-HRMS are presented in Figure I.3. Workflows used for exposomics applications are, per their general structure, quite similar to workflows used for metabolomics applications[91, 92]. However, in metabolomics, the focus is put on endogenous

chemicals only, with blood concentrations up to eight orders of magnitude above blood concentrations of exogenous chemicals (e.g. steroids or lipids found at ~1 mg/mL compared to industrial pollutants found at ~10 pg/mL) [5, 20]. In exposomics approaches, both are of interest although with a focus on exogenous compounds. This wide range of concentrations implies adaptations to the workflow at every step to ensure that low-abundant compounds are lost neither to ion suppression (first analytical step) nor to inadequate noise levels (first bioinformatics step). The several steps of the workflow are presented in the following sections.

2.1. Acquisition of the chemical fingerprint

Optimizing the acquisition of a chemical fingerprint involves two main steps, namely sample preparation and sample analysis. Regarding sample preparation, to date, there are no universal guidelines recommended for exposomics applications on human biological matrices. Recently, the HBM4EU initiative included for the first time a work package dedicated to suspect and non-target screening in human biological samples. The first steps towards a harmonization of sample preparation practices for suspect and non-targeted screening have been documented[89, 93]. These initial advancements allowed identifying crucial points of vigilance that must be carefully considered with NTA. These critical points include the starting volume, which should be minimized while retaining sufficient sensitivity performances, the extraction method, which greatly impacts the sensitivity versus selectivity compromise further described below, and the inclusion (or lack thereof) of a deconjugation step, which is traditionally used in targeted methods applied on urine samples but may lead to added variability[89, 93]. However, no consensus has yet been reached considering the complexity of the task and, importantly, the diversity of research objectives (e.g. exposure assessment in blood-derived matrices). This can be explained by several reasons. Firstly, while these matrices all contain high-abundant endogenous compounds which are likely to cause matrix-related troubles, they each have their specificities, thus possibly influencing the choice of the most appropriate sample preparation techniques. These specificities are even visible on matrices that may appear similar initially, such as blood serum and blood plasma, or even blood plasmas obtained with different anticoagulants[94]. Secondly, as no sample preparation method can comprehensively cover the width of compounds constituting the chemical exposome, it is beneficial to the community as a whole to explore different methods on similar (or even identical) samples. These developments condition the feasibility of implementing operational workflows combining different sample preparation methods, while still meeting the miniaturization requirements encountered in the case of valuable biological samples with limited availability.

The most commonly used sample preparation method (SPM) in metabolomics is protein precipitation (PPT)[95-98]. As proteins is one of the major classes of compounds in blood-derived

samples, notably regarding abundance, their elimination is the minimum sample purification necessary to reduce matrix effect and preserve the analytical system integrity (e.g. extending column life). This method was historically favored as it is simple, fast and highly non-selective, which is particularly sought after in non-targeted approaches. However, there are other classes of compounds highly abundant in plasma and serum samples that are not eliminated through this process, such as phospholipids and lysophospholipids. The gain in compound detection obtained from the low selectivity may therefore be compensated by the loss of signal due to ion suppression[99]. Moreover, issues with the analytical system such as clogging or poor column life may be exacerbated by the still complex PPT samples[99]. This partly explains the growing interest in solid phase extraction and filtration plates such as protein and phospholipid removal (PLR) methods in HRMS-based exposomics.

PLR methods have gained traction in the last few years as sample delipidation combined with deproteinization as they allow decreasing ion-suppression phenomena and extending LC-MS system life[83]. These methods specifically retain phospholipids through sometimes undivulged mechanisms, presumably relying on interactions between the packed-bed structure and polar esterified phosphate group found in phospholipids[100]. PLR methods have been shown to enhance analyte detection of non-lipid compounds compared to PPT methods[97, 101], and have been described as complementary to PPT in terms of metabolome coverage[95].

Other sample preparation method such as supported liquid extraction (SLE) also allow further sample purification. SLE methods aim to purify samples by using the affinity of compounds of interest for one solvent over another (both solvents being immiscible). These methods are similar to liquid-liquid extraction (LLE), with the exception that a solid media is used to support the extraction, replacing the interface traditionally formed between the two immiscible solvent. SLE methods are favored in the case of blood-derived sample preparation due to often high emulsification in the case of LLE, as well as an easier miniaturization of the sample volume needed[102]. LLE methods, such as the Bligh-Dyer[103] or the Folch[104] method, have been successfully used in metabolomics and lipidomics approaches to simultaneously cover non-polar lipids and polar metabolites[105, 106]. While primarily aimed at non-polar compounds due to the natures of the solid media and the extraction solvent, SLE methods have been reported to perform adequately on more polar compounds[107]. As other mentioned SPM are often more geared towards polar compounds, the use of SLE may allow the observation of another facet of the chemical exposome.

Lastly, solid phase extraction (SPE) methods have vastly been used for biological matrices[95, 98, 99, 101], as they offer a high level of sample purification and hugely limit matrix effects. However, despite the expected drastic decrease in ion suppression and for preserving UHPLC

columns, there is a concern for excessive method selectivity leading to a loss of information. Moreover, the overall complexity of SPE protocols allow more room for human error. These concerns have however been dampened by previous studies using non-targeted metabolomics approaches, where it was determined that the sometimes-reduced recovery of specific compounds was not necessarily associated with total loss of relevant information, especially when considering the possibility of increased concentration of extracts[96, 99].

Other sample preparation methods seem promising despite their limited reported use, such as solid phase micro extraction (SPME), which is reported to allow the recovery of compounds with a wide range of physical-chemical properties and limiting samples handling steps[108, 109].

Sample preparation for NTA are especially challenging to optimize, as there is no set list of compounds of interest on which to rely to ensure adequate performance. Moreover, it is less simple to monitor external contamination compared to targeted approaches. A systematic assessment of sample preparation performance for HRMS-based exposomics applications should therefore be conducted to document its impact on the observed chemical space. Consequently, a performance assessment of the sample preparation step will be the subject of one of the chapters of this PhD.

2.2. Data processing

Data processing for non-targeted approaches is the next decisive step in the workflow. This step involves transforming chromatographic and spectral data to a list of features; each attributed a m/z, a Rt, and an area for each analyzed sample. This step is critical since the rest of the workflow, especially annotation, is based on the feature list generated at this point. Its optimization is therefore paramount to ensure the correct detection and integration of features of interest. In the case of exposomics applications, with low-abundant compounds in complex matrices, it is particularly important to ensure that the data processing allows the disentangling of these signals from the noise. Very few to no studies are available regarding the optimization of this step for HRMS-based exposomics.

Data processing is conducted in four main steps: firstly, the signal is translated to peaks in each analyzed sample (i.e. peak picking). Peaks of all included samples are then aligned to obtain a single peak list. Missing values are filled if peaks were missed in some samples during the initial step (i.e. gap filling). Lastly, areas are normalized to ensure inter-sample comparability. A representation of this process is presented in Figure I.4.

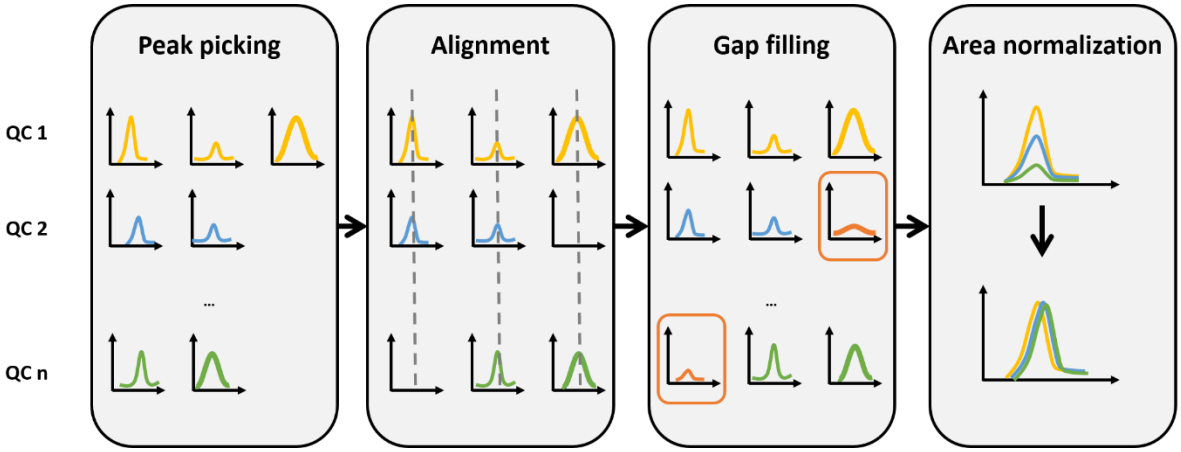

Figure I.4 – Main steps of the non-targeted data processing workflow, comprising of peak picking, alignment, gap filling, and normalization. These steps are presented on quality control (QC) samples. Various strategies and algorithms are available for the peak picking step, the alignment step and the normalization step, as detailed in Chapter II.

A number of open source software are available for non-targeted data processing, among which are the commonly used XCMS operated under an R environment[110] or online[111], MZmine2[112], MS-DIAL[113], and XCMS galaxy-based Workflow4Metabolomics[88]. Vendors also provide non-targeted data processing software, such as Mass Profiler Professional from Agilent, MetaboScape from Bruker, Compound Discoverer and TraceFinder from Thermo, MarkerView and XCMSplus from Sciex, or Progenesis QI from Waters. This multiplicity of tools, while beneficial to allow tailoring data processing to each application's need, also leads to an absence of guidelines regarding the preferential use of a particular software tool for select applications, or even regarding the parameter settings that should be used[114]. This is exacerbated by the lack of consensus regarding reporting data processing parameters in the literature, possibly explained by the fact that highly customizable processing workflows entail a large number of parameters to set and report.

While several data processing workflow optimizations and comparisons are available in the literature[114-118], they are tailored towards metabolomics applications. However, as compounds of interest in exposomics applications are often low abundant, the suggested optimized parameters may lead to failure to correctly identify peaks of interest. Parameters such as noise threshold, peak width or maximum authorized asymmetry should be adjusted to account for peaks presenting different characteristics to those classically encountered in metabolomics. As for the sample preparation method, the data processing method should be thoroughly evaluated to ensure that important chemical information is not lost at this stage.

2.3. Interbatch correction

When performing large-scale exposomics applications, the chemical analysis may be performed in several batches during several weeks. The collected data may suffer from systematic variability in Rt and signal[119] due to LC-ESI-HRMS analytical drifts, which may result in loss of data (e.g. sensitivity loss). These analytical issues, alongside data processing problems such as incorrect binning can lead to inaccuracy of further statistical analyses[120]. To correct the analytical drift in terms of retention time and intensity, interbatch correction should be implemented. While interbatch correction is usually considered part of the data processing step (i.e. alignment and normalization steps), the methods commonly used for these steps may not be sufficient to account for low-abundant compounds.

Traditional alignment processes only rely on a Rt tolerance value which is applied across samples, i.e. peaks with the same m/z value (within a m/z tolerance) in different samples will be considered as one feature if their Rt value is identical within this user-set tolerance value. The issue with this approach is that when a drift phenomenon is observed, the tolerance would have to be set at a high value to account for difference between first and last samples. This could lead to the alignment of peaks that are in fact two distinct compounds with similar although not identical Rt values (i.e. aforementioned binning issue). Moreover, the often non-linear nature of Rt drift with LC methods puts the relevance of a fixed Rt tolerance value into question. To address this issue, various data processing software provide additional Rt correction algorithms. These algorithms usually rely on peaks present in most or all samples to perform the Rt correction, whether they are user-specified (i.e. internal standards, such as for MS-DIAL), or chosen by the processing tool (i.e. adjustRtime - peakGroups algorithm available with XCMS[121]).

In targeted approaches, signal drift correction is usually performed by using internal standards[119]. However, in non-targeted approaches, these compounds only represent a fraction of the features, which may not be representative of varying signal fluctuation between chemical classes[122]. Similarly, signal drift correction methods traditionally used for NTA, based on total intensity or intensity of most abundant features, may fail to account for differing variability between metabolite classes[119]. This observation led to the development of quality control (QC)-based methods, where a sample constituted of pooled aliquots of all samples is repeatedly injected throughout the batches and used as a reference point[92, 119, 122-124]. Although not often compared in the literature, algorithms relying on all features of QC samples such as batchCorr[119] are reported to outperform internal standard drift correction and other linear sequence corrections[92, 119, 125].

This interbatch correction step, while often incorporated into the data processing workflow, should be carefully considered to ensure that the obtained feature list can be relied on for statistical analysis, and later non-targeted screening and/or suspect screening.

2.4. Statistical analysis

As non-targeted exposomics applications can generate datasets comprised of tens of thousands of features, statistical analyses are helpful to identify and prioritize features significantly altered between samples, either for further characterization (e.g. MS2 acquisition) or for annotation. Both univariate and multivariate analyses can be used to this end. The choice of a strategy for statistical analyses highly depends on the study design (e.g. case-control study, exposed vs non-exposed, etc.) as well as the nature and volume of data collected alongside the biological samples (e.g. socio-economic, geographical, clinical data).

While the non-targeted chemical exposome characterization is by nature multivariate (i.e. simultaneous observations of multiple variables), the high dimensionality of the generated datasets entails a high proportion of sparse information leading possible loss of multivariate model performance[126]. Multivariate data-driven dimension reduction techniques, such as principal component analysis (PCA) or partial least squares (PLS) can be used to describe the exposome or to evaluate exposome-health associations[127]. These approaches can describe the exposome by combining variables (i.e. exposures) that tend to occur simultaneously into independent components. These components describe the main patterns discriminating individuals or groups of individuals. However, due to the complex nature of the data, it may be difficult to summarize it with a reasonable number of patterns[127]. Establishing correlations between exposures and health is also challenging. Indeed, correlations between exposures, described as the exposome correlation structure, is largely dependent on the study settings and strongly affect the statistical method's performance in differentiating between true patterns predictors and correlated covariates[127, 128]. The low interpretability of generated independent components and the limited possibility of adjusting for confounders are also challenges for these multivariate exposome statistical analyses. Overall, despite the lack of guidelines, the growing implementation of large-scale exposomics studies will allow to better assess the performance of statistical exposome methods in varying contexts.

Multivariate approaches can be complemented by univariate approaches, which consider each feature individually. While considerably easier to implement, univariate statistical analyses in a multivariate context requires several adaptations to correct the dramatically increased false positive results (type I error). Indeed, as the number of hypotheses tests increases, so does the probability of wrongly rejecting the null hypothesis. To limit this multiple testing problem, p-

values generated by parametric or non-parametric tests (depending on whether the data is normally distributed and homoscedastic[129]) can be corrected. The Bonferroni correction, for instance, aims to strictly limit the amount of type I errors, although sometimes at the expense of type II errors (i.e. false negatives, or wrongly accepting the null hypothesis). Since missing a true significant difference is a concern, other approaches such as the false discovery rate (FDR) are often preferred[126]. Briefly, FDR correction adjusts p-values based on the initial non-corrected p-values and on the distribution of p-values among all the considered tests. To do so, a critical value is computed as a function of the feature's significance rank, the total number of tests and the chosen false discovery rate (usually 5%); the largest p-value that is inferior to to this critical value, as well as all smaller p-values, are significant. An adjusted p-value that is a function of the raw p-value, the feature's significance rank and the total number of tests can be computed. Both of these corrections can be performed depending on the application, although FDR corrections such as Benjamini-Hochberg are often preferred for non-targeted approaches[126, 129]. Vinaixa et al. (2012) [126] provide a detailed and comprehensive workchart to help navigate the implementation of univariate analyses for non-targeted data.

While methodological challenges still exist, univariate and multivariate statistical methods can improve the efficiency and reliability of the non-targeted workflow. These methods allow prioritizing features of interest for further investigation through various annotation strategies.

2.5. Annotation: non-targeted and suspect screening

Non-targeted HRMS-based methods, while not entirely comprehensive in their coverage of the chemical exposome, still generate a large amount of data. Exposomics datasets often include 10,000 to 50,000 features, as even low-abundant peaks are of potential interest. The annotation step aims to assign chemical identities to the detected signals with varying confidence levels depending on the amount and nature of the gathered elements of proof[130]. The consensus ranking currently used by the HRMS-based non-targeted community is the one proposed by Schymanski et al. (2014), where the highest confidence level is achieved by matching exact mass, MS/MS fragmentation pattern and Rt to a standard compounds, as schematized in Figure I.5[130]. It should be noted that the development of new methodological tools in the last few years, such as retention time prediction models[131-134], raise the question of updating this ranking system to account for other predictors. Due to a combination of the high volume of data generated, time restrictions and limited access to standards (for financial or availability reasons), it is usually admitted than less than 10% of features are identified[5].

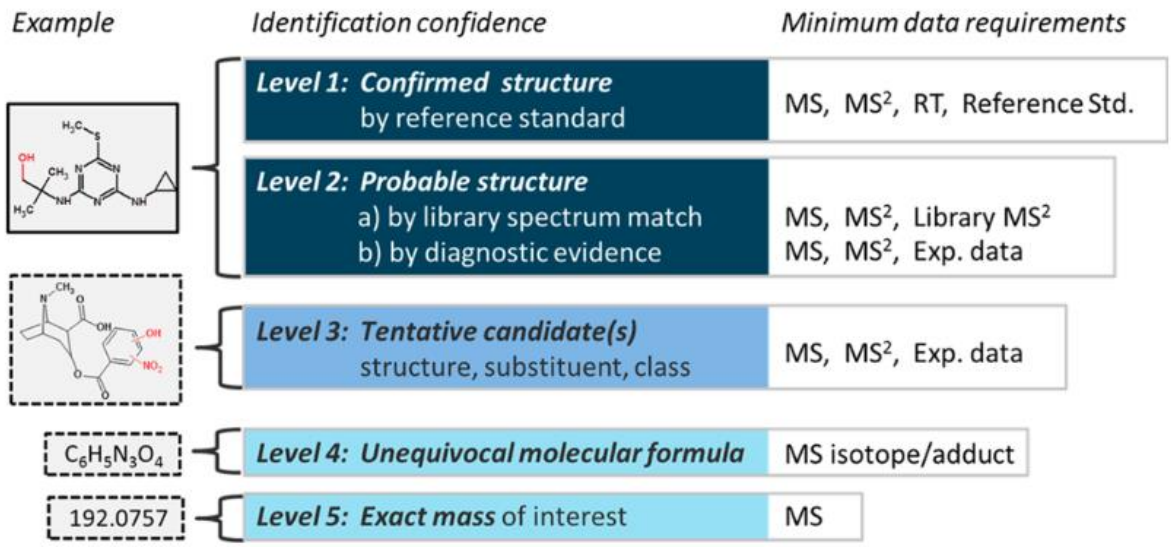

Figure I.5 – Identification confidence levels in high-resolution mass spectrometry proposed by Schymanski et al., ES&T, 2014.

Assigning chemical identities to features can be performed through two main approaches, namely non-targeted screening and suspect screening. Both of these approaches aim to identify new and/or infrequently investigated markers of chemical exposure through different methodologies.

Non-targeted screening consists in unambiguously identifying a feature's identity with no prior reference knowledge. This task is incredibly complex, as the number of tentative candidates, even restricted by a chemical formula, can still be extremely high; for example, a saturated alkane such as $C_{10}H_{22}$ already presents 75 possible isomers and 136 possible stereoisomers. Moreover, strong knowledge on analytical chemistry and biochemistry are needed to assess the plausibility of a given candidate; precise structure elucidation may require the use of other analytical techniques, such as nuclear magnetic resonance. Another bottleneck of annotation is the large size of non-targeted datasets, which cannot be entirely annotated. This can be managed by using statistical analyses to prioritize features of interest for non-targeted screening.

Suspect screening is performed by using one or more lists of compounds suspected to be present in a sample (e.g. expected dietary or occupational biomarkers), which is compared using several criteria to the feature list generated during the previous steps. This comparison is usually done through the comparison of chemical descriptors (e.g. m/z, Rt, isotopes[93, 135]) and correlation/clustering methods. Several automatized solutions based on this principle have been developed over the last few (e.g. CAMERA, MolNetEnhancer ProbMetab, and MetAssign)[136-139]. The use of biological matrices has also led to the use of biological correlation

(i.e. implication in the same pathways, etc.) to assist automatized suspect screening (e.g. xMSannotator[140]). Another powerful chemical predictor is the MS2 fragmentation pattern, which is also widely used to assist suspect screening in some software tools (e.g. MS-DIAL[113], msPurity[141], and DecoMetDIA[142]). While these annotation tools sometimes allow to directly use publicly available databases such as HMDB[143] or KEGG[144], the suspect screening strategy can be less time-consuming if the list of suspects is prioritized depending on varying criteria (e.g. chemical class, toxicity, production volume, etc.). Moreover, these databases were initially designed for metabolomics. They may therefore only be moderately relevant for exposomics applications. While these databases are still relevant to monitor a biological reaction to an exposure (i.e. biomarkers of effect), other databases such as the Blood Exposome Database[145] , Exposome Explorer[146] or CECscreen[147] may be better suited to identify biomarkers of exposure.

While suspect screening strategies have been greatly improved in the last few years[77], the use of sufficient and relevant chemical predictors is needed to decrease the rate of false positives and therefore limit the number of putative annotations that need manual curation. Furthermore, there is still a high need for the automation of this process, as few tools are available to implement suspect screening approaches. Most of the existing tools rely on highly discriminating MS2 fragmentation patterns, which can be difficult to obtain for less commonly investigated environmental contaminants. The use of *in silico* fragmentation algorithms, such MetFrag[148] or CFM-ID[149], can help bridge the gap between the number of potential substances of interest and the available experimental spectra[150]. When standards are available, local reference libraries can also be built, and the acquired spectra can be submitted to large databases such as MassBank[151]. As MS2 acquisitions may be difficult to trigger in the case of low-abundant compounds such as xenobiotics, the suspect screening process can rely on other MS1 predictors. For instance, as for theoretical MS2 fragmentation models, several algorithms for retention time prediction have been developed and evaluated[132, 133, 152, 153], such as PredRet[134], the retention time index RTI[154], Retip[131], and linear regression models using the octanol-water partition coefficient[155]. While predicted Rt values are not as reliable as experimental values, they are still helpful in combination with other chemical predictors to decrease the rate of false positive annotations. A chapter of this PhD work was dedicated to the development of a suspect screening tool relying on MS1 predictors to partly automatize this process and efficiently prioritize features of interest.

2.6. Semi-quantification

To date, NTA mainly provide qualitative data, i.e. presence/absence of a given biomarker, and semi-quantitative data, i.e. area fold changes between samples. This is a hindrance for the application of NTA in epidemiological studies, which heavily rely on quantitative data to perform statistical exposure-health associations, as well as for risk assessment purposes, which also consider quantitative data.

Semi-quantification relies on the hypothesis of a linear concentration-response relationship. Although rarely perfectly accurate over large concentration ranges[156], these relative comparisons are useful to establish fold change values and investigate statistically significant features for prioritization. Indeed, semi-quantitative results from univariate statistical analysis (corrected for multiple comparisons) have been used to compare emerging contaminants levels in human blood[157]. The same methodology has also been applied to investigate emerging contaminants in environmental matrices[158]. While normalization approaches can improve comparability between samples[156], differing ionization potentials make cross-chemical comparisons based on estimated concentrations difficult with this approach[5, 156, 159].

To generate quantitative data using NTA in epidemiological studies, two main types of approaches can be considered: quantification by surrogate standard or response modeling from chemical structure[156]. Quantification by surrogate standard consists in constructing calibration curves with a list of reference standard deemed representative of the chemicals of interest, and pairing them when similar analytical behaviors are expected (e.g. a parent compound and a metabolite, compounds from the same chemical class, etc.). However, due to intrinsic analytical variance and model uncertainty, these approaches have been reported to yield highly variable inaccuracies depending on the considered compound, and seem challenging to apply for predictive purposes[160].

Since there are several limitations to the surrogate standard assignment approach, models aiming to model compounds' ionization response based on their structure and properties (i.e. hydrophobicity, molecular weight, etc.) have been developed[159, 161]. An important consideration for these approaches is that the constructed models will only be usable in their validity domain, which is conditioned by the diversity (or lack thereof) of the training dataset. This implies that quantitative data would have to be acquired for compounds from different chemical classes, with a wide range of physical-chemical properties, functional groups, etc. in both ionization modes and in matrix to yield a robust model. Such an approach was carried out by Liigand et al. (2020) which allowed an rather low prediction error on compound concentration (i.e. averaging at two-fold)[160], encouraging further investigations of these approaches using

ionization response predictions to provide reliable quantitative data even with non-targeted approaches.

2.7. Reporting

Reporting NTA data can be challenging as no consensus format exists as of yet. While most reports contain the 1 to 5 confidence levels as described by Schymanski et al. (2014)[130], there may be some discrepancies between laboratories depending on interpretation. For instance, it may be relevant to add information regarding predicted Rt values, which are not taken into account in the existing annotation and reporting standards[130], or to document potential deconjugation steps implemented in the sample preparation procedure[93]. Moreover, the ever-evolving technologies, prediction models and methodological approaches may lead to annotations not fitting in any described categories, as is the case for annotations supported by predicted retention times or biotransformation products[5]. All the elements of proof used to assign the considered chemical identity should therefore be reported, along with any additional information that may support plausibility (e.g. production volume) or justify a further prioritization (e.g. toxicity). A common template for the reporting of non-targeted and suspect screening results is currently developed in the HBM4EU initiative[93]. Other recommendations are available in the literature, such as those from Dumas et al. (2022) which include providing m/z, Rt, molecular ion species detected, and fold change values, to ensure providing sufficient analytical, statistical and biological information to allow a full understanding of results[162]. Additionally, providing raw data and associated metadata through online workflows such as XCMSOnline[111] or Workflow4metabolomics[88], on through data repositories such as MetaboLights[163] or Metabolomics Workbench[164] helps enhance cooperation and further tool and database development[165].

2.8. Conclusion

While structurally inspired from workflows developed for metabolomics, HRMS-based exposomics workflow must be adapted and optimized for these specific exposure assessment applications. While highly challenging, this workflow optimization allows ensuring that each step's impact on the produced results is thoroughly investigated, and ideally vastly reduced. A systematic evaluation and optimization of the solutions available for every item of this workflow is necessary to implement robust large-scale applications that are minimally biased, and provide a wide view of the chemical exposome. Epidemiological cohort-based studies associating exposomics data and health data can therefore be carried out and shed some light on the complex links between environmental factors and non-communicable diseases.

3. When non-targeted and suspect screening meet epidemiology: first large-scale applications, achievements and remaining challenges

3.1. Large-scale applications and achievements

To date, there have been no large-scale applications of non-targeted and/or suspect screening approaches in epidemiological studies. However, several exposomics research initiatives have appeared in the last decade in Europe and worldwide (even though not all of them implemented NTA based on HRMS) (

Table I.1).

European projects (FP7) (2012-2017)	***Project name*** **Main objective**	**Funding (Million euros)**
HELIX	*The Human Early-Life Exposome* Novel tools for integrating environmental exposures during early life and child health across Europe	**11.3**
EXPOsOMICs	*Enhanced exposure assessment and omic profiling* Developing a new approach to assess environmental exposures, focusing on air and water pollution	**11.6**
NIEHS projects (USA)	***Project name*** **Main objective**	**Funding (Million euros)**
CHEAR	*The Children's Health Exposure Analysis Resource* Implementing the exposome concept in children's health studies	**34**
HHEAR	*Human Health Exposure Analysis Resource* Capturing the effects of environmental exposures on human health outcomes across the life course	**35**
European projects (H2020) (2017-2022)	***Project name*** **Main objective**	**Funding (Million euros)**
HBM4EU	*The European Human Biomonitoring Initiative* Coordinating and advancing human biomonitoring in Europe to provide evidence for chemical policy making	**74.9**
The European Human Exposome Network (2020-2025)	***Project name*** **Main objective**	**Funding (Million euros)**
ATHLETE	*Advancing tools for human early life-course exposome research and translation* Developing a human exposome toolbox to evaluate the effects of environmental exposure	**12.0**
EXPANSE	*Exposome powered tools for healthy living in urban settings* Maximizing one's health in a modern urban environment	**12.0**

Table I.1 – Research initiatives investigating the links between the chemical exposome and health. NIEHS: National Institute of Environmental Health Sciences. Adapted from David et al., m/s, 2021.

3.1.1. From 2012 to 2017

The emergence of the exposome concept has motivated the funding of several European and international projects aiming to characterize the exposome at a wide scale. Among the first large-scale research projects, two European projects funded in part by the European Commission through the seventh Framework Programme (FP7) were launched in 2012. Firstly, the HELIX project set out to characterize early-life exposures to multiple environmental factors and associate them with omics biomarkers and health outcomes. Methodological tools such as spatial models and exposure monitors were used to evaluate exposure to physical factors such as surrounding green spaces, noise and radiation. Other tools such as questionnaires and chemical analysis were used to assess early-life exposure to a wide range of environmental chemicals including various persistent organic pollutants (polychlorinated biphenyls, dichlorodiphenyldichloroethylene, hexachlorobenzene, polybrominated diphenyl ethers, perfluroalkyl substances), non-persistent pollutants (phtalates, phenols, organophosphate pesticides), and various metals[166]. These chemical parameters were measured in blood using GC-MS-based methods (aforementioned persistent pollutants), in urine using LC-MS-based methods (aforementioned non-persistent pollutants) or hair (mercury). In addition, the links between indirect measurements conducted on environmental samples and direct measurements conducted on biological matrices were investigated to give new insights on future exposure assessments. The HELIX project implicated six European birth cohorts (more than 30,000 mother-child pairs), with a subcohort of more than 1,300 mother-child pairs for which biomarkers, omics signatures and child health outcomes were measured at ages 6-11[167]. This project required a total budget of 11.3 million euros (8.6 million euros from FP7), and allowed to establish several significant environment-health outcomes associations such as perfluoroalkyl substances and cardiometabolic factors[39], and multiple exposures (including chemical mixtures) and cognitive function[168].

Secondly, the EXPOsOMICs project aimed to characterize exposure to air and water contaminants for more than 3000 participants (including newborns, children and adults) from 14 European regions, and to establish links with adverse health outcomes such as cardiovascular diseases, respiratory diseases and type II diabetes. Real-time monitors measuring notably fine particulate matter and innovative models were used to assess exposure to air pollution. Water contamination, on the other hand, was assessed through the determination of disinfection by-products notably in drinkable water and biological matrices such as urine. Omics data was also generated from biological samples obtained from the highly

exposed participants to potentially identify biomarkers of risk and better understand chemical compounds' mechanisms of action ("meet in the middle" approach[169]). It was funded for over 11.6 million euros, with a contribution of more than 8.7 million euros from the European Commission.

These projects, both closed in 2017, undertook the characterization of the chemical exposome at a large scale. However, while a large number of determinants were investigated, they still relied on targeted measurements of known toxicants.

During the same period, sizable infrastructures dedicated to the characterization of the exposome were set up in the United States of America. In 2013, the National Institute of Environmental Health Sciences (NIEHS, USA) funded HERCULES, an environmental health sciences center dedicated to supporting environmental health research through the development of new tools and technologies. This Core center was the first of its kind focused on the exposome concept. This platform supported many research projects throughout the years, providing targeted and high-resolution metabolomics analyses (aiming to identify both biomarkers of effect and exposure), as well as support regarding data analysis[38, 170]. Its funding was renewed for a second cycle in 2017[171]. In 2015, NIEHS also launched the Children's Health Exposure Analysis Resource (CHEAR), a large-scale infrastructure to allow researchers working specifically on children's health to incorporate the concept of exposome to their research[172]. Using targeted and high-resolution metabolomics, this infrastructure allowed researchers to characterize the chemical exposome of over 50,000 children in over 30 studies investigating the links between environmental exposures and adverse health outcomes such as asthma, obesity, autism, etc. In 2019, the Human Health Exposure Analysis Resource (HHEAR) was in turn launched to expand the characterization of the chemical exposome to other time windows of vulnerability during adulthood[173].

3.1.2. From 2017 to 2020 (extended to 2022)

The European Human Biomonitoring Initiative (HBM4EU) was launched in 2017 with a contribution from the European Commission of almost 50 million euros through the Horizon 2020 program (75 million euros of funding in total). This project is a joint effort of 30 countries to coordinate and harmonize human biomonitoring practices to improve the evaluation of the actual exposure of citizens to chemicals and to better understand the effect of mixtures on human health. Its main objectives included the harmonization of procedures for HBM to improve data comparability for policy makers, establishing links between chemical exposures and health outcomes, and adapting risk assessment procedures to account for multiple sources. It was one of the first major projects to include, in addition to targeted approaches

relying on lists of priority substances, non-targeted and suspect screening of biological matrices to characterize environmental exposures. HBM4EU's sixteenth work package titled "Emerging chemicals" is specifically dedicated to the harmonization and implementation of NTA[72, 76, 93]. The work carried out in the context of this work package has contributed to the field of the non-targeted characterization of the exposome on several aspects, notably the establishment of a list of chemicals of emerging concern[76, 93, 147], and recommendations on practices harmonization[93]. At a larger scale, this project has allowed, amongst other results, establishing recommendations for the harmonization of the use of HBM data in risk assessment[49, 174], as well as HBM guidance values for chemicals such as phthalates[175] or cadmium compounds[176].

In this context, the European-wide research program PARC (Partnership for the Assessment of Risks from Chemicals) was developed. This partnership established under Horizon Europe and co-funded for 400 million euros, will last 7 years. It implicates 200 partners, including 3 European Union agencies. PARC's main objectives revolve around the consolidation of the European Union's research capacity for chemical risk assessment to improve the protection of human and environmental health. Work package 4 of this program is specifically dedicated to the exposure and monitoring of chemicals through the development of innovative tools and methods to perform HBM and environmental monitoring. Non-targeted and suspect screening methods will be implemented in this work package.

Lastly, in 2018, the European Strategy Forum on Research Infrastructures (ESFRI) identified a gap in the health and food domain and recommended building an infrastructure dedicated to research surrounding the human exposome at a European level. This prompted the setup of Environmental Exposure Assessment Research Infrastructure (EIRENE), which includes more than 50 partners from 17 countries (including the United Kingdom and the United States of America). EIRENE was added to ESFRI's roadmap in 2021. It aims to bring together complementary capacities of partners to improve exposome research and achieve the high-throughput characterization of the human exposome. Another important aim of EIRENE is the translation of research results towards innovation and policymaking.

3.1.3. <u>From 2020 onwards</u>

In 2020, a large-scale network of projects focused on studying the impact of environmental exposure on human health, the European Human Exposome Network (EHEN), was launched. It was funded from Horizon 2020 for over 106 million euros, and aims to protect citizen's health and well-being from environmental factors. It consists of 9 research projects implicating 126 research groups from 24 countries. One of its main objectives is to develop a Findable

Accessible Interoperable Reusable (FAIR) toolbox for exposome research. This toolbox will notably include innovative tools for the assessment of the exposome, and new methodologies to associate this data to health data.

The Advancing Tools for Human Early Lifecourse Exposome Research and Translation (ATHLETE) project, as a part of EHEN, was launched in 2020. ATHLETE's objectives include the setting-up of a Europe-wide prospective cohort to cover the first 20 years of life using 17 already existing cohorts, measuring multiple environmental exposures and linking it to children's biological responses[177]. A work package dedicated to the non-targeted screening of emerging chemicals will, amongst other analyses including targeted screening, use LC- and GC-HRMS to perform non-targeted and suspect screening on the HELIX subcohort (1 300 individuals). Simultaneously launched in 2020, the EXposome Powered tools for healthy living in urbAN SEttings (EXPANSE) project was also funded for 12 million euros through the Horizon 2020 European program. It involves 20 partners in Europe and in the United States of America working together to identify factors influencing human health in urban environments[178]. EXPANSE includes four main study types: administrative cohorts, adult cohorts, matures birth cohorts, and urban labs with data collected 55 million, 2 million, 30 000 and 5 000 individuals respectively. Biological data was collected for all study types except administrative cohorts. Non-targeted screenings on 10 000 blood samples will be performed using both LC- and GC-HRMS. Both the ATHLETE and EXPANSE project will integrate multiple omics datasets to uncover exposome-health relationships, and allow expanding knowledge on biological pathways. Moreover, exposome-health associations will be explore with epidemiological approaches, as clinical data is available for individuals in the cohorts.

Although these many large-scale EU and international initiatives have been launched to decipher the impact of the chemical exposome on human health, to date, there are no large-scale epidemiological applications of non-targeted or suspect screening approaches. However, there are some studies using non-targeted or suspect screening approaches to characterize the chemical exposome and establish links with endogenous compounds to investigate the effect of various exposures on biological pathways[32, 179, 180]. While these studies constitute the crucial first steps towards conducting epidemiological analyses to investigate associations between environmental chemical exposures and adverse fetal health outcomes (e.g. preterm birth, low birth weight, preeclampsia)[32], breast cancer[179] or liver diseases[180], they report several remaining limitations in achieving this goal.

3.2. Remaining limitations

3.2.1. Statistical power in non-targeted applications

As mentioned above, no large-scale applications of non-targeted and suspect screening are described to date, and this is explained by many major methodological issues. The major challenge for the application of NTA in epidemiological studies is statistical power. Indeed, statistical power in these applications is limited by the large and unknown number of determinants (i.e. exposures) investigated[127]. This is further exacerbated by the fact that high-dimensional collinear data is generated through these approaches[126, 127].

It should be noted that this issue is also prevalent for EWAS conducted using targeted approaches. Indeed, when considering that the association sizes are often low to moderate, and that a substantial proportion of substance concentrations is below the limit of detection, high sample sizes are needed to achieve sufficient power[127, 181 167, 169]. A study investigating the link between 128 environmental contaminants and semen quality found in a post-hoc power analysis that sample size requirements when using a Bonferroni or a FDR correction were of at least 1795 and 925 men respectively, thus determining that many existing cohorts were vastly underpowered to undertake EWAS-like approaches[181].

Regardless of whether targeted or non-targeted approaches are undertaken to characterize the exposome, high sample sizes can be difficult to achieve for various reasons: limited funding for sample collection and analysis, analytical platform availability, loss of follow-up[167, 169], or investigation of rare diseases with low frequencies[180]. Theoretical and methodological studies are therefore still required to overcome the critical challenge of statistical power for use of NTA in epidemiological research.

3.2.2. The incomplete annotation process

Despite the many available tools and databases, the annotation process is still tedious and incomplete. It requires many steps, including searching for mass spectral information in databases, verifying the potential match to the observed feature, and even in some cases, such as isotope elucidation, using non-traditional additional approaches such as using other analytical techniques. While it would not be necessary to annotate the entirety of datasets, annotating only the statistically significant features can still remain an arduous task. For instance, Walker et al. (2021) described identifying 54 compounds associated to primary sclerosing cholangitis, resulting in only one high-confidence match. This can be partly explained by the fact that, to date, the main annotation approach used is suspect screening, since it requires fewer resources and has potential for automation. However, a critical aspect

of suspect screening is the construction of the reference database against which features are compared.

The Matthew effect is a psychological phenomenon described as maintaining prominence of items (i.e. compounds) that have been prominent in the past[183]. This bias includes the prioritization of compounds only based on previously researched compounds and the interpretation of lack of data as a null concentration. While NTA were specifically designed to overcome the restriction of set lists of well-researched compounds of interest, the data generated from these approaches must be made interpretable by expanding knowledge on compounds not traditionally investigated, including through the acquisition of MS2 spectra. Indeed, while hundreds of thousands of compounds are known to be in our environment, it is estimated that only 0.57-3.6% of them have spectral information available[150]. Ongoing efforts for the harmonized and collaborative acquisition of MS2 spectra must therefore be maintained and even expanded.

3.2.3. Interpretability of results: toxicology and determinants of exposure

Once an association between an environmental exposure and an adverse health outcome has been established, additional steps must be taken to understand the nature of this association. Indeed, in datasets as highly collinear as non-targeted exposomics data, it may be difficult to disentangle true predictors of health status and correlated covariates. Additional assays such as high-throughput toxicity screenings models may help to ascertain the effect of an environmental compound on various biological pathways[32]. To this end, the ToxCast program was launched by the U.S. Environmental Protection Agency (EPA) in 2006 to use computational chemistry, high-throughput screening and toxicogenomics technologies to predict toxicity and prioritize chemicals for limited *in-vivo* tests[184]. The data generated by this program is freely available and allows having preliminary data on the predicted toxicity of over 4,400 chemicals. As chemical mixtures may have synergic effects, additional developments must be made to allow these toxicological approaches to integrate multiple compounds. The implementation of toxicological approaches in exposomics is needed both to improve mechanistic understanding of chemicals' effects on human health and to translate these findings into regulatory measures in risk assessment[185].

Lastly, the detection and identification of new toxicants to which humans are exposed raise the question of the determinants of exposure. Indeed, to implement public health policies and limit the exposure to such compounds, the major sources of exposure must be identified. This can be challenging, as there are often multiple sources and confounding factors. To date, this task is mostly accomplished by using detailed questionnaires[23] that allow collecting large amounts

of data regarding socio-demographic features, diet, lifestyle, etc. Although this method is not ideal due to the data being subject to potential recall and reporting biases[186], it is often the most cost-effective way to obtain a starting point to establish the determinants of a given exposure.

Overall, there are still some key conceptual and methodological obstacles to implements NTA for epidemiological studies, including the unresolved question of statistical power, the tedious and incomplete annotation process, and the limited interpretability of the generated data. To address those issues, collaborative efforts must be maintained regarding the generation of additional knowledge, as well as regarding the development of new data processing and statistical methodologies needed to uncover the potential of NTA to investigate the etiology of diseases.

4. Conclusion

This first chapter illustrates the significance of the exposome concept to investigate the etiology of non-communicable chronic diseases, as genetic factors are not sufficient to explain alone their emergence. This exposome concept, combined with the advancement of technologies such as HRMS, paved the way for a change of paradigm for exposure assessment to chemical mixtures and emerging contaminants. Indeed, the development of new non-targeted approaches has allowed envisioning a characterization of the chemical exposome without establishing set lists of prioritized chemicals, but with an (ideally) unbiased vision. However, many technological barriers that come with the non-targeted characterization of the human internal chemical exposome remain. The many necessary methodological choices, which include the choice of matrix, analytical platform, sample preparation and parametrization of bioinformatics tools used for data processing have a hard-to-discern impact on the observable chemical space, which in turn may limit the applicability of these novel approaches in epidemiological studies. Indeed, the diverse and dynamic nature of the chemical exposome are both considerable obstacles to its exhaustive characterization. The combination of different biological matrices (i.e. urine, blood, placenta, hair, etc.), analytical platforms (i.e. LC-HRMS, GC-HRMS, etc.) and sample preparation methods is necessary to encompass the wide range of chemicals that constitute the chemical exposome. Moreover, the data processing and annotation algorithms are not yet fully efficient to translate the non-targeted chemical fingerprints to a list of identified chemical compounds, which is a hindrance to the application of NTA at a large scale. Lastly, the extensive amount of data generated by NTA is also a challenge to establish links between the characterized exposures and the considered health outcomes.

These many conceptual and methodological challenges for the application of non-targeted approaches to epidemiological studies are slowly being addressed through the efforts of independent laboratories and regional and worldwide collaborations. In 2020, the European Human Exposome Network was launched with the aim to bring together 9 research projects studying the impact of environmental exposure on human health. It is partly funded by the European Commission for over 100 million euros, and involves 126 research groups from 24 countries. Closely following in 2021, the Research Infrastructure for EnvIRonmental Exposure assessmeNt in Europe (EIRENE RI) entered in the European Strategy Forum on Research Infrastructures (ESFRI) roadmap. This European research infrastructure connects 50 research institutions from 17 countries and aims to support large-scale research on human health and the environment, way of life, diet, exercise, economic pressures and psychosocial problems. These initiatives hold great promises for supporting the development and harmonization of new methodologies aiming to bridge the gaps in knowledge regarding the impact of environmental exposures on human health, as it is a complex task only achievable through the collaboration of multiple partners focusing on its different aspects.

In this context of rising global interest, this PhD thesis project was focused on developing and optimizing a workflow for the non-targeted LC-ESI-HRMS characterization of the chemical exposome in blood plasma and serum samples. This was conducted by optimizing the acquisition of the chemical fingerprint and notably the sample preparation step, as well as the data processing step for the characterization of low-abundant environmental compounds in complex matrices. Moreover, a suspect screening workflow was developed to improve the efficiency of the annotation step, which remains an important bottleneck for the implementation of NTA. Lastly, a proof-of-concept study was conducted on serum samples from Breton adolescents to demonstrate this workflow’s efficiency to characterize the chemical exposome at a large scale.

References

1. Collins F.S., Morgan M.*, et al.*, The Human Genome Project: Lessons from Large-Scale Biology. *Science* **2003,** *300* (5617), 286-90.
2. Consortium* I.H.G.S., Finishing the euchromatic sequence of the human genome. *Nature* **2004,** *431*, 931-45.
3. Consortium I.H.G.S., Initial sequencing and analysis of the human genome. *Nature* **2001,** *409*, 860-921.
4. Bodmer W.,Bonilla C., Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **2008,** *40* (6), 695-701.
5. David A., Chaker J.*, et al.*, Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **2021,** *156*, 106630.
6. Vineis P., Schulte P.*, et al.*, Misconceptions about the use of genetic tests in populations. *The Lancet* **2001,** *357* (9257), 709-12.
7. Wild C.P., Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **2005,** *14* (8), 1847-50.
8. Wild C.P., The exposome: from concept to utility. *Int J Epidemiol* **2012,** *41* (1), 24-32.
9. Miller G.W.,Jones D.P., The nature of nurture: refining the definition of the exposome. *Toxicol Sci* **2014,** *137* (1), 1-2.
10. Jones D.P., Sequencing the exposome: A call to action. *Toxicol Rep* **2016,** *3*, 29-45.
11. Vineis P., Robinson O.*, et al.*, What is new in the exposome? *Environ Int* **2020,** *143*, 105887.
12. Rappaport S.M., Implications of the exposome for exposure science. *J Expo Sci Environ Epidemiol* **2011,** *21* (1), 5-9.
13. Wang Z., Walker G.W.*, et al.*, Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environ Sci Technol* **2020**.
14. REACH *REACH Registration statistics*; REACH: 2022.
15. Fillol C., Oleko A.*, et al.*, Exposure of the French population to bisphenols, phthalates, parabens, glycol ethers, brominated flame retardants, and perfluorinated compounds in 2014-2016: Results from the Esteban study. *Environ Int* **2021,** *147*, 106340.
16. Casas M., Chevrier C.*, et al.*, Exposure to brominated flame retardants, perfluorinated compounds, phthalates and phenols in European birth cohorts: ENRIECO evaluation, first human biomonitoring results, and recommendations. *Int J Hyg Environ Health* **2013,** *216* (3), 230-42.
17. Barnett-Itzhaki Z., Esteban Lopez M.*, et al.*, A review of human biomonitoring in selected Southeast Asian countries. *Environ Int* **2018,** *116*, 156-64.
18. Murawski A., Roth A.*, et al.*, Polycyclic aromatic hydrocarbons (PAH) in urine of children and adolescents in Germany - human biomonitoring results of the German Environmental Survey 2014-2017 (GerES V). *Int J Hyg Environ Health* **2020,** *226*, 113491.
19. David A., Chaker J.*, et al.*, Acetaminophen metabolism revisited using non-targeted analyses: Implications for human biomonitoring. *Environ Int* **2021,** *149*, 106388.
20. Rappaport S.M., Barupal D.K.*, et al.*, The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* **2014,** *122* (8), 769-74.
21. Huang Z., Li H.*, et al.*, Target and Suspect Screening of Urinary Biomarkers for Current-use Pesticides: Application of a Simple Extraction Method. *Environ Toxicol Chem* **2022,** *41* (1), 73-80.
22. Dereumeaux C., Mercier F.*, et al.*, Identification of pesticides exposure biomarkers for residents living close to vineyards in France. *Environ Int* **2022,** *159*, 107013.
23. Bonvallot N., Jamin E.L.*, et al.*, Suspect screening and targeted analyses: Two complementary approaches to characterize human exposure to pesticides. *Science of The Total Environment* **2021,** *786*.

24. Assens M., Frederiksen H.*, et al.*, Variations in repeated serum concentrations of UV filters, phthalates, phenols and parabens during pregnancy. *Environ Int* **2019,** *123*, 318-24.
25. Locatelli M., Furton K.G.*, et al.*, An FPSE-HPLC-PDA method for rapid determination of solar UV filters in human whole blood, plasma and urine. *J Chromatogr B Analyt Technol Biomed Life Sci* **2019,** *1118-1119*, 40-50.
26. Ma Y., Liu H.*, et al.*, The adverse health effects of bisphenol A and related toxicity mechanisms. *Environ Res* **2019,** *176*, 108575.
27. Pathak R.,Dikshit A.K., Atrazine and Human Health. *IJE* **2011,** *1* (1), 14-23.
28. Vorkamp K., Castano A.*, et al.*, Biomarkers, matrices and analytical methods targeting human exposure to chemicals selected for a European human biomonitoring initiative. *Environ Int* **2021,** *146*, 106082.
29. Vermeulen R., Schymanski E.L.*, et al.*, The exposome and health: Where chemistry meets biology. *Science* **2020,** (367), 392-6.
30. Di Renzo G.C., Conry J.A.*, et al.*, International Federation of Gynecology and Obstetrics opinion on reproductive health impacts of exposure to toxic environmental chemicals. *Int J Gynaecol Obstet* **2015,** *131* (3), 219-25.
31. SOULIER C., BOITEUX V.*, et al.*, La spectrométrie de masse haute résolution pour la recherche de micropolluants organiques dans l'environnement. *Techniques Sciences Méthodes* **2021**, 43-54.
32. Panagopoulos Abrahamsson D., Wang A.*, et al.*, A Comprehensive Non-targeted Analysis Study of the Prenatal Exposome. *Environ Sci Technol* **2021**.
33. Niedzwiecki M.M., Walker D.I.*, et al.*, The Exposome: Molecules to Populations. *Annu. Rev. Pharmacol. Toxicol* **2019,** *59* (1), 6.1–6.21.
34. Wang A., Gerona R.R.*, et al.*, A Suspect Screening Method for Characterizing Multiple Chemical Exposures among a Demographically Diverse Population of Pregnant Women in San Francisco. *Environ Health Perspect* **2018,** *126* (7), 077009.
35. Rappaport S.M.,Smith M.T., Epidemiology. Environment and disease risks. *Science* **2010,** *330* (6003), 460-1.
36. Barker D.J.P.,Osmond C., Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *The Lancet* **1986,** *327* (8489), 1077-81.
37. Haugen A.C., Schug T.T.*, et al.*, Evolution of DOHaD: the impact of environmental health sciences. *J Dev Orig Health Dis* **2015,** *6* (2), 55-64.
38. Tan Y., Barr D.B.*, et al.*, High-resolution metabolomics of exposure to tobacco smoke during pregnancy and adverse birth outcomes in the Atlanta African American maternal-child cohort. *Environ Pollut* **2022,** *292* (Pt A), 118361.
39. Papadopoulou E., Stratakis N.*, et al.*, Prenatal and postnatal exposure to PFAS and cardiometabolic factors and inflammation status in children from six European cohorts. *Environ Int* **2021,** *157*, 106853.
40. Garcia E., Stratakis N.*, et al.*, Prenatal and childhood exposure to air pollution and traffic and the risk of liver injury in European children. *Environ Epidemiol* **2021,** *5* (3), e153.
41. Franken C., Koppen G.*, et al.*, Environmental exposure to human carcinogens in teenagers and the association with DNA damage. *Environ Res* **2017,** *152*, 165-74.
42. Buttke D.E., Sircar K.*, et al.*, Exposures to endocrine-disrupting chemicals and age of menarche in adolescent girls in NHANES (2003-2008). *Environ Health Perspect* **2012,** *120* (11), 1613-8.
43. Steckling N., Gotti A.*, et al.*, Biomarkers of exposure in environment-wide association studies - Opportunities to decode the exposome using human biomonitoring data. *Environ Res* **2018,** *164*, 597-624.
44. Raffy G., Mercier F.*, et al.*, Oral bioaccessibility of semi-volatile organic compounds (SVOCs) in settled dust: A review of measurement methods, data and influencing factors. *J Hazard Mater* **2018,** *352*, 215-27.
45. Dulio V., Koschorreck J.*, et al.*, The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): let's cooperate! *Environmental Sciences Europe* **2020,** *32* (1).

46. Wei W., Bonvallot N.*, et al.*, Bioaccessibility and bioavailability of environmental semi-volatile organic compounds via inhalation: A review of methods and models. *Environ Int* **2018,** *113*, 202-13.
47. Szabo K., Emőke Teleky B.*, et al.*, Bioaccessibility of microencapsulated carotenoids, recovered from tomato processing industrial by-products, using in vitro digestion model. *Lwt* **2021,** *152*.
48. Hoppin J.A., Adgate J.L.*, et al.*, Environmental exposure assessment of pesticides in farmworker homes. *Environ Health Perspect* **2006,** *114* (6), 929-35.
49. Louro H., Heinala M.*, et al.*, Human biomonitoring in health risk assessment in Europe: Current practices and recommendations for the future. *Int J Hyg Environ Health* **2019,** *222* (5), 727-37.
50. Zidek A., Macey K.*, et al.*, A review of human biomonitoring data used in regulatory risk assessment under Canada's Chemicals Management Program. *Int J Hyg Environ Health* **2017,** *220* (2 Pt A), 167-78.
51. Dereumeaux C., Fillol C.*, et al.*, The French human biomonitoring program: First lessons from the perinatal component and future needs. *Int J Hyg Environ Health* **2017,** *220* (2 Pt A), 64-70.
52. Remer T., Montenegro-Bethancourt G.*, et al.*, Long-term urine biobanking: storage stability of clinical chemical parameters under moderate freezing conditions without use of preservatives. *Clin Biochem* **2014,** *47* (18), 307-11.
53. Elliott P., Peakman T.C.*, et al.*, The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* **2008,** *37* (2), 234-44.
54. Ferland S., Cote J.*, et al.*, Detailed Urinary Excretion Time Courses of Biomarkers of Exposure to Permethrin and Estimated Exposure in Workers of a Corn Production Farm in Quebec, Canada. *Ann Occup Hyg* **2015,** *59* (9), 1152-67.
55. Hernandez A.F., Lozano-Paniagua D.*, et al.*, Biomonitoring of common organophosphate metabolites in hair and urine of children from an agricultural community. *Environ Int* **2019,** *131*, 104997.
56. Husoy T., Andreassen M.*, et al.*, The Norwegian biomonitoring study from the EU project EuroMix: Levels of phenols and phthalates in 24-hour urine samples and exposure sources from food and personal care products. *Environ Int* **2019,** *132*, 105103.
57. Haug L.S., Sakhi A.K.*, et al.*, In-utero and childhood chemical exposome in six European mother-child cohorts. *Environ Int* **2018,** *121* (Pt 1), 751-63.
58. Esteban M.,Castano A., Non-invasive matrices in human biomonitoring: a review. *Environ Int* **2009,** *35* (2), 438-49.
59. Hardy E.M., Dereumeaux C.*, et al.*, Hair versus urine for the biomonitoring of pesticide exposure: Results from a pilot cohort study on pregnant women. *Environ Int* **2021,** *152*, 106481.
60. Coppola L., Cianflone A.*, et al.*, Biobanking in health care: evolution and future directions. *J Transl Med* **2019,** *17* (1), 172.
61. Ronningen K.S., Paltiel L.*, et al.*, The biobank of the Norwegian Mother and Child Cohort Study: a resource for the next 100 years. *Eur J Epidemiol* **2006,** *21* (8), 619-25.
62. Khmiri I., Cote J.*, et al.*, Toxicokinetics of bisphenol-S and its glucuronide in plasma and urine following oral and dermal exposure in volunteers for the interpretation of biomonitoring data. *Environ Int* **2020,** *138*, 105644.
63. Calafat A.M., Contemporary Issues in Exposure Assessment Using Biomonitoring. *Curr Epidemiol Rep* **2016,** *3* (2), 145-53.
64. Dvorakova D., Pulkrabova J.*, et al.*, Interlaboratory comparison investigations (ICIs) and external quality assurance schemes (EQUASs) for flame retardant analysis in biological matrices: Results from the HBM4EU project. *Environ Res* **2021,** *202*, 111705.
65. Engel S.M.,Wolff M.S., Causal inference considerations for endocrine disruptor research in children's health. *Annu Rev Public Health* **2013,** *34*, 139-58.

66. Paulzen M., Goecke T.W.*, et al.*, Pregnancy exposure to quetiapine - Therapeutic drug monitoring in maternal blood, amniotic fluid and cord blood and obstetrical outcomes. *Schizophr Res* **2018,** *195*, 252-7.
67. Ostrea E.M., Jr., Bielawski D.M.*, et al.*, Combined analysis of prenatal (maternal hair and blood) and neonatal (infant hair, cord blood and meconium) matrices to detect fetal exposure to environmental pesticides. *Environ Res* **2009,** *109* (1), 116-22.
68. Fannin M.,Kent J., Origin stories from a regional placenta tissue collection. *New Genet Soc* **2015,** *34* (1), 25-51.
69. Makin J., Blount K.*, et al.*, The Global Pregnancy Collaboration (CoLab) Biobank of rare placentas. *Placenta* **2021,** *114*, 50-1.
70. Concheiro M., Gutierrez F.M.*, et al.*, Assessment of biological matrices for the detection of in utero cannabis exposure. *Drug Test Anal* **2021,** *13* (7), 1371-82.
71. Concheiro M.,Huestis M.A., Drug exposure during pregnancy: analytical methods and toxicological findings. *Bioanalysis* **2018,** *10* (8), 587-606.
72. Pourchet M., Narduzzi L.*, et al.*, Non-targeted screening methodology to characterise human internal chemical exposure: Application to halogenated compounds in human milk. *Talanta* **2021,** *225*, 121979.
73. Dubocq F., Karrman A.*, et al.*, Comprehensive chemical characterization of indoor dust by target, suspect screening and nontarget analysis using LC-HRMS and GC-HRMS. *Environ Pollut* **2021,** *276*, 116701.
74. Boudah S., Olivier M.F.*, et al.*, Annotation of the human serum metabolome by coupling three liquid chromatography methods to high-resolution mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci* **2014,** *966*, 34-47.
75. Delaporte G., Cladière M.*, et al.*, Untargeted food chemical safety assessment: A proof-of-concept on two analytical platforms and contamination scenarios of tea. *Food Control* **2019,** *98*, 510-9.
76. Pourchet M. Development of non-targeted approaches to evidence emerging chemical hazard - Identification of new biomarkers of internal human exposure, in order to support human biomonitoring and the study of the link between chemical exposure and human health. ONIRIS, 2020.
77. Gonzalez-Gaya B., Lopez-Herguedas N.*, et al.*, Suspect and non-target screening: the last frontier in environmental analysis. *Anal Methods* **2021,** *13* (16), 1876-904.
78. Milman B.L.,Zhurkovich I.K., The chemical space for non-target analysis. *TrAC Trends in Analytical Chemistry* **2017,** *97*, 179-87.
79. Jandera P.,Janas P., Recent advances in stationary phases and understanding of retention in hydrophilic interaction chromatography. A review. *Anal Chim Acta* **2017,** *967*, 12-32.
80. Beschnitt A., Schwikowski M.*, et al.*, Towards comprehensive non-target screening using heart-cut two-dimensional liquid chromatography for the analysis of organic atmospheric tracers in ice cores. *J Chromatogr A* **2022,** *1661*, 462706.
81. Hemmler D., Heinzmann S.S.*, et al.*, Tandem HILIC-RP liquid chromatography for increased polarity coverage in food analysis. *Electrophoresis* **2018,** *39* (13), 1645-53.
82. Yusa V., Millet M.*, et al.*, Analytical methods for human biomonitoring of pesticides. A review. *Anal Chim Acta* **2015,** *891*, 15-31.
83. Chetwynd A.J.,David A., A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. *Talanta* **2018,** *182*, 380-90.
84. Mazur D.M., Detenchuk E.A.*, et al.*, GC-HRMS with Complementary Ionization Techniques for Target and Non-target Screening for Chemical Exposure: Expanding the Insights of the Air Pollution Markers in Moscow Snow. *Sci Total Environ* **2021,** *761*, 144506.
85. Valvi D., Walker D.I.*, et al.*, Environmental chemical burden in metabolic tissues and systemic biological pathways in adolescent bariatric surgery patients: A pilot untargeted metabolomic approach. *Environ Int* **2020,** *143*, 105957.

86. Ruiz-Delgado A., Plaza-Bolanos P.*, et al.*, Advanced evaluation of landfill leachate treatments by low and high-resolution mass spectrometry focusing on microcontaminant removal. *J Hazard Mater* **2020,** *384*, 121372.
87. Alharbi O.M.L., Basheer A.A.*, et al.*, Health and environmental effects of persistent organic pollutants. *Journal of Molecular Liquids* **2018,** *263*, 442-53.
88. Giacomoni F., Le Corguille G.*, et al.*, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **2015,** *31* (9), 1493-5.
89. Caballero-Casero N., Belova L.*, et al.*, Towards harmonised criteria in quality assurance and quality control of suspect and non-target LC-HRMS analytical workflows for screening of emerging contaminants in human biomonitoring. *TrAC Trends in Analytical Chemistry* **2021,** *136*.
90. Monteiro Bastos da Silva J., Chaker J.*, et al.*, Improving Exposure Assessment Using Non-Targeted and Suspect Screening: The ISO/IEC 17025: 2017 Quality Standard as a Guideline. *Journal of Xenobiotics* **2021,** *11* (1), 1-15.
91. Klavus A., Kokla M.*, et al.*, "notame": Workflow for Non-Targeted LC-MS Metabolic Profiling. *Metabolites* **2020,** *10* (4).
92. Ribbenstedt A., Ziarrusta H.*, et al.*, Development, characterization and comparisons of targeted and non-targeted metabolomics methods. *PLoS One* **2018,** *13* (11), e0207082.
93. Pourchet M., Debrauwer L.*, et al.*, Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. *Environ Int* **2020,** *139*, 105545.
94. Barri T.,Dragsted L.O., UPLC-ESI-QTOF/MS and multivariate data analysis for blood plasma and serum metabolomics: Effect of experimental artefacts and anticoagulant. *Analytica Chimica Acta* **2013,** *768*, 118-28.
95. Rico E., Gonzalez O.*, et al.*, Evaluation of human plasma sample preparation protocols for untargeted metabolic profiles analyzed by UHPLC-ESI-TOF-MS. *Anal Bioanal Chem* **2014,** *406* (29), 7641-52.
96. Tulipani S., Mora-Cubillos X.*, et al.*, New and vintage solutions to enhance the plasma metabolome coverage by LC-ESI-MS untargeted metabolomics: the not-so-simple process of method performance evaluation. *Anal Chem* **2015,** *87* (5), 2639-47.
97. Tulipani S., Llorach R.*, et al.*, Comparative analysis of sample preparation methods to handle the complexity of the blood fluid metabolome: when less is more. *Anal Chem* **2013,** *85* (1), 341-8.
98. Vuckovic D., Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Anal Bioanal Chem* **2012,** *403* (6), 1523-48.
99. David A., Abdul-Sada A.*, et al.*, A new approach for plasma (xeno)metabolomics based on solid-phase extraction and nanoflow liquid chromatography-nanoelectrospray ionisation mass spectrometry. *J Chromatogr A* **2014,** *1365*, 72-85.
100. Ahmad S., Kalra H.*, et al.*, HybridSPE: A novel technique to reduce phospholipid-based matrix effect in LC-ESI-MS Bioanalysis. *J Pharm Bioallied Sci* **2012,** *4* (4), 267-75.
101. Sitnikov D.G., Monnin C.S.*, et al.*, Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. *Sci Rep* **2016,** *6*, 38885.
102. Ramesh B., Manjula N.*, et al.*, Comparison of conventional and supported liquid extraction methods for the determination of sitagliptin and simvastatin in rat plasma by LC-ESI-MS/MS. *J Pharm Anal* **2015,** *5* (3), 161-8.
103. Bligh E.G.,Dyer W.J., A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology* **1959,** *8* (37).
104. Folch J., Lees M.*, et al.*, A Simple Method for the Isolation and Purification of Total Lipides from Animal Tissues. *Journal of Biological Chemistry* **1957,** *226* (1), 497-509.

105. Ulmer C.Z., Jones C.M.*, et al.*, Optimization of Folch, Bligh-Dyer, and Matyash sample-to-extraction solvent ratios for human plasma-based lipidomics studies. *Anal Chim Acta* **2018,** *1037*, 351-7.
106. Fei F., Bowdish D.M.*, et al.*, Comprehensive and simultaneous coverage of lipid and polar metabolites for endogenous cellular metabolomics using HILIC-TOF-MS. *Anal Bioanal Chem* **2014,** *406* (15), 3723-33.
107. Whiley L., Godzien J.*, et al.*, In-vial dual extraction for direct LC-MS analysis of plasma for comprehensive and highly reproducible metabolic fingerprinting. *Anal Chem* **2012,** *84* (14), 5992-9.
108. Bessonneau V., Ings J.*, et al.*, In vivo microsampling to capture the elusive exposome. *Sci Rep* **2017,** *7*, 44038.
109. Vuckovic D.,Pawliszyn J., Systematic evaluation of solid-phase microextraction coatings for untargeted metabolomic profiling of biological fluids by liquid chromatography-mass spectrometry. *Anal Chem* **2011,** *83* (6), 1944-54.
110. Smith C.A., Want E.J.*, et al.*, XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006,** (78), 779-87.
111. Tautenhahn R., Patti G.J.*, et al.*, XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem* **2012,** *84* (11), 5035-9.
112. Pluskal T., Castillo S.*, et al.*, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**.
113. Tsugawa H., Cajka T.*, et al.*, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.
114. Tugizimana F., Steenkamp P.A.*, et al.*, A Conversation on Data Mining Strategies in LC-MS Untargeted Metabolomics: Pre-Processing and Pre-Treatment Steps. *Metabolites* **2016,** *6* (4).
115. Myers O.D., Sumner S.J.*, et al.*, Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal Chem* **2017,** *89* (17), 8689-95.
116. Alboniga O.E., Gonzalez O.*, et al.*, Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* **2020,** *16* (1), 14.
117. Li Z., Lu Y.*, et al.*, Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Anal Chim Acta* **2018,** *1029*, 50-7.
118. Rafiei A.,Sleno L., Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Commun Mass Spectrom* **2015,** *29* (1), 119-27.
119. Brunius C., Shi L.*, et al.*, Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* **2016,** *12* (11), 173.
120. Veselkov K.A., Vingara L.K.*, et al.*, Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal Chem* **2011,** *83* (15), 5864-72.
121. Prince J.T.,Marcotte E.M., Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal Chem* **2006,** *78*, 6140-52.
122. Dunn W., Wilson I.*, et al.*, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012,** *4* (18), 2249–64.
123. Evans A.M., O'Donovan C.*, et al.*, Dissemination and analysis of the quality assurance (QA) and quality control (QC) practices of LC-MS based untargeted metabolomics practitioners. *Metabolomics* **2020,** *16* (10), 113.
124. Want E.J., Wilson I.D.*, et al.*, Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* **2010,** *5* (6), 1005-18.

125. Fernandez-Ochoa A., Quirantes-Pine R.*, et al.*, A Case Report of Switching from Specific Vendor-Based to R-Based Pipelines for Untargeted LC-MS Metabolomics. *Metabolites* **2020,** *10* (1).
126. Vinaixa M., Samino S.*, et al.*, A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites* **2012,** *2* (4), 775-95.
127. Santos S., Maitre L.*, et al.*, Applying the exposome concept in birth cohort research: a review of statistical approaches. *Eur J Epidemiol* **2020,** *35* (3), 193-204.
128. Agier L., Portengen L.*, et al.*, A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations. *Environ Health Perspect* **2016,** *124* (12), 1848-56.
129. Di Guida R., Engel J.*, et al.*, Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **2016,** *12*, 93.
130. Schymanski E.L., Jeon J.*, et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.
131. Bonini P., Kind T.*, et al.*, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020,** *92* (11), 7515-22.
132. Aalizadeh R., Nika M.C.*, et al.*, Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *J Hazard Mater* **2019,** *363*, 277-85.
133. Cao M., Fraser K.*, et al.*, Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **2015,** *11* (3), 696-706.
134. Stanstrup J., Neumann S.*, et al.*, PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. *Anal Chem* **2015,** *87* (18), 9421-8.
135. Moschet C., Piazzoli A.*, et al.*, Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. *Anal Chem* **2013,** *85* (21), 10312-20.
136. Kuhl C., Tautenhahn R.*, et al.*, CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* **2012,** *84* (1), 283-9.
137. Silva R.R., Jourdan F.*, et al.*, ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics* **2014,** *30* (9), 1336-7.
138. Daly R., Rogers S.*, et al.*, MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **2014,** *30* (19), 2764-71.
139. Ernst M., Kang K.B.*, et al.*, MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* **2019,** *9* (7).
140. Uppal K., Walker D.I.*, et al.*, xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal Chem* **2017,** *89* (2), 1063-7.
141. Lawson T.N., Weber R.J.*, et al.*, msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Anal Chem* **2017,** *89* (4), 2432-9.
142. Yin Y., Wang R.*, et al.*, DecoMetDIA: Deconvolution of Multiplexed MS/MS Spectra for Metabolite Identification in SWATH-MS-Based Untargeted Metabolomics. *Anal Chem* **2019,** *91* (18), 11897-904.
143. Wishart D.S., Feunang Y.D.*, et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **2018,** *46* (D1), D608-D17.
144. Kanehisa M., The KEGG database. *Novartis Found Symp* **2002,** *247*, 91-101; discussion -3, 19-28, 244-52.
145. Barupal D.K.,Fiehn O., Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. *Environmental Health Perspectives* **2019,** *127* (9).

146. Neveu V., Moussy A.*, et al.*, Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res* **2017,** *45* (D1), D979-D84.
147. Meijer J., Lamoree M.*, et al.*, An annotation database for chemicals of emerging concern in exposome research. *Environ Int* **2021,** *152*, 106511.
148. Ruttkies C., Schymanski E.L.*, et al.*, MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **2016,** *8*, 3.
149. Allen F., Pon A.*, et al.*, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* **2014,** *42* (Web Server issue), W94-9.
150. Oberacher H., Sasse M.*, et al.*, A European proposal for quality control and quality assurance of tandem mass spectral libraries. *Environmental Sciences Europe* **2020,** *32* (1).
151. Horai H., Arita M.*, et al.*, MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **2010,** *45* (7), 703-14.
152. McEachran A.D., Mansouri K.*, et al.*, A comparison of three liquid chromatography (LC) retention time prediction models. *Talanta* **2018,** *182*, 371-9.
153. Barron L.P.,McEneff G.L., Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods. *Talanta* **2016,** *147*, 261-70.
154. Aalizadeh R., Thomaidis N.S.*, et al.*, Quantitative Structure-Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. *J Chem Inf Model* **2016,** *56* (7), 1384-98.
155. Bade R., Bijlsma L.*, et al.*, Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. *Talanta* **2015,** *139*, 143-9.
156. McCord J.P., Groff L.C.*, et al.*, Quantitative non-targeted analysis: Bridging the gap between contaminant discovery and risk characterization. *Environment International* **2022,** *158*.
157. Plassmann M.M., Fischer S.*, et al.*, Nontarget Time Trend Screening in Human Blood. *Environmental Science & Technology Letters* **2018,** *5* (6), 335-40.
158. Hollender J., Schymanski E.L.*, et al.*, Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ Sci Technol* **2017,** *51* (20), 11505-12.
159. Kruve A., Strategies for Drawing Quantitative Conclusions from Nontargeted Liquid Chromatography-High-Resolution Mass Spectrometry Analysis. *Anal Chem* **2020,** *92* (7), 4691-9.
160. Liigand J., Wang T.*, et al.*, Quantification for non-targeted LC/MS screening without standard substances. *Sci Rep* **2020,** *10* (1), 5808.
161. Liigand P., Liigand J.*, et al.*, Ionisation efficiencies can be predicted in complicated biological matrices: A proof of concept. *Anal Chim Acta* **2018,** *1032*, 68-74.
162. Dumas T., Courant F.*, et al.*, Environmental Metabolomics Promises and Achievements in the Field of Aquatic Ecotoxicology: Viewed through the Pharmaceutical Lens. *Metabolites* **2022,** *12* (2).
163. Haug K., Cochrane K.*, et al.*, MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* **2020,** *48* (D1), D440-D4.
164. Sud M., Fahy E.*, et al.*, Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **2016,** *44* (D1), D463-70.
165. Schymanski E.L., Kondic T.*, et al.*, Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *J Cheminform* **2021,** *13* (1), 19.
166. Vrijheid M., Slama R.*, et al.*, The human early-life exposome (HELIX): project rationale and design. *Environ Health Perspect* **2014,** *122* (6), 535-44.

167. Maitre L., de Bont J.*, et al.*, Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. *BMJ Open* **2018,** *8* (9), e021311.
168. Julvez J., Lopez-Vicente M.*, et al.*, Early life multiple exposures and child cognitive function: A multi-centric birth cohort study in six European countries. *Environ Pollut* **2021,** *284*, 117404.
169. Vineis P., Chadeau-Hyam M.*, et al.*, The exposome in practice: Design of the EXPOsOMICS project. *Int J Hyg Environ Health* **2017,** *220* (2 Pt A), 142-51.
170. Chang C.J., Barr D.B.*, et al.*, Per- and polyfluoroalkyl substance (PFAS) exposure, maternal metabolomic perturbation, and fetal growth in African American women: A meet-in-the-middle approach. *Environ Int* **2022,** *158*, 106964.
171. Niedzwiecki M.M.,Miller G.W., HERCULES: An Academic Center to Support Exposome Research. In *Unraveling the Exposome*, 2019; pp 339-48.
172. Balshaw D.M., Collman G.W.*, et al.*, The Children's Health Exposure Analysis Resource: enabling research into the environmental influences on children's health outcomes. *Curr Opin Pediatr* **2017,** *29* (3), 385-9.
173. Viet S.M., Falman J.C.*, et al.*, Human Health Exposure Analysis Resource (HHEAR): A model for incorporating the exposome into health studies. *Int J Hyg Environ Health* **2021,** *235*, 113768.
174. Apel P., Rousselle C.*, et al.*, Human biomonitoring initiative (HBM4EU) - Strategy to derive human biomonitoring guidance values (HBM-GVs) for health risk assessment. *Int J Hyg Environ Health* **2020,** *230*, 113622.
175. Lange R., Apel P.*, et al.*, The European Human Biomonitoring Initiative (HBM4EU): Human biomonitoring guidance values for selected phthalates and a substitute plasticizer. *Int J Hyg Environ Health* **2021,** *234*, 113722.
176. Lamkarkach F., Ougier E.*, et al.*, Human biomonitoring initiative (HBM4EU): Human biomonitoring guidance values (HBM-GVs) derived for cadmium and its compounds. *Environ Int* **2021,** *147*, 106337.
177. Vrijheid M., Basagana X.*, et al.*, Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview. *Environ Epidemiol* **2021,** *5* (5), e166.
178. Vlaanderen J., de Hoogh K.*, et al.*, Developing the building blocks to elucidate the impact of the urban exposome on cardiometabolic-pulmonary disease: The EU EXPANSE project. *Environ Epidemiol* **2021,** *5* (4), e162.
179. Bessonneau V., Gerona R.R.*, et al.*, Gaussian graphical modeling of the serum exposome and metabolome reveals interactions between environmental chemicals and endogenous metabolites. *Sci Rep* **2021,** *11* (1), 7607.
180. Walker D.I., Juran B.D.*, et al.*, High-Resolution Exposomics and Metabolomics Reveals Specific Associations in Cholestatic Liver Diseases. *Hepatol Commun* **2021**.
181. Chung M.K., Buck Louis G.M.*, et al.*, Exposome-wide association study of semen quality: Systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes. *Environ Int* **2019,** *125*, 505-14.
182. Rathahao-Paris E., Alves S.*, et al.*, High resolution mass spectrometry for structural identification of metabolites in metabolomics. *Metabolomics* **2015,** *12* (1).
183. Anna S., Sofia B.*, et al.*, The dilemma in prioritizing chemicals for environmental analysis: known versus unknown hazards. *Environ Sci Process Impacts* **2016,** *18* (8), 1042-9.
184. Dix D.J., Houck K.A.*, et al.*, The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* **2007,** *95* (1), 5-12.
185. Barouki R., Audouze K.*, et al.*, The exposome and toxicology: a win-win collaboration. *Toxicol Sci* **2021**.
186. Loo R.L., Chan Q.*, et al.*, A comparison of self-reported analgesic use and detection of urinary ibuprofen and acetaminophen metabolites by means of metabonomics: the INTERMAP Study. *Am J Epidemiol* **2012,** *175* (4), 348-58.

Chapter II. Material and methods

1. Instrumental method development and optimization

A LC-ESI-HRMS SCIEX ExionLC™ Ultra-High Performance Liquid Chromatography (UHPLC) system (Framingham, USA) coupled to a high-resolution QTOF mass spectrometer SCIEX X500R equipped with a Turbo V ion source with a twin-sprayer ESI probe and a hybrid quadrupole time-of-flight mass spectrometer was used for all experiments. External calibration was systematically performed by infusion of AB SCIEX calibration mixtures for negative and positive ionization modes before all injections. The instrument was controlled by SCIEX OS software version 1.2. LC optimizations and development (e.g. columns, flow rates, solvent of injection) were made using a mix of standards spiked in solvent and plasma/serum to ensure a good analytical sensitivity and repeatability.

1.1. Mix of standards used for the optimizations

One of the main challenges of non-targeted method development is the width and depth of the chemical space intended to be observed. Indeed, compounds constituting the chemical exposome are extremely varied in both physical-chemical properties and concentrations in biological matrices. While there are indubitably less constraints in non-targeted methods regarding quantification performances compared to targeted methods, there is an added difficulty in ensuring a high coverage of the observable chemical space to characterize the chemical exposome as thoroughly as possible given the chosen analytical system.

To achieve this goal, a mix of 50 compounds, referred to as the optimization mix, was designed. These compounds were chosen to meet three main objectives:

(i) Belong to different chemical classes of interest in the context of an exposomics application in human biological matrices (i.e. endogenous compounds such as steroids and eicosanoids, and exogenous compounds such as pesticides and drugs).

(ii) Represent a wide range of physical-chemical properties (i.e. m/z and polarity) to cover the entire space of the LC method.

(iii) Cover both ESI (+) and ESI (−) ionization modes.

An overview of this compound set is presented in Figure II., while a detailed list is available in Appendix 1.1. To summarize, chosen compounds are distributed as follows: 14 endogenous compounds (1 neurotransmitter, 6 steroids and 7 eicosanoids) and 36 exogenous compounds (2 food compounds, 13 drugs, 19 pesticides, and 2 environmental pollutants linked to smoking). These compounds present monoisotopic mass values ranging between 133.0640 and 496.2607 Da and octanol-water partition coefficients (logP) ranging between 0.07 and

6.99. Overall, 36 compounds are better observed in ESI (+) mode, while 14 are better observed in ESI (−) mode.

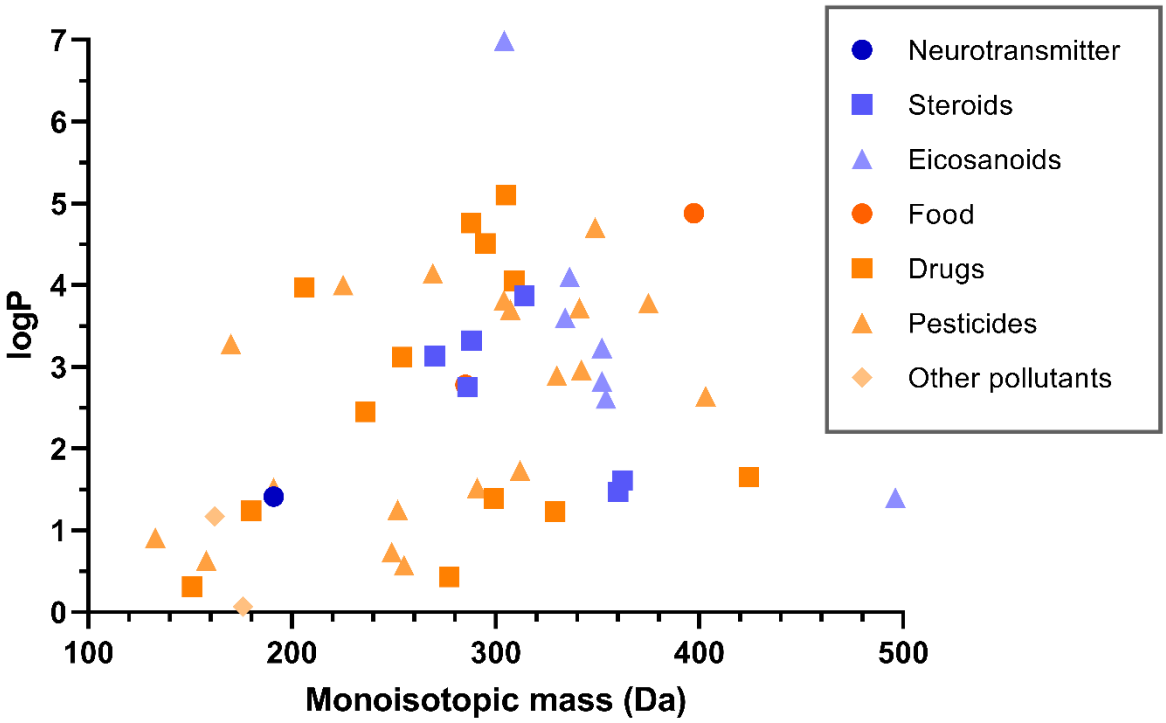

Figure II.1 – Overview of the physical-chemical properties of the 50-compound optimization mix, including endogenous compounds (in blue) and exogenous compounds (in orange). The octanol-water partition coefficient (logP) and the monoisotopic mass (Da) are presented.

This optimization mix was prepared at 1 µg/mL in methanol, and diluted and reconstituted in the optimized reconstitution phase (see paragraph 1.3.2) as needed for sample spiking or for injection in solvent. Usually, mix concentration was kept between 0.1 and 100 ng/mL in vial (whether in solvent or in matrix) to avoid excessive system contamination.

1.2. Quality assurance and quality control procedures

Several quality assurance/quality control procedures were implemented for non-targeted analyses. One solvent blank (i.e. acetonitrile/ultrapure water 90:10 (v/v)) and one extraction blank sample (i.e. preparation with UHPLC grade water instead of sample) were systematically injected with each batch. This allowed ensuring lack of carryover in the UHPLC system and monitoring the contamination linked to the sample preparation process respectively. Contamination linked to the sample preparation process for annotated compounds in particular was taken into account by verifying their presence in the extraction blank, and if so, subtracting the blank level from samples. Additionally, composite quality control samples were prepared and injected after blanks to equilibrate the analytical system, and periodically throughout the

batch (i.e. every 5-7 samples) to monitor the analytical drift and repeatability. In the case of multiple batches, samples were assigned randomly, and samples were injected randomly in all cases. Internal standards were systematically used in samples and monitored to assess analytical drift. MS2 acquisitions were performed at the end of each batch to generate fragmentation data for the annotation process.

1.3. LC method optimization

In non-targeted approaches, the chromatographic separation is important to optimize to ensure that sufficient chromatographic separation is achieved. Optimizing the LC method parameters is a crucial step, since ESI sources are prone to phenomena such as ion suppression, particularly in complex biological matrix. To limit the impact of ion suppression and maximize sensitivity performances, the LC method should be optimized to reduce co-elution, which can be done by increasing chromatographic dilution for instance.

A base gradient was set as follows for a flow rate of 0.1 mL/min: 0-2.5 min, 10-20% B; 2.5-20 min, 20-30% B; 20-38 min, 30-45% B; 38-45 min, 45-100% B; 45-55 min, 100% B; 55-60 min, 10% B, for a total run time of 60 minutes. While run times in metabolomics are typically shorter (from 15 to 30 minutes[1-3]), some studies rely on longer methods (from 45 to 85 minutes[4, 5]) to increase Rt stability or to ensure sufficient separation between isomers through less steep gradients. Moreover, the need for high sensitivity often entails lower flow rates[6] for better sample decomplexification[7], which in turn leads to longer run times. A 60-minute run was determined adequate as it notably allowed to separate isomers prostaglandins D2 and E2 (Rt values of 16.25 min and 15.52 min respectively). The test of different flow rates was performed with comparable gradients, with the adjustment of times to allow the flow of an identical amount of solvent.

Using the optimization mix, three parameters of the LC method were then optimized; firstly, two reverse phase columns (both 1.8µm, 150mm Acquity HSS T3, Waters, with diameters of 2.1mm and 1.0mm) were tested. These assays were done conjointly with the flow rate optimization, as two flow rates were tested for each column. Lastly, the organic phase percentage of the reconstitution solvent was optimized.

1.3.1. Column diameter and flow rate optimization

The optimization mix was spiked post-extraction (protein precipitation) in a serum homogenate and injected in two quantities (20 and 200 pg) using a 2.1 mm diameter column and a 1.0 mm diameter column. Each column was tested using two flow rates (0.3 and 0.15 mL/min for the 2.1 mm column, and 0.1 and 0.05 mL/min for the 0.1mm column). Compounds' areas were

integrated manually using SCIEX OS, and area coefficient of variation (CV) values were calculated from four replicate injections. A generic elution gradient of water (A) and acetonitrile (B) both supplemented with 0.01% of formic acid was used. Oven temperature was maintained at 40°C for all experiments. The results are summarized in Table II.1, and detailed results are available in Appendix 1.2.

	Ø 2.1 mm		**Ø 1.0 mm**	
	0.30 mL/min	0.15 mL/min	0.10 mL/min	0.05 mL/min
20 pg	1.17 e+4 (2.4%)	1.96 e+4 (2.1%)	2.75 e+4 (2.0%)	4.42 e+4 (2.0%)
200 pg	1.70 e+5 (2.1%)	2.50 e+5 (1.7%)	3.08 e+5 (1.7%)	3.80 e+5 (2.2%)

Table II.1 – Median area (and area CV) of compounds from the optimization mix injected in four replicate depending on the column diameter and flow rate.

Although results were compound-dependent, the overall trend showed that area values increased as flow rate (and column diameter) decreased. This led to the favoring of the 1.0 mm diameter column, as sufficient pressure could be achieved using lower flow rates. However, it was also observed that using this column, area repeatability and retention time stability were significantly improved using the 0.10 mL/min flow rate compared to 0.05 mL/min. Retention time being a key factor for accurate binning during the data processing, and because sensitivity was already improved with this flow rate, the 0.10 mL/min flow rate with the 1.0 mm column were kept as analytical conditions.

1.3.2. Reconstitution phase optimization

Prior to the injection, samples are often evaporated then reconstituted for conservation, concentration or composition purposes. The reconstitution phase composition's impact on the metabolome coverage in non-targeted analyses has been demonstrated, and more precisely the relevance of using 100% water as a reconstitution phase compared to 100% methanol and 50:50 water:methanol[8]. For this optimization, acetonitrile was preferred as it is the gradient's organic phase. Considering the range of polarities present in the mix, seven compositions were compared (i.e. from 25:75 to 100:0 water:acetonitrile). The comparison was performed on serum homogenates prepared by protein precipitation and spiked with the optimization mix at 100 ng/mL. Two parameters were determined:

(i) the percentage of compounds which attained the largest area with each reconstitution phase, and

(ii) the percentage of compounds with areas above the median area for all reconstitution phases.

The results are summarized in Table II.2. It was observed that 31% of compounds had the largest area value using the 70:30 (water:acetonitrile) composition, which was the best performance for this criterion. However, this condition is moderately distant from the chromatographic method's initial conditions (90:10 water:acetonitrile), which significantly affected peak shapes for some compounds as shown in Figure II.2.A. This would be an issue for the data processing step, as such an irregular peak shape would lead to poorer integrations.

Reconstitution phase (water:acetonitrile)	**Percentage of compounds with largest area for each composition (%)**	**Percentage of compounds with area for each composition above the median area (%)**
25:75	13	29
50:50	2	27
60:40	4	35
70:30	31	60
80:20	19	69
90:10	21	54
100:0	10	25

Table II.2 – Impact of the reconstitution phase composition on areas of 50 compounds spiked in serum homogenates and injected on UHPLC-ESI-QTOF in positive and negative ionization modes.

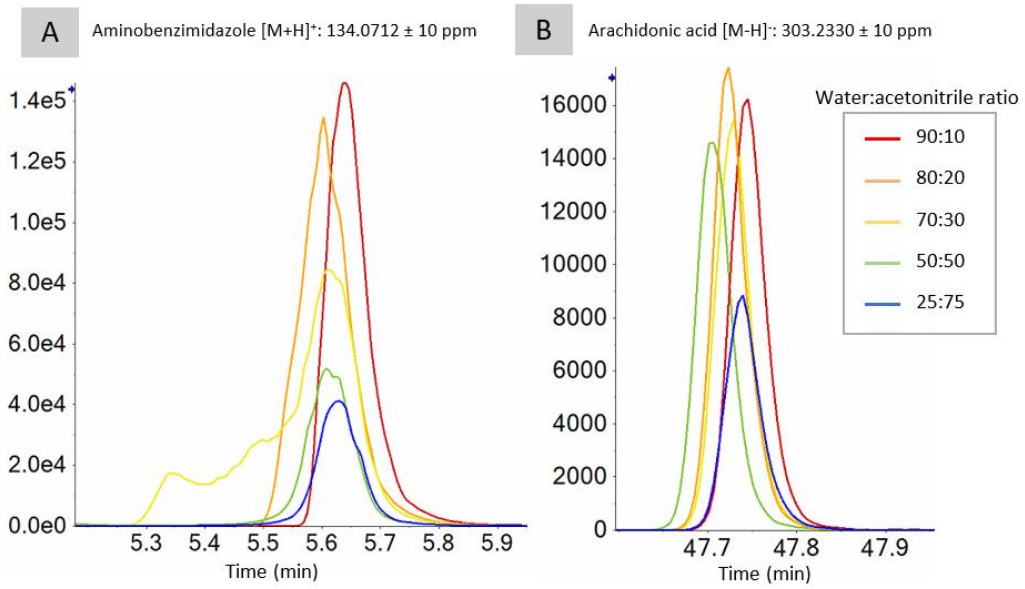

Figure II.2 – Extracted ion chromatogram for Aminobenzimidazole (logP = 0.91) in ESI (+) mode (A) and Arachidonic acid (logP = 6.98) in ESI (−) mode (B) depending on the reconstitution phase composition (generated with a m/z tolerance of 10 ppm).

It was then established that 69% of compounds had areas in the 80:20 (water:acetonitrile) phase above the median area in all reconstitution phases. This composition allowed retaining a satisfying peak shape as it was closer to the initial chromatographic conditions, and was therefore kept as the optimized reconstitution phase composition.

It is worthy to note that the 100% water condition was not the best reconstitution phase in this case, contrary to what was suggested in the literature, although it still presented significantly better results than the 50:50 composition. This difference with the literature may be explained by the difference in the organic phase (acetonitrile here versus methanol in the literature), as well as the fact that in this work, the effects of reconstitution phase composition were evaluated on a set of compounds, as opposed to being done at the non-targeted scale as presented in the litterature[8].

Overall, optimization of the chromatographic separation was not the main aim of this PhD but as a critical step in non-targeted LC-HRMS analyses, it was important to ensure that it was possible to observe a wide range of compounds using this method. Moreover, it was also important to check the repeatability of the retention time using the mix spiked in blood serum for this method as it could affects further data processing steps such as compound annotation, which may rely on such a parameter.

1.4. MS optimization

1.4.1. MS acquisition

Full-scan mass spectra was acquired in both – and + ESI modes for all samples. The mass range was set between 50-1100 m/z. The MS analysis was performed using original ESI source settings: temperature 550°C, ionspray voltage 4,5kV (-4,5kV in negative mode), declustering potential 80V (-80V in negative mode), accumulation time 300 ms, spray N_2 gas 35 arbitrary units, heat conduction gas 35 arbitrary units, curtain gas 7 arbitrary units, collisionally activated dissociation gas 7 arbitrary units. Run time was set at 60 min in coherence with the LC method. For all the experiments in this PhD, injections of samples were always performed first in full scan to obtain the most comprehensive chemical fingerprint without affecting the sensitivity as explained below, and then a selection of samples were re-injected using MS2 for further work on structural elucidation.

1.4.2. MS2 acquisition

MS2 acquisitions were performed in addition to MS acquisitions on randomly selected samples. The choice to separate these two acquisitions was made to obtain higher accumulation times for both analyses, thus attaining better sensitivity performances. Sensitivity

was prioritized over run time, as the aim was to constitute digital archives possibly extensively re-usable, with accurate chemical information even for low-abundant compounds. MS2 acquisitions were performed either using data dependent acquisition or data independent acquisitions.

1.4.2.1. Data dependent acquisition

Data dependent acquisition was performed using SCIEX's Information Dependent Acquisition (IDA) methods. IDA experiments allow data analysis concomitantly to its acquisition, changing conditions accordingly; the selection of precursor ions on which dependent scans are performed is made during the analysis. This results in the acquisition of often high quality fragmentation spectra on a selected number of precursors. Since the aim was to obtained MS2 data for the highest numbers of chemicals potentially present at low concentrations (e.g. exogenous chemicals), the number of maximum precursor ions per scan was optimized by comparing four threshold values (i.e. 10, 20, 50 and 100) for the MS2 analysis of the 50-compound optimization mix at 10 ng/mL in plasma. Three parameters were compared: firstly, the percentage of compounds successfully triggering MS2 analysis, secondly, the percentage of compounds for which a usable MS2 spectra is obtained (i.e. intensity of at least one fragment over 20 counts per second), and thirdly, the median number of spectra acquired by compound. The results are summarized in Table II.3 below.

Maximum precursor ions (per scan)	**Compounds triggering MS2 analysis (%)**	**Compounds for which a usable spectra is obtained (%)**	**Median number of acquired spectra by compound**
10	52	46	4
20	84	78	4
50	80	60	3
100	80	48	1

Table II.3 – Results of the Information Dependent Analysis (IDA) method optimization through the selection of adequate maximum precursor ions per scan for the mix injected at 10ng/ml on QTOF in ESI (−) and ESI (+) modes

It was observed more than 80% of compounds spiked at 10 ng/ml in plasma triggered MS2 analysis whenever the maximum number of precursor ions per scan was set above 20. On the other hand, the best performance on number of acquired spectra by compound was achieved with lower thresholds. Better performance for higher thresholds was expected, as low-abundant compounds in complex matrices are more likely to be picked when a higher number of candidate ions are authorized. However, legible spectra for those compounds were mostly obtained at lower thresholds. This may be explained by the fact that the set MS2 accumulation

time of 100 ms per scan was divided between fewer acquisitions in the case of lower thresholds, thus resulting in better sensitivity. This led to the choosing of 20 maximum precursor ions per scan.

IDA experiments were performed in both ESI (–) and (+) modes, using the following source settings: MS1 accumulation time 250 ms, MS2 accumulation time 100 ms, collision energy 35 eV) and ESI (+) ionization modes is presented in Table II.4.

Window index	ESI (−)	ESI (+)
1	49.5 – 59.4	49.5 – 60.5
2	58.4 – 68.9	59.5 – 74.5
3	67.9 – 85.7	73.5 – 80.0
4	84.7 – 114.0	79.0 – 99.3
5	113.0 – 149.7	98.3 – 109.3
6	148.7 – 177.8	108.3 – 135.0
7	176.8 – 200.7	134.0 – 161.8
8	199.7 – 245.7	160.8 – 199.6
9	244.7 – 269.4	198.6 – 240.4
10	268.4 – 310.9	239.4 – 268.9
11	309.0 – 323.5	267.9 – 324.5
12	322.5 – 346.1	323.5 – 367.8
13	345.1 – 388.1	366.8 – 395.9
14	387.1 – 454.2	394.9 – 425.6
15	453.2 – 515.1	424.6 – 474.7
16	514.1 – 569.7	473.7 – 506.7
17	568.7 – 593.4	505.7 – 533.0
18	592.4 – 677.9	532.0 – 577.1
19	676.9 – 844.6	576.1 – 771.8
20	844.6 – 999.9	770.8 – 999.9

Table II.4 - Example of SWATH windows generated by the vendor SWATH windows calculator on plasma quality control samples in ESI (−) and ESI (+) ionization modes

SWATH experiments were performed in both – and + ESI modes, using the following source settings: MS1 accumulation time 80 ms, MS2 accumulation time 30 ms, collision energy set as a ramp evolving from 20 to 50 eV (35±15 eV), cycle time 469 ms, mass range 50-1100 m/z.

2. Sample preparation methods for non-targeted exposomics

This section aims to provide more details on the principle of the different techniques used in this PhD, however, the thorough investigation of the impact of sample preparation on the extraction of the components of the chemical exposome using blood plasma and blood serum

samples will be done in Chapter III. As non-targeted methods aim to accurately detect a high number of unknown compounds in a given sample, the choice of a sample preparation technique is particularly challenging. Indeed, it is often recommended that non-targeted approaches rely on minimal sample preparation procedure to avoid loss of potential compounds of interest. However, when exploring the chemical exposome with complex biological matrices using LC, issues such as ion suppression may arise, resulting in a need for efficient sample purification. Moreover, human biological matrices are often only available in small quantities, meaning that SPM should use minimal matrix amount while allowing sufficient concentration to keep high sensitivity performances.

Based on the methods used in the literature, the investigation of the chemical exposome using blood plasma and blood serum samples may be done using, at least, four major types of SPM, from least to most selective: protein precipitation (PPT), supported liquid extraction (SLE), protein and phospholipid removal (PLR), and solid phase extraction (SPE). As mentioned earlier, a systematic evaluation of the impact of the SPM for non-targeted exposomics analyses is presented in Chapter III; the following paragraphs introduce the advantages of each type of SPM through a generic outline of the associated protocol. They each offer a different balance between sensitivity and selectivity, thus potentially offering a different vision of the chemical space.

2.1. Protein precipitation

The use of PPT methods is widespread in both metabolomics and recent exposomics applications[2, 3, 7]. It is the least selective of all the listed SPM types, as it only consists of precipitation the proteins present in the sample with a solvent (often methanol, acetonitrile, or a mixture of both) used at a 1:1 to 4:1 ratio compared to the sample volume[9-13]. The mixtures are then left for one hour at -20°C to allow precipitation to occur, after which a centrifugation is performed. The operating principle is schematized in Figure II.3.The supernatants are then collected and may be evaporated and reconstituted as needed, usually with a concentration factor of 1 to 3[9-13].

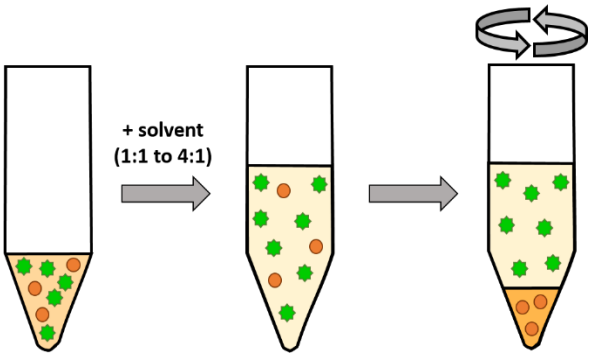

Figure II.3 – Operating principle of protein precipitation. Samples are mixed with an organic solvent (usually methanol or acetonitrile) at a solvent:sample ratio of 1:1 to 4:1. After a prolonged contact, centrifugation allows forming a protein pellet (in orange) and the purified supernatant can be used.

2.2. Protein and Phospholipid removal

PLR methods allow an additional sample purification compared to PPT, as phospholipids and lysophospholipids elimination is performed in addition to protein removal. PLR methods are performed by using a solid phase in a cartridge or a plate as a filter through which the sample (sometimes diluted with an organic solvent) must be passed. Most proteins, phospholipids and lysophospholipids should be retained on the stationary phase and leave a purified sample. Vendors often recommend a prior deproteinization of the sample to avoid saturation of the packed-bed structure. Acidification of the sample (e.g. 1% with formic acid) is also recommended to help protein precipitation. Moreover, vendors often recommend using a specific solvent to maximize performance. A schematized protocol is presented in Figure II.4.

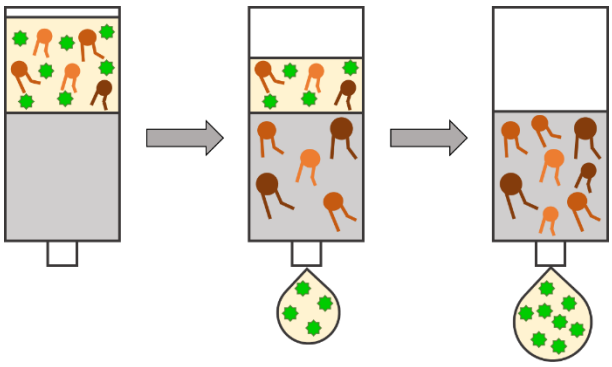

Figure II.4 – Operating principle of solid phase extraction. Samples are filtered through a stationary phase that retains phospholipids (in shades of orange) and leaves other compounds (in green) pass through.

Following this protocol, the filtrate can be evaporated and reconstituted with a concentration factor varying between 1 and 5[14, 15]. It is usually possible to increase the concentration factor compared to PPT, as further sample purification is achieved, resulting in less concern for

clogging, carry-over and matrix effect. The exact retention mechanism of the sorbents is not known, although some of them are hypothesized to retain the phosphate group inherent with all phospholipids with zirconia atoms on the stationary phase through Lewis acid-base interactions[16]. However, other mechanisms (such as apolar retention) may affect the retention of compounds. This will be systematically evaluated in Chapter III. Systematic evaluations of blood-derived sample preparation methods for HRMS-based chemical exposomics

2.3. Supported Liquid Extraction

SLE is also performed using a solid sorbent, which in this case acts as an interface between two immiscible liquid phases. The whole sample is loaded on the sorbent, which the aqueous sample soaks. As the entirety of the sample is retained on the sorbent, it is critical to ensure that a sufficient amount is used to soak the total volume. The sorbent is then washed using the extraction solvent, selectively eluting the analytes. The extraction solvent is often hexane, ethyl acetate, or methyl tert-butyl ether (MTBE), as they are immiscible with aqueous matrices. In the context of this PhD, only the Isolute SLE (Biotage) was used with MTBE, as per the vendor's recommendations. Compounds that have a high affinity to the extraction solvent will be carried, while other compounds will be retained by the solid media. A schematized operating principle is presented in Figure II.5.

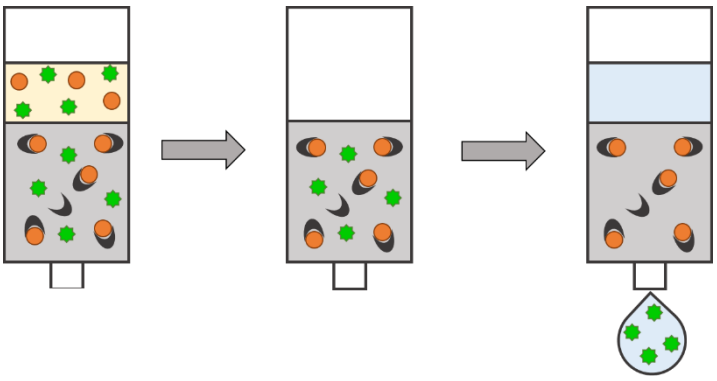

Figure II.5 - Operating principle of supported liquid extraction. The sample is loaded onto a sorbent, which retains the entire sample. The analytes are then selectively eluted using an immiscible organic solvent.

2.4. Solid Phase Extraction

SPE is a selective SPM that is performed using a multi-step protocol in order to remove interferents (e.g. proteins, salts) and concentrate potential compounds of interest. It first requires conditioning the solid phase, followed by sample loading. The solid phase is then rinsed with an aqueous solvent to eliminate interferents, and dried. Lastly, an extraction solvent is used to recover compounds of interest previously retained by the solid phase. A schematized operating principle is presented in

Figure II.6. This standard protocol contains significantly more steps than any other mentioned SPM, which may lead to poorer repeatability. However, it provides significant sample purification, and is traditionally used in targeted approaches to improve sensitivity. It is therefore a key type of SPM to evaluate when using human biological matrices. As eluates have a high purity level, similar concentration ratios to those used for PLR are considered, i.e. between 1 and 5[9-11].

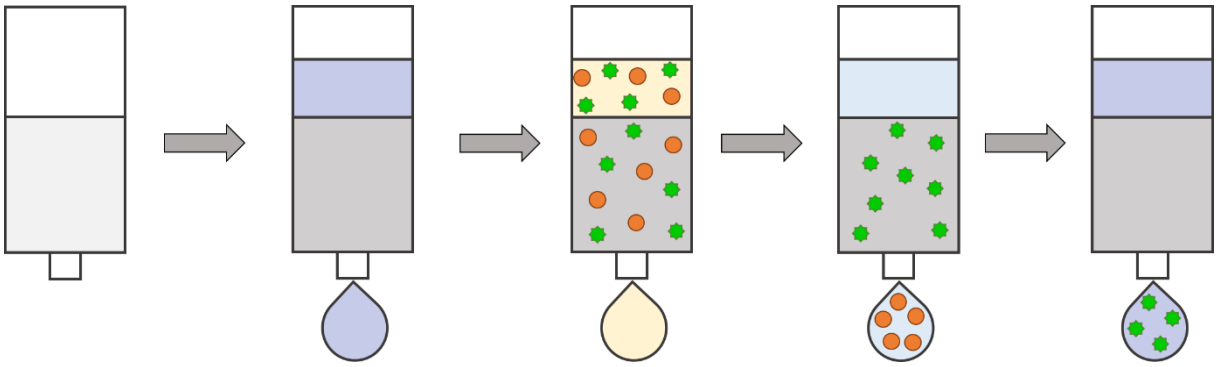

Figure II.6 - Operating principle of solid phase extraction. The solid phase is conditioned, followed by sample loading. Interferents are washed usually using water, and compounds of interest are eluted using an elution solvent.

3. Data processing methods for non-targeted exposomics

This section presents how the data processing of a HRMS chemical fingerprint is used to translate this acquired data to a list of features. As this list is used as a basis for the annotation and/or suspect screening steps, it is critical to ensure that the data processing steps taken allow the proper recovery of all the detected signals, including the low-abundant ones. Data processing is a crucial step as poor parameter optimization may result in the propagation of errors on the subsequent workflow steps. Moreover, it is a complex step that involves many substeps, each achievable through various algorithms that are not all implemented in the chosen data processing software. Therefore, its optimization for the intended application is critical, especially in the case of an interest in low-abundant compounds in complex matrices, where data quality may be limited due to sensitivity issues. Like for sample preparation, a thorough evaluation and optimization of the data processing step for the detection of low-abundant compounds in complex matrices is presented in Chapter IV. In the next paragraphs, the used data processing tools along with the four major steps and algorithms implemented successively during this work's non-targeted data processing are presented: peak detection, alignment, gap filling and normalization.

3.1. Data processing tools

Five data processing tools were used in the frame of this PhD work: MarkerView, Progenesis QI for Metabolomics, MZmine2, XCMS, and MS-DIAL 4.0. The first four tools were optimized and compared; detailed results are available in Chapter IV.

MarkerView and Progenesis QI are vendor software provided by AB SCIEX and Waters, respectively. MZmine2[17] and MS-DIAL[18] are open source solutions with graphical user interfaces, and XCMS[19] is an open source R-based package. While vendor software are usually more user-friendly compared to open source software, they often operate in a black box-like fashion, with little to no information on the algorithms and parameters used to process the data. The following paragraphs detail the different algorithms available for each major data processing steps in open source software, as this information is not available for vendor software.

3.2. Peak picking

The first data processing substep is peak picking, during which features are detected in each individual sample.

First, MS spectra are individually centroided (i.e. represented by a single value, often the mass peak apex[20]). Different algorithms are available depending on the chosen data processing software, such as *centWave* and *Wavelet transform* algorithms in XCMS and MZmine2 respectively or *ADAP* in MZmine2. The first two cited algorithms are continuous wavelet transform (CWT) algorithms based on matching m/z peaks to a “Mexican hat” or “Ricker” wavelet model[17]. These algorithms have been reported as particularly well-suited for noisy data[17]. Automated Data Analysis Pipeline (ADAP), on the other hand, is a complete data processing pipeline as underlined in the name. Although there is little available information on its specific mechanisms, the peak detection module within this pipeline is described to be particularly efficient in reducing false positive peak detection compared with CWT algorithms[21]. Several parameters used to perform this step critically affect the data processing results, in particular the peak width (usually required as a minimum value or as a range) and the noise threshold.

Then, close-to-identical m/z values observed over consecutive scans are combined into chromatogram objects. These objects might be either a single peak, or a group of peaks with similar m/z and Rt. They therefore need to be deconvoluted into individual peaks. Several algorithms may also be used for deconvolution, relying on finding local minima or using the chromatogram curve’s second derivative (i.e. Savitzky-Golay algorithm) to establish

boundaries between peaks[17, 22]. MZmine2 offers both of these options, while it is still unclear on which strategy XCMS's *refineChromPeaks* function relies.

A schematized representation of the peak picking process is available in Figure II.7.

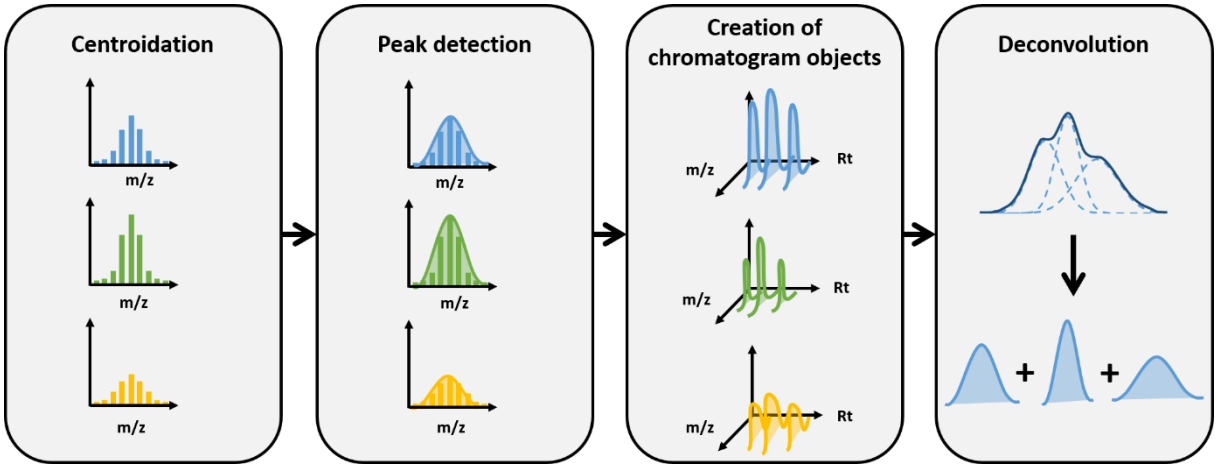

Figure II.7 – Representation of the peak picking process, which consists of four steps: centroidation, peak detection, creation of chromatogram objects and deconvolution.

The most critical parameters for this step are usually the minimum peak width (or range of peak widths) that should be expected by the software, and the noise threshold. In this PhD work, these parameters were tested with default values and optimized for the detection of low-abundant chemicals in complex biological matrices. Given the LC method used in the context of this PhD, the minimal and maximal peak width values were found to be 6-10 s and 50 s respectively, depending on the software. Minimizing the noise threshold (i.e. setting it at 0-10 depending on what is allowed by the software) was also found to provide the best results.

3.3. Alignment

Once individual peaks have been detected for each sample, an alignment must be performed to establish the common features among different samples. In this step, peaks with identical m/z and Rt (with a user-determined tolerance range) across samples are matched across the samples. A Rt correction can also be implemented at this stage; indeed, analytical drift on Rt is a frequent issue, and it is possible to adjust the data by shifting signals to align them between samples. MZmine2 offers the Join aligner and the RANSAC aligner. The first one only relies on the tolerance ranges specified by the user, with no additional adjustment. The second one (RANdom Sample Consensus, RANSAC) is an iterative algorithm which adjusts parameters from a mathematical model based on random observations, and checks the fit. It was determined that this algorithm provide a significantly better alignment performance than the Join aligner[17]. XCMS also offers two alignment algorithms called obiwarp and peakGroups.

Obiwarp relies on a center sample against which all other samples are aligned[23]. The peakGroups algorithm is based on peaks present in most or all samples. With this algorithm, the retention time deviation of peaks is established using a linear or a polynomial model. The obtained model is then extended to close peaks that are not present in all samples. This algorithm is presumably similar or identical to the one used in MS-DIAL, considering the requested parameters. However, in MS-DIAL, only specific user-determined features (often internal standards) are used to establish the linear or polynomial model.

These different alignment strategies were tested and evaluated for exposomics applications. In the case of our selected analytical system, the tolerance ranges chosen for m/z and Rt were of 10 ppm and 2 min respectively. These values were determined based on vendor recommendation (m/z tolerance are typically set lower with Orbitrap analyzers compared to QTOF analyzers for instance) and visual examination of the raw data. An in-depth detail of software parametrization is presented in Chapter IV.

3.4. Gap filling

Following the alignment, the obtained feature matrix might contain missing data (i.e. no peak detected for a m/z × Rt combination in one or more samples). This may either be due to an absence of signal, or to a failure of the peak detection algorithm during the first data processing substep. All data processing software can therefore proceed to the gap filling step, where raw signal in the m/z × Rt region of interest is extracted, integrated and added to the matrix. This is usually done by exploring the raw data, but may also be performed with data collected at the peak picking stage. For the work presented in this manuscript, this step was systematically performed. As there is no parametrization for this step, it did not require optimization.

3.5. Normalization

Normalization of the feature areas is the last critical substep of data processing. It is often required to perform statistical analysis or to report any semi-quantitative data, as there is a need for area comparability between samples. While XCMS does not support normalization at this time, both MS-DIAL and MZmine2 offer to normalize feature areas through a user-specified list of reference compounds that should have identical areas in all samples (often internal standards). MZmine2 also offers linear normalization, where all areas are divided by a normalization factor (e.g. average intensity or total raw signal). In the context of this PhD, normalization strategies based on total ion chromatogram were systematically attempted with software that allowed it (i.e. MarkerView an MZMine2) and compared to raw results to determine relevance. This is notably demonstrated in Chapter V.

In the context of this PhD work, a thorough comparison of data processing tools for this purpose was implemented. This work is presented in Chapter IV. This allowed demonstrating that adjustments still need to be made to these tools to be suited for exposomics applications, and that vendor software, while opaque, can be an efficient solution to non-targeted data processing.

4. Annotation methods and tools

At the end of the data processing step, a feature list each characterized by a m/z, a Rt and one area per sample is obtained. The last critical step of the non-targeted workflow is to link these features to chemical identities. This link may be formed in two ways: non-targeted screening, and suspect screening. Both have been used during this PhD, even though suspect screening was predominantly used. Non-targeted screening was notably used for the NORMAN Network's first collaborative trial in biota, as mentioned in the "Scientific valorization chapter, paragraph 4.

4.1. Non-targeted screening: statistical analysis

Non-targeted screening aims to assign a chemical identity to an experimental feature with no pre-existing idea regarding the compound's structure. This approach is highly challenging, as unequivocal structural elucidation of a compound requires advanced knowledge and means in many fields, such as mass spectrometry, nuclear magnetic resonance, organic chemistry, biochemistry, and bioinformatics. However, it is also very promising as a mean of expanding knowledge regarding the chemical exposome by uncovering entirely uninvestigated compounds.

In this work, univariate analyses were performed under an R environment (version 3.6.3). Individual features were compared between samples by performing unpaired *t*-tests and computing *p*-values with an Adaptive Benjamini-Hochberg (ABH)[24] correction for multiple comparisons. Features presenting lowest adjusted *p*-values (i.e. < 0.01) were prioritized for the annotation process. Multivariate analyses were also performed to compare sample groups and establish whether there was an observable and explainable discrimination between groups. To this end, unsupervised Principal Component Analysis (PCA) and Partial Least Square-Discriminant Analysis (PLS-DA) were implemented under an R environment[25].

4.2. Suspect screening tool

As mentioned in Chapter I, suspect screening is an approach that consists in linking experimental features to compounds that are suspected of being present in the sample *a posteriori*. The establishment of this link is a time-consuming task that has the potential to be

at least partly automatized. An important part of this PhD was devoted to develop a fully automated software using several chemical descriptors and developing intermediate and global confidence scoring.

The developed suspect screening software is a Python software tool first developed in 2019 in LERES to assist suspect screening approaches using MS1 analyses. It aims to perform an automatized pre-annotation of processed datasets obtained from liquid LC coupled to HRMS analyses. To this end, confidence indices (CI) were constructed to score the proximity between experimental features and suspects. This proximity is established through three chemical predictors, each scored individually: the classically used m/z, isotopic fit (which combines m/z and relative abundance fit) and Rt. Pre-annotated features need further manual curation based on fragmentation patterns found in either MS1 or MS2 acquisitions, isotopic pattern (particularly in the case of the presence of a bromine and/or chlorine atoms), and plausibility.

4.2.1. Suspect screening predictors

Suspect screening approaches aim to link experimental features to a list of compounds post-analysis. Linking features to suspects can be done through various indicators, such as the often-used MS2 fragmentation pattern[18, 26]. In the following paragraph, the three predictors implemented in the in-house tool and their relevance for MS1 suspect screening are presented. They will be further developed in Chapter IV.

4.2.1.1. Mass-to-charge ratio

The mass-to-charge ratio (m/z) is the basis of all annotation and suspect screening approaches. Indeed, the precision of exact masses generated by HRMS analyses allow significantly restraining the number of chemical formulas that may be associated with a given signal. This predictor is therefore fundamental to implement a suspect screening approach.

4.2.1.2. Isotopic fit

Another parameter that can be used to elucidate a compound's chemical formula is its isotopic pattern. Indeed, the presence of certain atoms such as bromine, chlorine, or sulfur in a molecule is reflected in the compound's isotopic pattern due to the $^{81}Br/^{79}Br$, $^{37}Cl/^{35}Cl$, and $^{34}S/^{32}S$ ratio values of approximately 1.00, 0.32 and 0.05 respectively. This information is particularly relevant in the case of some compounds classes such as pesticides, which often include one or more bromine or chlorine atoms.

4.2.1.3. Retention time

While m/z and isotopic pattern can give a reliable indication of a compound's chemical formula, other compound characteristics can be explored. Retention time (Rt) is an indication of a compound's affinity to the column's stationary phase compared to the mobile phase; in the case of many LC-HRMS systems, this translates to the compound's polarity (even though caution must be taken to not overgeneralize). This parameter is represented by a logP value, which can allow a distinction between two compounds having an identical chemical formula. Despite the potential of such a predictor, the retention time is not often implemented in annotation or suspect screening software currently available, except with a user-specified library containing experimental retention times[17, 18]. While the experimental retention time is the ultimate parameter to reach a level 1 annotation according to Schymanski *et al.*[27], it requires the use of a standard injected on the same system. Yet, acquiring standards for a large number of compounds is not feasible due to limitations in terms of both financial resources and commercial availability. Thus, Rt values may also be predicted through various algorithms such as RTI[28], Retip[29], or classically-used linear regressions using logP values[30]. Although these predicted values are less reliable than experimental values, they can help prioritize the most likely annotation of a feature and drastically reduce false positives. To date, no major screening tool implements the use of predicted Rt values to assist this process, which is why it was implemented in the in-house software.

These suspect screening predictors are used to score the similarity between features and suspects. Individual scores are then combined to a global confidence index that indicates the overall similarity between the feature and suspect. The schematized operating principle is presented in Figure II.8. To operate, the in-house software is structured in two main complementary modules: a library that regroups all the suspect compounds' theoretical properties, and a suspect screening module that matches experimental features to said suspects. The next paragraphs detail the last updated version of the in-house software (version 2.0). A presentation of its first version is available in Chapter IV paragraphs 4.4 and 4.5.

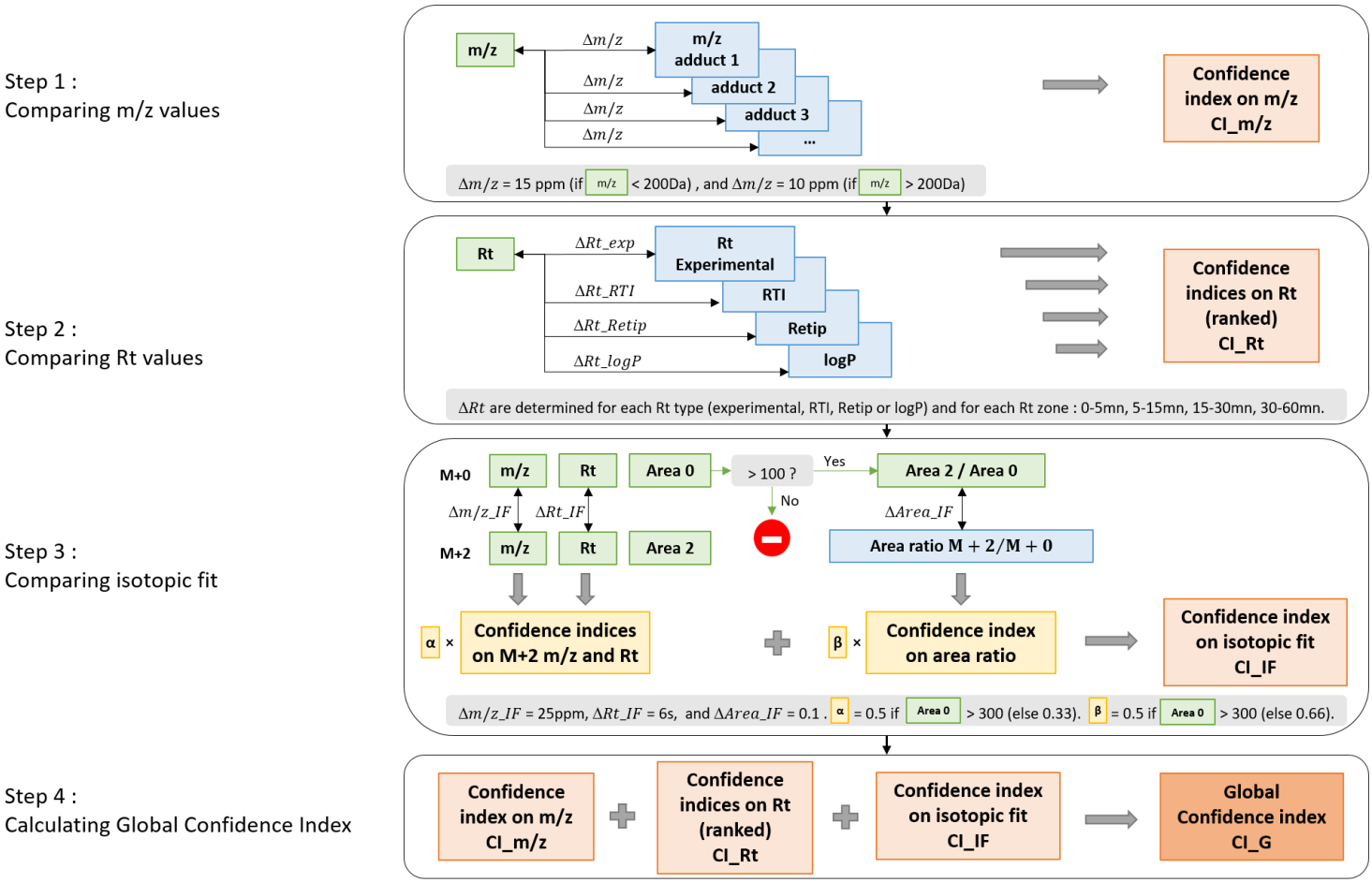

Figure II.8 – Schematized operating principle of the in-house annotation workflow in four steps: comparing successively m/z, Rt and isotopic fit, then generating a global scoring.

4.2.2. Library module: generating suspects data

The library module allows indexing and computing the reference data for the list of suspects. Every compound listed in the library must be linked to a chemical formula, a unique identifier such as the SMILES, and if available, Rt values (experimental or predicted) and an logP value. The library then outputs data regarding the compound's m/z, theoretical isotopic pattern, and Rt. Indeed, the chemical formula allows calculating nine exact masses: the monoisotopic mass, the masses of four positively charged adducts ($[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$), and the masses of four negatively charged adducts ($[M-H]^-$, $[M-H_2O-H]^-$, $[M+Cl]^-$, $[M+FA-H]^-$). Moreover, the formula allows the computing of theoretical isotopologue probabilities P_0, P_1, and P_2 (i.e. first, second, and third isotopologue) as well as their masses M_0, M_1 and M_2 through a polynomial-based algorithm adapted from the MIDAs software[31]. Four parameters are computed and presented to the user: mass differences M_1-M_0 and M_2-M_0, as well as probability ratios P_1/P_0 and P_2/P_0. Lastly, the logP value given by the user can be used to predict a Rt

value under the condition that the library contains at least 20 compounds that have both an experimental Rt and a logP value indicated.

The library used for this work contains close to 6000 compounds, which were compiled from various sources:

(i) Xenobiotics previously reported as detected in blood plasma or serum in the literature[32-35];

(ii) Compounds reported in open access databases Human Metabolome Database[36], Exposome Explorer[37], Foodb[38], and the Normal Suspect List Exchange[39]

The compounds listed in the library can be modified depending on the research question. Once all the predictors' data is calculated, the suspect list can be compared to the experimental features in the suspect screening module through the computing of confidence indices.

4.2.3. Suspect screening module: computing confidence indices

The suspect screening module requires a feature list obtained from any data processing tool, containing the following columns: m/z, Rt, and areas for all analyzed samples. Each feature is compared to compounds from the suspect list through confidence indices computed on the three predictors presented in paragraph 4.2.1. CI values are computed according to Equation II.1.

$$\mathrm{CI_i} = 1 - \frac{\left| \frac{\mathrm{i_{feature}} - \mathrm{i_{suspect}}}{\mathrm{i_{suspect}}} \right|}{\Delta_\mathrm{i}}$$

Equation II.1 - Expression of Confidence Indices (CI) for all predictors (i= m/z, Rt, or A_n/A_0 ratio, where A_n refers to the area of the nth isotopologue). Δ_i is a confidence interval and is specifically defined for each predictor as the maximal acceptable deviation from the reference value.

4.2.3.1. Mass-to-charge ratio

Depending on the ESI mode specified by the user, the software compares the feature's m/z with one of the two sets of adducts generated by the library. This predictor acts as a filter, as suspects with m/z values outside the confidence interval are eliminated as potential annotations. The confidence interval $\Delta_{m/z}$ is based on instrumental precision. It takes the value of 15 or 10 ppm depending if the m/z is strictly lower than 200 Da or over 200 Da respectively.

4.2.3.2. Isotopic fit

The matching between a feature and a suspect's isotopic fit is evaluated in a stepwise manner. At first, the software determines which isotopologue should be investigated. As mentioned

earlier, compounds containing chlorine, bromine, or sulfur atoms present a distinctive isotopic pattern involving a high abundance of the second isotopologue. If one of these compounds is the considered suspect, as well as if a $[M+Cl]^-$ adduct is considered, the software will focus on the second isotopologue. The first isotopologue will be considered for all other compounds and adducts.

Then, the software will establish whether there is a feature in the dataset that may be the M+n (n=1 or 2 for first or second isotopologue) of the annotated signal. To do so, it will compare the mass differences between two features and the suspect's theoretical M_n-M_0 value computed by the library module. A first temporary CI is computed based on the m/z difference proximity between suspect and feature, with the same $\Delta_{m/z}$ values as the ones presented in paragraph 3.2.3.2. A second temporary CI is also computed based on the Rt proximity between the annotated feature and the M+n, with a strict Δ_{Rt} value of 6 seconds (0.1 min). Indeed, isotopologues should be detected exactly at the same time; the confidence interval is set to take the instrument's and the data processing tool's uncertainties into account. The two temporary CI are averaged to obtain a first intermediate CI, referred to as "M+n identification CI".

Once the M+n feature is identified, area ratios are compared under the condition that the area of the M+0 (i.e. the annotated feature) is superior to 100. This is because low areas are often poorly integrated, resulting in inaccurate ratio values. If this is not verified, only the intermediate M+n identification CI is displayed. Else, the area ratios are compared and a second intermediate CI is computed for abundances with a $\Delta_{A2/A0}$ value of 0.1. The determination of the confidence interval is based on a regression of experimental area ratio values against theoretical area ratio values for 98 compounds. The root mean square error (RMSE) was calculated and the confidence interval was established at 3 RMSE to encompass 99.7% of projected data points (assuming normal distribution and applying statistics' empirical rule). A detailed explanation is presented in Chapter IV, paragraphs 4.4.3 and 5.2.2.

Lastly, an overall CI for isotopic fit is computed as a weighed sum of the two intermediate CI for M+n identification and abundance. For the reason cited earlier regarding limited confidence in the integration of small areas, the ponderation is determined based on the area of the M+n. Indeed, the two CI are weighed identically if the M+n area is higher than 20, else the M+n identification CI is weighed at 1/3 and the abundance CI is weighed at 2/3.

4.2.3.3. Retention time

As previously mentioned, the in-house software supports up to four Rt value per compound in the library: one experimental Rt, an up to three predicted Rt. All CI for Rt are computed using the standard formula presented in Equation II.1 with the appropriate Δ_{Rt} values.

In the software's initial version, the Δ_{Rt} value for experimental Rt was determined manually based on analytical Rt variability. Detailed explanations regarding these Rt prediction models are available Chapter IV, paragraphs 4.4.2 and 5.2.1. Briefly, compounds from the optimization mix spiked in plasma and serum samples (n=8) as well as isotopically labeled compounds (listed in Appendix 1.1) spiked in 16 plasma and serum samples were used to determine the standard deviation (SD) on Rt values. The chromatogram was divided in four sections based on observable variability as analytical variability in Rt is heterogeneous. The Δ_{Rt} value was constructed by selecting the highest compound Rt SD for each section, to avoid excessive stringency, and multiplying by three. In the software's current version, the suspect screening module is able to automatically compute Δ_{Rt} values based on a user-filled Excel sheet containing triplicate Rt data for at least 20 known compounds. The user may also specify their desired way of sectioning the chromatogram, or leave the standard sectioning of the chromatogram in quarters.

Regarding the three predicted Rt, for this work, they were obtained through an in-house regression model based on logP, the quantitative structure-retention relationship-based tool RTI[28], and machine learning-based tool Retip[29]. These three prediction models were evaluated and compared based on a set of 134 compounds presented in Appendix 1.3, which allowed ranking them from most reliable (RTI) to less reliable (logP). Detailed explanations regarding these Rt prediction models are also available Chapter IV, paragraphs 4.4.2 and 5.2.1. Briefly, the Δ_{Rt} values for all predicted retention times were established manually by comparing experimental Rt and predicted Rt when both values were available. Absolute differences between these two values were calculated and the standard deviation of each model's prediction within each predetermined chromatogram section was established. These values were multiplied by three to obtain the Δ_{Rt} values, each specific to a model and a chromatogram section.

The Rt CI was computed for all available Rt values for a given suspect, whether experimental or predicted. However, the global CI combining all predictors was computed using only the CI associated to the most reliable Rt available (i.e. experimental, then RTI, then Retip, then logP).

The in-house software was designed to compute CI on three chemical predictors to establish the similarity between features and suspects. The global CI is then computed as an average

of all the available CI values. Thus, each suggested annotation generated by the software is scored between 0 and 1 on all the mentioned predictors, as well as overall. The global CI is also preceded by a “G3”, “G2” or “G1” mention, which accounts for the number of predictors taken into account in its computing.

In the context of exposomics applications, where compounds of interest are often low-abundant in complex matrices and therefore often do not trigger MS2 acquisition, a tool such as this software which relies on MS1 predictors is a valuable help in assisting pre-annotation. Indeed, while manual curation is still required to confirm or infirm the suggested annotations, its discriminating scoring system allows prioritizing plausible annotations by drastically reducing false positives.

4.2.4. Manual curation

The in-house software was created to assist suspect screening approaches that provides pre-annotations. These suggested chemical identities must then be manually curated to rule out false positives (i.e. incorrectly identified chemical). This manual curation process comprises four main steps. Firstly, the extraction blank is manually checked to ensure that the compound’s presence is not linked to contamination during the sample preparation process. If the compound is present in the blank, the blank area is subtracted from the area in the samples. Secondly, the feature’s isotopic pattern is verified to ensure coherence with the suggested chemical formula (i.e. verification of whether the investigated m/z is a pseudomolecular ion, and of the isotopic ratios in case it was not performed by the software). Thirdly, the suspect’s fragmentation pattern should be compared to a reference spectra, which can be obtained through online databases[40], or through in silico fragmentation models[41, 42]. This pattern is used to partly or entirely confirm molecular structure (e.g. positional isomers or diastereoisomers may not be distinguishable). Other parameters such as polarity (via logP-predicted retention time) may help narrow the suggested annotation. Lastly, the plausibility is verified through a database search of the suggested formula, and a comparison of the pre-annotation with other possible close structural matches. For instance, if there is a strong structural resemblance between a well-documented endogenous compound and an exogenous compound never documented in blood, plausibility would dictate to rule in favor of the former.

5. Biological samples

The optimized workflow was then applied in a large-scale application. Initially, this application was to be made using blood plasma samples obtained from a Danish mother-child cohort dating back to 1988-89. More specifically, 256 blood plasma samples from pregnant women

linked with their daughter's clinical data 20 years later were selected through a collaboration with the Rigshospitalet (Copenhagen, DK) with David Kristensen. Reproductive health data for the daughters was also collected and available. This cohort would have therefore allowed linking data for environmental exposures during the prenatal period and reproductive health. However, due to the unforeseen pandemic circumstances and unresolved ethical procedures on the epidemiological side, samples from the local Breton Pélagie cohort were used. This cohort, initially built as a longitudinal study to measure exposure to organic pollutants during the pregnancy, included 3,500 women pregnant between 2002 and 2005 in Brittany. One of the follow-ups occurred when the children turned 12, at which time a questionnaire was provided to obtain physical growth data and pubertal stage. Additional clinical parameters such as growth, adiposity, visual function and oral-dental health were evaluated on a subset of 500 12-year-olds. Serum samples were collected from 250 12-year-olds at this time to measure sex hormones and to assess exposure to organic contaminants. Serum samples from 125 boys were used in this PhD work to perform a suspect screening approach to characterize the human internal chemical exposome.

References

1. Geller S., Lieberman H.*, et al.*, A systematic approach to development of analytical scale and microflow-based liquid chromatography coupled to mass spectrometry metabolomics methods to support drug discovery and development. *J Chromatogr A* **2021,** *1642*, 462047.
2. Panagopoulos Abrahamsson D., Wang A.*, et al.*, A Comprehensive Non-targeted Analysis Study of the Prenatal Exposome. *Environ Sci Technol* **2021**.
3. Hou Y., He D.*, et al.*, An improved detection and identification strategy for untargeted metabolomics based on UPLC-MS. *J Pharm Biomed Anal* **2020,** *191*, 113531.
4. Whiley L., Godzien J.*, et al.*, In-vial dual extraction for direct LC-MS analysis of plasma for comprehensive and highly reproducible metabolic fingerprinting. *Anal Chem* **2012,** *84* (14), 5992-9.
5. Sandra K., Pereira Ados S.*, et al.*, Comprehensive blood plasma lipidomics by liquid chromatography/quadrupole time-of-flight mass spectrometry. *J Chromatogr A* **2010,** *1217* (25), 4087-99.
6. Chetwynd A.J.,David A., A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. *Talanta* **2018,** *182*, 380-90.
7. David A., Chaker J.*, et al.*, Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **2021,** *156*, 106630.
8. Lindahl A., Saaf S.*, et al.*, Tuning Metabolome Coverage in Reversed Phase LC-MS Metabolomics of MeOH Extracted Samples Using the Reconstitution Solvent Composition. *Anal Chem* **2017,** *89* (14), 7356-64.
9. Sitnikov D.G., Monnin C.S.*, et al.*, Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. *Sci Rep* **2016,** *6*, 38885.
10. Rico E., Gonzalez O.*, et al.*, Evaluation of human plasma sample preparation protocols for untargeted metabolic profiles analyzed by UHPLC-ESI-TOF-MS. *Anal Bioanal Chem* **2014,** *406* (29), 7641-52.
11. Tulipani S., Llorach R.*, et al.*, Comparative analysis of sample preparation methods to handle the complexity of the blood fluid metabolome: when less is more. *Anal Chem* **2013,** *85* (1), 341-8.
12. Yang Y., Cruickshank C.*, et al.*, New sample preparation approach for mass spectrometry-based profiling of plasma results in improved coverage of metabolome. *J Chromatogr A* **2013,** *1300*, 217-26.
13. Want E.J., O'Maille G.*, et al.*, Solvent-Dependent Metabolite Distribution, Clustering, and Protein Extraction for Serum Profiling with Mass Spectrometry. **2006**.
14. David A., Lange A.*, et al.*, Disruption of the Prostaglandin Metabolome and Characterization of the Pharmaceutical Exposome in Fish Exposed to Wastewater Treatment Works Effluent As Revealed by Nanoflow-Nanospray Mass Spectrometry-Based Metabolomics. *Environ Sci Technol* **2017,** *51* (1), 616-24.
15. Tulipani S., Mora-Cubillos X.*, et al.*, New and vintage solutions to enhance the plasma metabolome coverage by LC-ESI-MS untargeted metabolomics: the not-so-simple process of method performance evaluation. *Anal Chem* **2015,** *87* (5), 2639-47.
16. Ahmad S., Kalra H.*, et al.*, HybridSPE: A novel technique to reduce phospholipid-based matrix effect in LC-ESI-MS Bioanalysis. *J Pharm Bioallied Sci* **2012,** *4* (4), 267-75.
17. Pluskal T., Castillo S.*, et al.*, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**.
18. Tsugawa H., Cajka T.*, et al.*, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.
19. Smith C.A., Want E.J.*, et al.*, XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006,** (78), 779-87.

20. Samanipour S., Choi P.*, et al.*, From Centroided to Profile Mode: Machine Learning for Prediction of Peak Width in HRMS Data. *Anal Chem* **2021,** *93* (49), 16562-70.
21. Du X., Smirnov A.*, et al.*, Metabolomics Data Preprocessing Using ADAP and MZmine 2. *Methods Mol Biol* **2020,** *2104*, 25-48.
22. Boulet J.C., Meudec E.*, et al.*, High-resolution mass spectrometry (HRMS): Focus on the m/z values estimated by the Savitzky-Golay first derivative. *Rapid Commun Mass Spectrom* **2021,** *35* (6), e9036.
23. Prince J.T.,Marcotte E.M., Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal Chem* **2006,** *78*, 6140-52.
24. Li A.,Barber R.F., Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Statist. Soc. B* **2019,** *81*, 45–74.
25. Liland K.H., Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry* **2011,** *30* (6), 827-41.
26. Lawson T.N., Weber R.J.*, et al.*, msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Anal Chem* **2017,** *89* (4), 2432-9.
27. Schymanski E.L., Jeon J.*, et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.
28. Aalizadeh R., Thomaidis N.S.*, et al.*, Quantitative Structure-Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. *J Chem Inf Model* **2016,** *56* (7), 1384-98.
29. Bonini P., Kind T.*, et al.*, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020,** *92* (11), 7515-22.
30. McEachran A.D., Mansouri K.*, et al.*, A comparison of three liquid chromatography (LC) retention time prediction models. *Talanta* **2018,** *182*, 371-9.
31. Alves G., Ogurtsov A.Y.*, et al.*, Molecular Isotopic Distribution Analysis (MIDAs) with adjustable mass accuracy. *J Am Soc Mass Spectrom* **2014,** *25* (1), 57-70.
32. Rappaport S.M., Barupal D.K.*, et al.*, The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* **2014,** *122* (8), 769-74.
33. Schulz M., Iwersen-Bergmann S.*, et al.*, Therapeutic and toxic blood concentrations of nearly 1,000 drugs and other xenobiotics. *Crit Care* **2012,** *16* (4), R136.
34. Assens M., Frederiksen H.*, et al.*, Variations in repeated serum concentrations of UV filters, phthalates, phenols and parabens during pregnancy. *Environ Int* **2019,** *123*, 318-24.
35. Haug L.S., Sakhi A.K.*, et al.*, In-utero and childhood chemical exposome in six European mother-child cohorts. *Environ Int* **2018,** *121* (Pt 1), 751-63.
36. Wishart D.S., Feunang Y.D.*, et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **2018,** *46* (D1), D608-D17.
37. Neveu V., Moussy A.*, et al.*, Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res* **2017,** *45* (D1), D979-D84.
38. Foodb. www.foodb.ca.
39. Aalizadeh R., Alygizakis N.*, et al.*, Merged NORMAN Suspect List: SusDat. v0.4.1 ed.; 2022.
40. Horai H., Arita M.*, et al.*, MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **2010,** *45* (7), 703-14.
41. Ruttkies C., Schymanski E.L.*, et al.*, MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **2016,** *8*, 3.
42. Allen F., Pon A.*, et al.*, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* **2014,** *42* (Web Server issue), W94-9.

Chapter III. Systematic evaluations of blood-derived sample preparation methods for HRMS-based chemical exposomics

1. Context and summary

This chapter was published as an original paper as first author in the journal *Analytical Chemistry*: **Chaker, J.**, Kristensen, D. M., Halldorsson, T. I., Olsen, S.F., Monfort, C., Chevrier, C., Jégou, B., David, A.* (2022). Comprehensive Evaluation of Blood Plasma and Serum Sample Preparations for HRMS-Based Chemical Exposomics: Overlaps and Specificities. *Anal Chem* **(IF=6.8)**, 94(2), 866–874.

The non-targeted characterization of biological samples strongly depends on the methodological choices made throughout the workflow. As the first critical step in the workflow, sample preparation must be diligently chosen and optimized. Indeed, this choice is highly decisive, as the compounds lost to sample preparation step cannot be recovered through any optimization of the following steps in the workflow. Since new data processing and annotation tools are continuously developed, it is crucial to obtain optimized HRMS fingerprints of often precious samples, that may be reprocessed to broaden the knowledge of the chemical exposome. When choosing and optimizing the SPM, the right middle ground has to be found between the sensitivity required to detect often low-abundant exogenous chemical compounds and the selectivity needed to eliminate highly abundant endogenous compounds responsible for ion suppression. As described in Chapter II paragraph 2, there are many categories of SPM available to prepare plasma or serum samples, with varying degrees of selectivity. The objectives of this chapter were to systematically evaluate the performance of twelve SPM to detect low-abundant compounds in complex biological matrices, and to document their effect on the visible chemical space.

In the following article, twelve SPM (seven PLR methods, three SPE methods, one SLE method and one PPT method) were systematically evaluated for the characterization of the chemical exposome through blood plasma and serum samples. This evaluation was performed based on the implementation of complementary criteria rarely used to evaluate non-targeted methods, namely quantitative (e.g. recovery, repeatability, matrix effect) systematically used in targeted approaches, and qualitative (e.g. time and ease of implementation) criteria. This evaluation process allowed documenting the observable analytical perimeter of the chemical exposome profiled with each of these SPM. Delineating the observable analytical perimeter of each SPM is crucial for further interpretation of HRMS datasets.

The SPM were evaluated using a stepwise approach. Firstly, the 50-compound set (optimization mix described in Chapter II, paragraph 1.1) was spiked at a mid-range level (i.e. 40 ng/mL) in serum samples, and recovery, repeatability and matrix effect were determined. Secondly, SPM suitable for this application were applied to serum and plasma samples spiked

with the same 50-compound set at a lower level (i.e. 10 ng/mL). Detection frequency, S/N, repeatability, spiking significance (i.e. significance of the difference in areas between spiked and non-spiked samples) and ease of implementation were evaluated, resulting on the further selection of two appropriate SPM. Lastly, those SPM were applied to cohort plasma and serum samples. Annotated compounds' areas were compared for the same samples prepared with one SPM or the other to assess the impact of the SPM choice on the visible chemical space. Results of these comparisons are described and discussed throughout this article. A simple sample preparation workflow involving both SPM was proposed to broaden the visible chemical space as they appear complementary.

Comprehensive evaluation of blood plasma and serum sample preparations for HRMS-based chemical exposomics: overlaps and specificities

Jade Chaker[a], David M. Kristensen[ab], Thorhallur Ingi Halldorsson[cd], Sjurdur Frodi Olsen[ce], Christine Monfort[a], Cécile Chevrier[a], Bernard Jégou[a†], Arthur David[a*]

[a] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35000 Rennes, France

[b] Department of Neurology, Danish Headache Center, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

[c] Center for Fetal Programming, Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark.

[d] The Unit for Nutrition Research, Faculty of Food Science and Nutrition, School of Health Sciences, University of Iceland, Reykjavik, Iceland.

[e] Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, United States of America.

* Corresponding author

To whom correspondence should be addressed:

Tel: +33 299022885

email: arthur.david@ehesp.fr

2. Abstract

Sample preparation of complex biological samples can have a substantial impact on the coverage of small molecules detectable using liquid chromatography-high-resolution mass spectrometry (LC-HRMS). This initial step is particularly critical for the detection of externally-derived chemicals and their metabolites (internal chemical exposome) generally present at trace levels. Hence, our objective was to investigate how blood sample preparation methods affect the detection of low-abundant chemicals and to propose alternative methods to improve the coverage of the human internal chemical exposome. We performed a comprehensive evaluation of twelve sample preparation methods (SPM) using phospholipid and protein removal plates (PLR), solid phase extraction plates (SPE), supported liquid extraction cartridge (SLE), and conventionally used protein precipitation (PPT). We implemented new quantitative and qualitative criteria for non-targeted analyses (detection frequency, recoveries, repeatability, matrix effect, low-level spiking significance, method detection limits, throughput and ease of use) to amply characterize these SPM in a step-by-step-type approach. As a final step, PPT and one PLR plate were applied to cohort plasma and serum samples injected in triplicate to monitor batch repeatability, and annotation was performed on the related datasets to compare the respective impacts of these SPM. We demonstrate that sample preparation significantly affects both the range of observable compounds and the level at which they can be observed (more than 40% of total feature only detected using one SPM). We propose to use PPT and PLR on the same samples by implementing a simple analytical workflow as their complementarity would allow the broadening of the visible chemical space.

Key words: Non-targeted exposomics, high-resolution mass spectrometry, sample preparation, plasma, serum

Graphical abstract

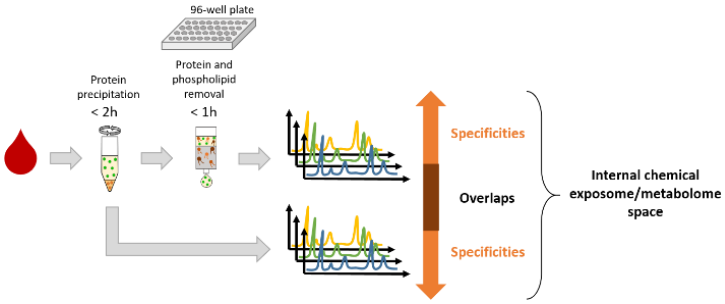

Figure III.1 – Graphical abstract of the research paper titled "Comprehensive evaluation of blood plasma and serum sample preparations for HRMS-based chemical exposomics: overlaps and specificities

3. Introduction

As the impact of environmental exposures and particularly chemical exposures to the global burden of chronic disease is uncovered[1, 2], the need for sensitive, robust and comprehensive detection of exogenous chemicals, their biotransformation products and their metabolites present as complex mixtures in human biological matrices grows. During the last few years, the technological progress regarding high-resolution mass spectrometry (HRMS) has allowed to simultaneously and reproducibly profile thousands of compounds (including both endogenous and exogenous chemicals) in biological samples using non-targeted approaches[3-6]. Concomitantly, significant developments and optimizations have been made on bioinformatics tools to improve their suitability to peak pick and annotate low-abundant chemicals in complex matrices, which are of particular interest for exposomics studies[7, 8]. However, optimizations are still lacking to ensure that the first analytical step of the workflow can profile unbiasedly the internal components of the human chemical exposome (i.e. exogenously derived chemicals accumulating in humans). A special focus on analytical methods allowing the detection of exogenous chemicals is necessary since concentrations of exogenous chemicals such as pesticides and plasticizers are generally 700 times lower than those of endogenous compounds in blood-derived samples[9, 10]. Considering the widespread use of liquid chromatography (LC) for compound separation coupled to HRMS, the presence of exogenous chemicals at trace levels in complex biological matrices (i.e. pg/ml) raises the question of sensitivity issues partially due to ion suppression[11]. Hence, a particular attention must be payed to the sample preparation step for exposomics applications to allow elimination of abundant interfering chemicals while ensuring minimal loss of compounds of interest. Furthermore, the determination of quantitative/qualitative parameters must be better defined to document the perimeter of the internal chemical exposome profiled with a given method[12-14].

The most commonly described sample preparation methods (SPM) for metabolomics applications of plasma or serum samples rely on solvent-based protein precipitation (PPT), and use cold methanol or acetonitrile with ratios of solvent-to-sample ratio between 1 and 4[11, 15-18]. For mid-range spiking concentrations (i.e. 800-5000 ng/mL), PPT was described as allowing high recovery rates[15], and producing more information-rich samples with a slight decrease in repeatability when using acetonitrile compared to methanol[11]. Overall, PPT is one of the least selective preparation methods. However, the presence of abundant compounds such as phospholipids in PPT extracted blood sample may be detrimental for the detection of low-abundant compounds[19] and/or method repeatability. Coupled with the need to extend column life and within batch analytical drifts, particularly in the case of high-throughput

applications, this has led to a growing interest in more selective SPM such as liquid-liquid extraction (LLE), phospholipid and proteins removal (PLR) methods, and solid phase extraction (SPE) methods[11, 15, 16, 19-22]. LLE offers sample decomplexification while maintaining good coverage among polar and non-polar compounds[23]. However, due to repeatability issues linked to emulsification and the need for high sample volume, supported liquid extraction (SLE) can be preferred to LLE for blood-derived sample preparation[24]. PLR and SPE allow further sample purification physically and chemically, as their packed-bed structure filter large precipitated proteins and aim to retain phospholipids[25]. When applied on samples with mid-range spiking concentrations, these SPM tend to perform better in terms of matrix effect than PPT[15], and have been described as complementary to PPT in terms of metabolome coverage[16].

Comparisons of SPM for plasma and serum samples to attain an optimal compromise between sensitivity and selectivity have been published, but have either relied on evaluating method performance at the non-targeted scale[16], or used only mid-concentration range spiking levels and endogenous spiking compounds (n < 20) [11, 15, 20] which is not suitable for exposomics applications. One study has however offered a performance evaluation for a SPE plate on exogenous compounds in lower concentrations[19]. To date and to the best of the authors' knowledge, there is no reported large-scale comparison of SPM for both blood plasma and serum oriented towards human chemical exposomics applications. Thus, the objective of this work is to evaluate twelve SPM for the chemical exposomics analysis of plasma and serum samples, with a focus on low-abundant compounds. Considering the complexity of human blood-derived samples in terms of number and concentration of chemicals, a large set of exogenous and endogenous spiking compounds (n=50) with a wide range of physical-chemical properties (0.07 ≤ logP ≤ 6.99 ; 133.0640 ≤ Monoisotopic mass (Da) ≤ 496.2607) was used to cover the chemical space[26]. Quantitative and qualitative criteria (i.e. respectively detection frequency, recoveries, repeatability, matrix effect, low-level spiking significance, method detection limits, and time of implementation, complementarity) were used to amply characterize these SPM in a step-by-step-type approach aiming to compare the reference PPT with alternative SPMs. The best-suited SPM were applied to cohort plasma (n=8) and serum (n=10) samples which were then injected in triplicate to monitor within batch repeatability, and annotation was performed on the related datasets to compare the respective impacts of these SPM on the obtained results at a larger scale.

4. Experimental section

4.1. Biological samples

Human blood plasma and serum bags used for method development were acquired from the French blood agency (Etablissement Français du Sang, EFS). For the final step of method validation, serum samples (n=10) were obtained from 12-year-old children from the PELAGIE cohort regrouping 3,421 women from Brittany (France) enrolled by gynecologists from the general population during early pregnancy between 2002 and 2006[27] and plasma samples (n=8) were obtained from a Danish mother-child cohort.

4.2. Sample preparation methods comparison

The ability of twelve SPMs to detect low-abundant chemicals in biological matrices were evaluated using a step-by-step comparison process. The methodology is presented in Figure III.2. First, a two-step procedure (including a SPM preselection step and then a comparison of preselected SPMs with the reference PPT) was conducted consecutively using sets of spiking experiments on homogenate plasma and serum samples. A mix of 50 spiking compounds was chosen to cover different chemical classes of contaminants (i.e. diet toxins, drugs, and pesticides) and metabolites (i.e. eicosanoids, neurotransmitters, and steroids). Labeled internal standards (IS) (n = 17, 100 ng/mL) were used throughout to monitor analytical variability attributed to UHPLC-ESI-QTOF injections (spiked post-extraction in the preselection phase) or sample preparation (spiked pre-extraction in the following phases). Suppliers and further physical-chemical data can be found in the Supporting Information (SI), Tables A1 and A2. The preparation methods selected through these two experiments were then applied to cohort serum (n=10) and plasma (n=8) samples and compared.

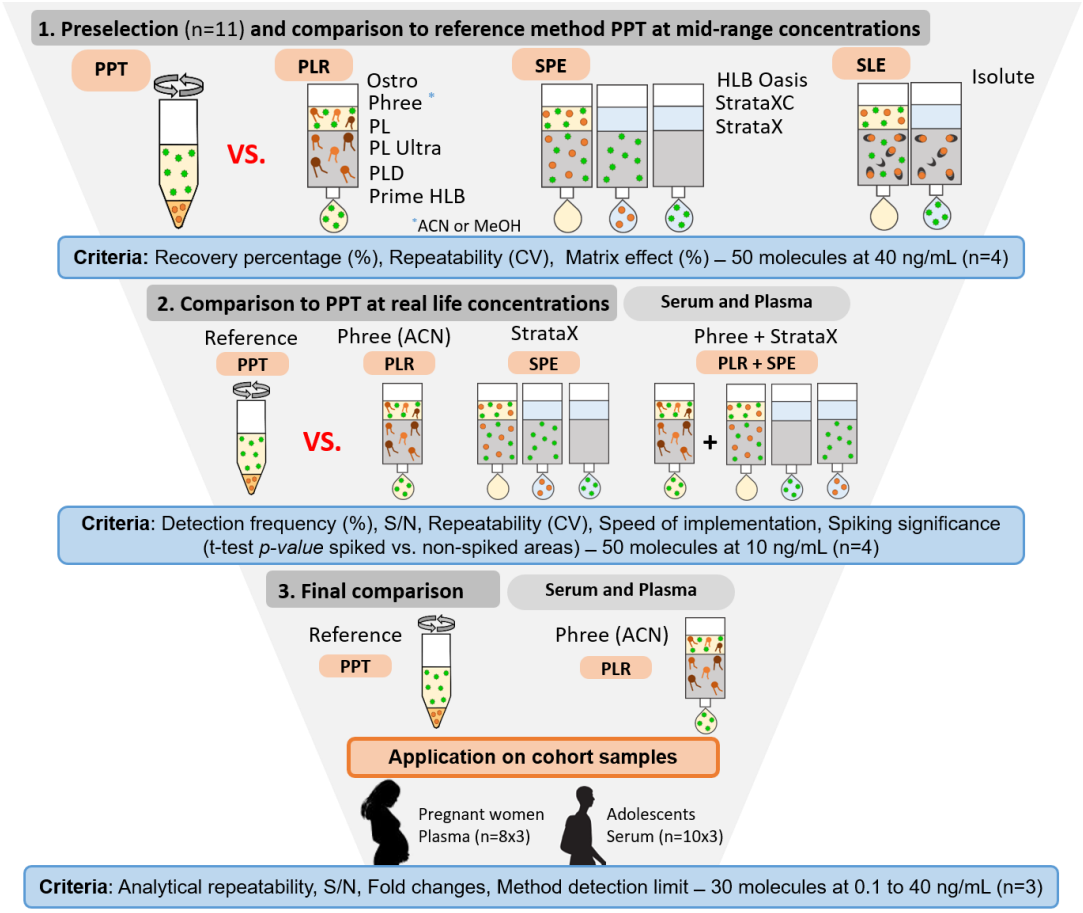

Figure III.2 – Diagram of the methodology used to compare sample preparation methods. Two low-level spiking experiments were conducted to compare various phospholipid and protein removal plates (PLR), solid phase extraction cartridges (SPE), and supported liquid extraction cartridge (SLE) among themselves, and to the classically used protein precipitation (PPT). The best-suited methods were selected using a set of qualitative and quantitative criteria, then applied to plasma and serum cohort samples to observe the impact of the sample preparation method on the visible chemical space.

4.2.1. Preselection

Seven procedures using phospholipid and protein removal (PLR) plates, three using solid phase extraction (SPE) plates, one using supported liquid extraction (SLE) cartridges, and conventionally used protein precipitation (PPT) (i.e. a total of twelve SPM) were implemented to prepare serum samples. Details on individual preparation procedures can be found in the SI. For each preparation method, homogenate serum samples (n=4) were spiked at 40 ng/mL using the 50-compound spiking set. Calibration samples (n=5, 20-150 ng/mL spiked after extraction) as well as an extracted matrix blank (n=1) and an extracted ultrapure water blank (n=1) were also prepared. Each batch was injected with calibration samples (n=5, 20-150 ng/mL) prepared in solvent. Absolute recovery percentage was calculated as the ratio of peak area of each compound in samples spiked before and after extraction. Repeatability was

assessed for each compound using the coefficient of variation (CV) of peak area on four replicates. Matrix effect (ME) was calculated as described in Equation III. for each compound at two concentration levels (lowest and highest points of calibration range).

$$ME[X,C]\ (\%) = \frac{A[X,C]_{solvent} - A[X,C]_{matrix}}{A[X,C]_{solvent}} * 100$$

Equation III.1– Matrix effect formula, where A is the peak area of a given compound X at a given concentration C.

SPM that were found adequate on all three criteria (i.e. recovery between 70-120%, repeatability below 20%, and low matrix effect) were preselected and further compared to the conventionally used solvent-based PPT.

4.2.2. Comparison to PPT at real-life concentrations

The preselected PLR plate (Phree – Acetonitrile (ACN)), the preselected SPE plate (StrataX), as well as a combination of these two preparation methods, were compared to PPT, which is a reference method for metabolomics[21, 22, 28]. For each of these four methods, plasma and serum homogenate samples (n=4 each) were spiked to a real life concentration (10 ng/mL) in plasma and serum. Background contamination was assessed using similar but non-spiked plasma and serum homogenates (n=4 each) and an extracted solvent blank (n=1). Detection frequency of compounds in spiked versus non-spiked samples and repeatability (using CV computations) were determined for each SPM. Signal-to-noise ratio (S/N) was retrieved for each compound and SPM. Spiking significance was assessed by computing *p*-values (unpaired *t*-tests) on compound IS-corrected areas in spiked versus non-spiked samples (threshold set at $p = 0.05$). Lastly, SPM were ranked on speed of implementation. Based on these criteria, two SPM were compared at the non-targeted scale on cohort samples.

4.2.3. Final comparison

The Phree PLR plate and PPT were used to prepare serum and plasma cohort samples (n=10 and 8, respectively). Batches included quality control (QC) samples and each sample was injected in triplicate. Analytical repeatability was assessed at the targeted scale using IS peak areas in QC and sample replicates, and at the non-targeted scale using the criteria proposed by Want et al.[28], according to which at least 80% of features found in at least 80% of QC should have a CV below 30%. Features varying significantly between the two SPM for each cohort were identified using *t*-tests (*p*-value threshold set at 0.01). These two data subsets were screened using an in-house automatized suspect screening tool[8] to characterize the impact of each SPM. Annotated features' S/N and fold changes (FC) between methods were also

reported. Details are available in Section 6. Further method characterization was achieved by determining the method detection limits (MDL). To this end, plasma and serum homogenate sample were spiked post-extraction at 0.1, 0.5, 1, 5, 10, 20, 40 ng/mL and were then injected in triplicate. MDL was determined as the lowest concentration with area CV lower than 10% and S/N higher than 100.

4.3. Data acquisition and quality assurance procedures

Samples were analyzed using a QTOF-MS (AB SCIEX X500R) interfaced with a UHPLC system (AB SCIEX ExionLC AD). Chromatographic separation was performed on injection volume of 2µL using an Acquity UHPLC HSS T3 C18 column (1.8µm, 1.0 × 150 mm) maintained at 40°C. Additional information regarding the chromatographic separation and (ESI) source parameters are available in the SI. Samples were analyzed with full scan experiments in both − and + ESI modes. MS/ MS fragmentation data were obtained by analysis of selected samples in sequential window acquisition of the theoretical mass spectrum (SWATH) or data dependent acquisition (DDA). Quality Control procedures are specified in the SI.

4.4. Data processing

4.4.1. Non-targeted data processing

Mass spectra acquired in full scan were processed using vendor software MarkerView v.1.3 (AB SCIEX). Main parameter values were set as: noise threshold of 10, minimal intensity of 20 counts, m/z tolerance of 10 ppm, retention time (Rt) tolerance of 2 min, minimum Rt of 1 min, no isotope filtering. This data processing workflow (i.e. software and parameters) was previously optimized and validated to detect low-abundant chemicals in blood plasma and serum samples[8]. Blank subtraction was performed by subtracting the solvent blank area from the sample's area for any given feature.

4.4.2. Targeted data processing

Manual peak integration for all spiked compounds and IS was achieved using vendor software Sciex OS v.1.6 (AB SCIEX).

4.5. Suspect screening and annotation

4.5.1. Suspect screening tool

Feature tables obtained through non-targeted data processing were screened using an in-house 6000-compound library mainly comprised of food intake biomarkers, pesticides (and metabolites), industrial pollutants, cosmetic ingredients, and pharmaceuticals/drugs (and

metabolites). An automatized in-house screening tool scoring proximity of m/z, Rt (experimental and predicted[29, 30]) and isotopic pattern between suspects and features was used[8]. Manual curation on MS/MS data was performed to confirm results obtained through the assisting suspect screening tool.

4.5.2. Annotation

Feature tables were uploaded into an R environment (version 3.6.3) to run univariate analyses. Statistical analyses were performed separately for each sample (i.e. individual), considering analytical replicates and two performed SPM. The impact of the SPM was assessed by performing unpaired *t*-tests and computing *p*-values with an Adaptive Benjamini-Hochberg (ABH) correction for multiple comparisons. Features presenting lowest adjusted *p*-values and a sample-to-blank area ratio of more than three for at least one sample were prioritized for the annotation process. Annotation was conducted manually, relying on chemical information databases[31, 32], experimental MS/MS databases[33], and in silico fragmentation prediction tools[34, 35]. Confidence levels based on recommendations made by Schymanski et al. (2014)[36] were provided in the SI, Tables A5a and A5b for serum and plasma samples respectively.

5. Results and discussion

5.1. Preselection of most suitable SPM

The twelve SPM performances regarding recovery, repeatability and matrix effect on 50 compounds spiked at 40 ng/mL in serum are presented in Figure III.3. Results for individual compounds are available in the SI, Table A3.

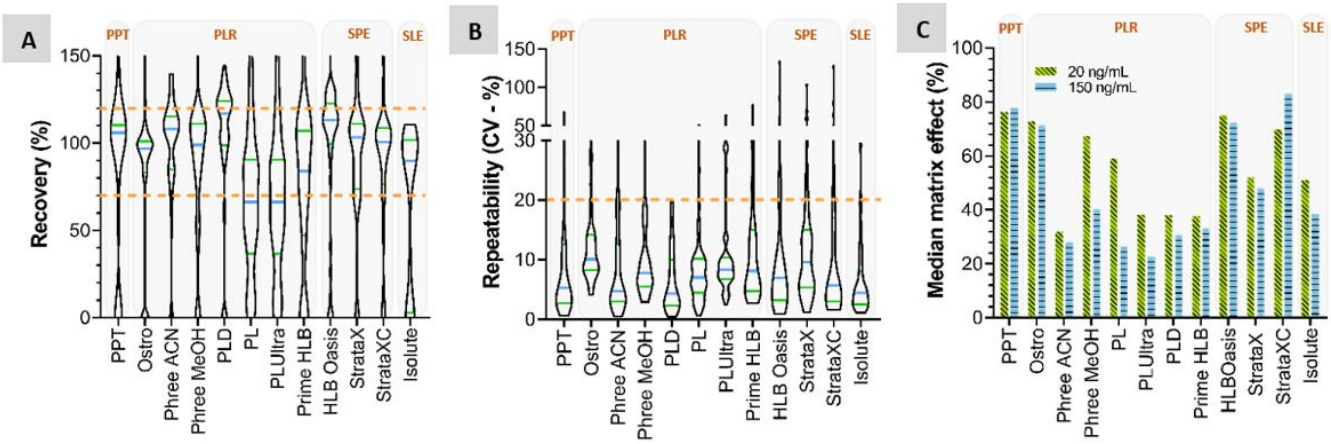

Figure III.3 – Comparison of the recovery (A), repeatability (B), and median matrix effect performances (C) of the eleven considered sample preparation methods using a 50-compound mix spiked in serum (n=4). Preparation methods include protein precipitation (PPT), phospholipid removal (PLR) plates, solid phase extraction (SPE) cartridges, and a supported liquid extraction (SLE) cartridge. For the recovery

and repeatability criteria, Q1, Q3 and median values are represented with two green lines and one blue line respectively. Median values for two spiking concentrations are presented for the matrix effect criterion.

Median spiked compound recovery varied between 56.3% (PLUltra) and 102.6% (PLD). PL and PLUltra are seemingly the least adequate SPM for the intended application, only allowing a median compound recovery of 61.7% and 56.3% respectively. SPM recovery performances for individual compounds indicated that PL and PLUltra specifically performed less adequately on polar compounds (0.07 ≤ logP ≤ 1.73). This may be explained by the fact that both of these plates retain phospholipids using a Lewis acid-base interaction between the stationary phase and the polar esterified phosphate group found in phospholipids[37]. However, due to lack of information on the phospholipid retention mechanism of other PLR plates, this hypothesis cannot be further investigated. The SLE cartridge did not seem adequate either for the intended application, as 20% of compounds were not recovered at all. Most of these non-recovered compounds (90%) were compounds usually favored in – ESI mode notably due to the presence of a common carboxylic acid group, which may suggest a less efficient desorption of such molecules when using this cartridge. Similarly, Prime HLB seemingly disadvantaged the recovery of compounds presenting a carboxylic acid group (100% of non-recovered compounds). This SPM also seemed inadequate for the recovery of selective serotonin reuptake inhibitors fluoxetine and paroxetine (8.8% and 1.5% recovery respectively), which may indicate a particular affinity of the sorbent for this class of compounds. It should be noted that eight compounds (i.e. 2-phenylphenol, acetylsalicylic acid, arachidonic acid, cotinine, nicotine, leukotriene D4, and prostaglandins D2 and J2) were generally poorly recovered (recovery below 70% for at least six SPM). As these compounds span across wide ranges of m/z (162.1167 ≤ Monoisotopic mass (Da) ≤ 496.2607) and Rt (3.76 ≤ Rt (min) ≤ 46.64), and share no common substructure, it appears that recovery in the case of low-level spiking in a complex matrix is partly compound-dependent with no evident generalization hypothesis. A similar observation regarding overall poor compound recovery regardless of the used extraction method was reported by Tulipani *et al.* (2015)[20].

Overall, five out of eleven methods (i.e. PLR plates Ostro, Phree with both solvents, StrataX and StrataXC) in addition to reference SPM PPT presented Q1 and Q3 recovery values comprised between 70% and 120%, constituting adequate performance for this criterion. Despite the generally satisfying recovery values obtained with these SPM, Ostro also tended to disadvantage compounds with a carboxylic acid group, although at a lesser level than Isolute or Prime HLB (14% of compounds were not recovered). Phree PLR plates mildly disadvantaged two thiophosphates, i.e. chlorpyrifos and diazinon (42.8-63.6% recovery),

regardless of the used solvent. Another thiophosphate, i.e. Malathion, was only recovered at 53.8% when using Phree with methanol. This insecticide, along with its precursor dimethyldithiophosphate, were also mildly to strongly disadvantaged by both Strata SPE cartridges (2.8-69.2% recovery). This tendency may indicate a need for a particular attention to thiophosphates when choosing and optimizing an SPM for non-targeted exposomics studies.

Observed repeatability on compound recovery was suitable for all SPM, with a calculated CV below 20% for 80% (HLB Oasis) to 100% (PLD) of spiked compounds. Lower interquartile ranges (i.e. difference between the third and first quartiles) were noted for PLR plates (3.4-9.2%) compared to SPE cartridges (9.1-13.1%). This suggests that PLR-based methods are more repeatable than SPE-based methods overall, which may be attributable to the higher complexity of SPE protocols (i.e. higher number of steps), as was previously suggested by Rico et al. (2014)[16].

Median matrix effects were highly variable among SPM, ranging from 31.9-75.0% (Phree ACN and PPT respectively) for the 20 ng/mL spiking level and from 22.6-83.0% (PLD and HLB Oasis respectively) for the 150 ng/mL spiking level. As expected, higher median matrix effect were observed with the lower spiking concentration for most SPM, with the exception of HLB Oasis (69.7-83.0% at 20 and 150 ng/mL). Additionally, PPT showed high matrix effect compared to other SPM, which was expected since it is the least selective. For PLR plates, Phree ACN performed best with a low median matrix effect at both spiking levels (31.9% and 28.0% at 20 and 150 ng/mL). It is to be noted that while Phree MeOH allowed similar performance on the recovery criterion, the use of methanol as a solvent exacerbated the observed matrix effect, in coherence with what was previously reported by Sitnikov *et al.* (2016)[15]. StrataX was the best-performing SPE cartridge at both spiking levels (52.0% and 47.9% at 20 and 150 ng/mL).

Overall, Phree ACN was the best compromise among PLR plates between high compound recovery, high repeatability and low matrix effect in the case of low-level spiking. Similarly, for SPE cartridges, StrataX was identified as the most appropriate given the considered criteria. Lastly, the SLE cartridge did not allow sufficient homogeneity in compound recovery to be selected for the next SPM comparison step.

5.2. Comparison to PPT at real-life concentrations

The preselected SPM Phree ACN and StrataX were compared to the commonly used solvent-based PPT on plasma and serum samples. Moreover, as relatively high matrix effects were observed namely for StrataX, a combination of both preselected SPM, further referred to as

Phree+StrataX, was carried out to attempt further purification of the samples. The SPM performances regarding spiked compound detection frequency, S/N, semi-quantification performance, detection significance, and speed of implementation were evaluated following a 10 ng/mL spiking of plasma and serum samples. Results are presented in Figure III.4. Results for individual compounds are available in the SI, Table A4.

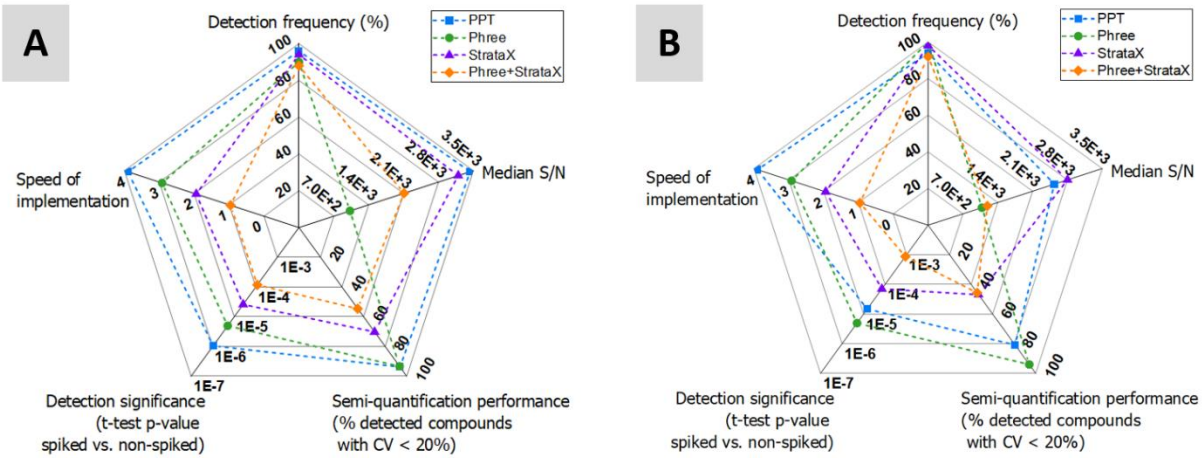

Figure III.4 – Sample preparation methods evaluation for the detection of 50 low-level spiked compounds in (A) serum and (B) plasma samples (n=4 each). Outer edges identify best performances.

Some differences were observed between matrices; indeed, median S/N values were lower for plasma for all SPM except Phree, and semi-quantification was poorer for this matrix when using PPT or StrataX. Observed areas are smaller in plasma samples overall (although not for all compounds), which could partly explain both the lower S/N values and area irregularities. This is consistent with prior reports of compound-dependent anticoagulant-caused ion suppression in plasma samples.[38]

All SPM allowed adequate spiked compounds detection frequencies in both matrices (88-96% of low-level spiked compounds detected in serum, 92-100% in plasma), although the combination of Phree ACN and StrataX systematically ranked last. Similarly, median S/N for spiked compounds were satisfying in all cases, ranging from 1024-3437 (Phree ACN-PPT respectively) in serum and 1082-2803 in plasma (Phree ACN-StrataX respectively). Lower S/N for SPM Phree ACN and Phree+StrataX seem to be partly linked to less detected signal overall with a more noticeable impact on peaks (compared to noise), presumably attributable to the common use of Phree ACN. The addition of an additional matrix purification step with the use of SPE cartridge StrataX allowed a better performance of Phree+StrataX compared to Phree ACN alone through a lower noise level in the case of serum.

Repeatability was assessed through semi-quantification performance, representing the percentage of detected compounds with CV ≤ 20% on 4 replicates. PPT and Phree ACN were

the only two SPM that allowed a suitable performance on both serum (94 and 93% respectively) and plasma (81 and 94% respectively). In coherence with the observations presented in the SPM preselection process, StrataX produced less repeatable results compared to Phree ACN, which is further reflected in the Phree+StrataX SPM. Moreover, lower semi-quantification performance values for these two SPM are once again not linked to overall higher CV values for all compounds, but rather to a stronger heterogeneity over the range of compounds. Indeed, CV interquartile ranges are of 4.0%, 6.7%, 13.0% and 18.0% for PPT, Phree, StrataX and Phree+StrataX respectively in serum (8.4%, 6.8%, 14.5% and 26.5% in plasma). High CV values (i.e. CV ≥ 25%) with the use of StrataX and Phree+StrataX SPM in serum were found for compounds that were discussed in the preselection process, such as selective serotonin reuptake inhibitors fluoxetine and paroxetine, as well as triphosphates chlorpyrifos and diazinon. StrataX also seemed to induce low repeatability for triazoles propiconazole and tebuconazole for this real-life-level spiking (10 ng/mL), which was not visible during the preselection phase (40 ng/mL). This observation, coupled with previous reports of comparable repeatability between PPT and SPE-based SPM at high spiking levels (800-5000 ng/mL) [11, 15], suggests the need for application-appropriate evaluations of SPM, as the detection of xenobiotics at real-life concentrations may be further hindered by the choice of an unfitting SPM.

All four SPM allowed the statistical differentiation ($p \leq 0.01$) of spiked compounds areas in spiked and non-spiked samples for both matrices for more than 75% of detected compounds. Overall, PPT and Phree ACN performed best for this criterion, followed by StrataX then Phree+StrataX. This is coherent with the data obtained on repeatability, as significance decreases with repeatability. Indeed, high p-values ($p \geq 0.01$) are generally observed on compounds with high CV values (e.g. diazinon in both matrices, paroxetine in serum, nicotine in plasma, etc.). Phree+StrataX also predictably ranked last regarding the speed of implementation criterion, as the multiplication of extraction steps to achieve further sample purification led to a longer sample preparation process.

Overall, PPT and Phree ACN both present similar and superior performances for the detection of low-level compounds in complex blood-derived matrices compared to StrataX and Phree+StrataX. The study design based on fifty spiked compounds did not allow to demonstrate any clear advantage on one compared to the other; a final comparison of these two SPM was made through their application to serum and plasma cohort samples to obtain a wider point of view on each method's impact on results of a non-targeted exposomics approach.

5.3. Final comparison with MDL determination and application on cohort samples

First, MDL were determined for PPT and Phree ACN on thirty xenobiotics, in plasma and serum. Results on individual compounds are presented in the SI, Table A6. Median MDL values were 0.1 and 0.3 ng/mL for Phree and PPT respectively in both matrices, which suggests lower matrix effect presumably linked to further sample purification with Phree. Contrary to this tendency, some compounds, such as chlorpyrifos and tebuconazole in plasma, present a higher MDL for Phree compared to PPT. Similarly, pravastatin is only detected in samples prepared with PPT in both matrices. Overall, these differences in MDL highlight that the chosen SPM has an effect on both the range of visible compounds and the level at which they are reliably observable.

Further comparison of PPT and Phree ACN was performed by using both SPM to prepare serum and plasma cohort samples (n=10 and 8, respectively). Quality control was performed on the injected batches, both at the targeted and non-targeted scales. Detailed results of the quality control criteria are presented in the SI, Figure S1. Repeatability was assessed at the non-targeted scale through area CV of features found in more than 80% of QC samples. For both SPM and both matrices, more than 80% of QC features presented area CV of less than 30%, which validates the criterion suggested by Want et al. (2010)[28]. Median area CV of all QC features was always less than 20% (11-13%). Similarly, median area CV of IS spiked in QC samples and in cohort samples was always less than 10% (respectively 2-6% and 2-8%). There was little observable difference between SPM or cohorts for these four quality control criteria regardless of the considered scale (i.e. targeted or non-targeted). Lastly, Euclidian distances between analytical replicates were computed. Although all values for median Euclidian distances were satisfactory (<12%), a difference was observed between cohorts, as plasma from the Danish cohort produced more repeatable results compared to serum from Pelagie for both SPM. Moreover, plasma samples prepared using PPT were more repeatable than those prepared using Phree (*p-value*<0.01), whereas no significant effect of SPM could be observed on serum samples.

Following the validation of quality control criteria, suspect screening was performed on the datasets obtained from both cohorts and both SPM using an in-house automatized suspect screening tool[8], followed by manual curation using fragmentation data. In total, 44 and 41 xenobiotics were annotated in the Pelagie serum samples and the Danish plasma samples, respectively. Maximum fold changes (FC) were computed between both SPM for all annotated

compounds, and are reported in Figure III.5. Additional information on individual annotations are available in the SI, Tables A5a and A5b for serum and plasma samples respectively.

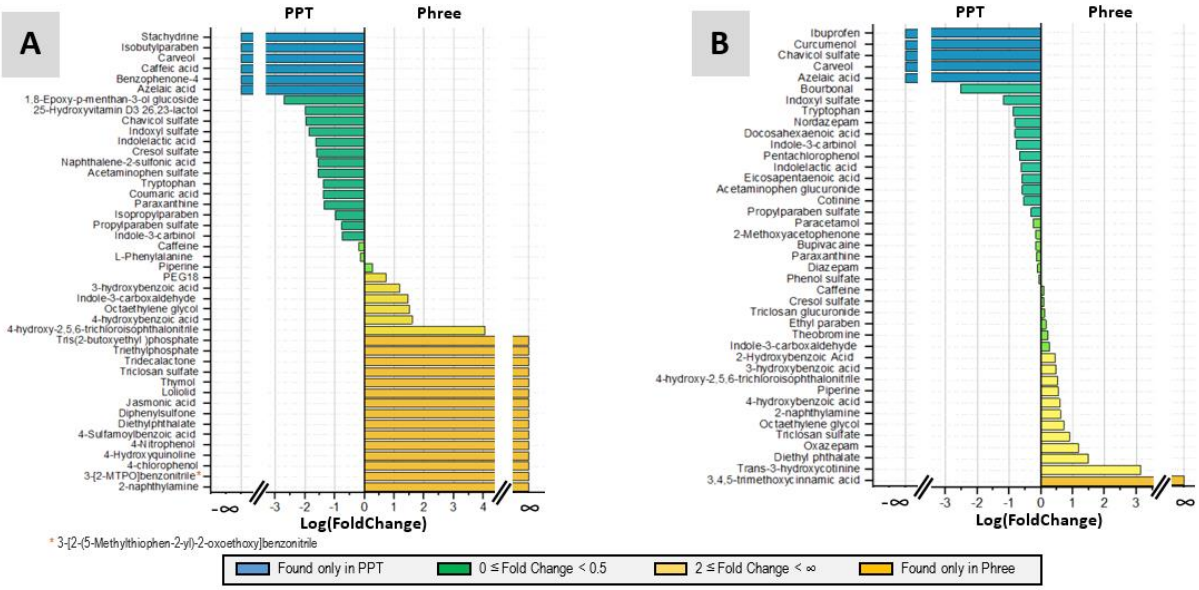

Figure III.5 – Comparison of annotated xenobiotics' areas in samples prepared with protein precipitation (PPT) and protein removal plate Phree in Pelagie serum samples (A) and Danish plasma samples (B). Logged values of fold changes (i.e. area ratio between Phree and PPT) are presented on the x-axis, where –∞ and +∞ values represent the absence of compounds in samples prepared with Phree and PPT, respectively. Bars on the left of the y-axis represent compounds presenting higher areas in PPT samples and vice-versa.

In serum, 93% of annotated xenobiotics presented FC values below 0.5 or above 2, whereas it was the case for only 70% of compounds annotated in plasma, seemingly suggesting a more pronounced effect of SPM on serum than on plasma. As this observation may be skewed by the low amount of annotations compared to the total number of features (>20,000), this tendency was further investigated by computing FC values on QC samples. Results are presented in Table III.1.

Fold change (FC) value	Pelagie serum samples	Danish plasma samples
0 (only in PPT)	38.0%	30.6%
0 < FC ≤ 0.5	9.5%	11.2%
0.5 < FC ≤ 2	28.7%	40.2%
2 < FC < ∞	7.8%	5.3%
∞ (only in Phree)	16.0%	12.6%

Table III.1– Percentage of features of quality control samples categorized by fold change value (i.e. area ratio of features in Phree and protein precipitation). Values are computed for Pelagie serum samples and Danish plasma samples.

Overall, features obtained in serum samples present more differences between the two considered SPM (i.e. FC values closer to the extremes) compared to what is observed in plasma samples. This may be explained in part by the presence of highly abundant and often multiply charged peptide peaks observed in serum samples prepared using PPT, which seem mostly retained during the sample preparation step for Phree samples. These peptide peaks are mostly observed within a specific Rt range (39-45 minutes), which is also the range where phospholipids and lysophospholipids (which are specifically retained by Phree plates) are observed. A comparative visualization of FC values organized by Rt value in serum and plasma is presented in the SI, Figure S2. These peaks are not as abundant in plasma samples prepared with protein precipitation, and therefore present less polarizing FC values. The differentiating presence of these dominating peptide peaks in serum compared to plasma has already been reported[38, 39]. Importantly, in both matrices, more than 40% of feature are only detected using one SPM (43.2-54.0% in plasma and serum, respectively). This highlights the complementarity of these SPM, as they only partially overlap. The use of both PPT and Phree therefore allow to broaden the visible chemical space.

Xenobiotics of various origin were detected using Phree and PPT SPM, including environmental pollutants (e.g. diethylphthalate and chlorothalonil metabolite 4-hydroxy-2,5,6-trichloroisophthalonitrile), compounds used in cosmetic formulations (e.g. octaethylene glycol, benzophenone-4 and various parabens), medication (e.g. paracetamol, diazepam and metabolite nordazepam), and dietary compounds (e.g. caffeine and metabolites, piperine, and flavoring agent bourbonal). This diversity of compounds in terms of polarity ($-0.9 \leq \log P \leq 6.4$), mass ($138.0316 \leq$ monoisotopic mass ≤ 766.4562) and chemical functions underlines the adequacy of these SPM for a wide chemical exposome coverage.

FC values were coherent (i.e. always favored by the same SPM or not favored by any SPM) for compounds detected in both serum and plasma cohort samples, such as tryptophan (FC values of 0.041 and 0.132 in serum and plasma respectively), or caffeine (FC values of 0.65 and 1.25 in serum and plasma respectively). Overall, there is no evident correlation between polarity, mass, or presence of any chemical function and favored detection by either SPM, which does not allow the anticipation of the SPM's effect on other compounds or classes of compounds. This observation underlines the critical need for orthogonal data when aiming for a thorough characterization of a sample, as choice of SPM conditions both the range (i.e. observed compounds) and depth (i.e. observed level) of the visible chemical space.

Documenting the perimeter of the profiled internal chemical exposome for each set of analytical conditions is particularly crucial when aiming for large-scale epidemiological applications. Indeed, non-targeted approaches may be used as exploratory work to identify

previously uninvestigated compounds that are either particularly prevalent or linked to any given health outcomes, potentially resulting in priority lists used in targeted assays focused on quantitation. Yet, the choice of SPM evidently skews the visible information obtained from a sample by either completely preventing the detection of certain compounds, or conditioning it to higher levels in matrix, which may never be reached due to low exposure and/or lack of bioaccumulation. This is not negligible when considering that low-level exposure may still result in toxicity in the case of chronic exposure or low-level exposure to biologically active compounds. For example, known potent toxicant pentachlorophenol is favored by PPT, as is toxicant metabolite triclosan sulfate. In light of this context, biological sample preparation for non-targeted approaches should ideally include multiple SPM to allow a more holistic view of the exposure. Considering the two retained SPM in the case of plasma and serum, this could be achieved by first performing a PPT, followed by a division of the extract between an injection as is (after proper reconstitution) and a further purification using a Phree PLR plate. As biological sample availability is often limited in volume, this suggested sample preparation workflow requires additional effort in miniaturization throughout the process, from the preparation in itself to the injection step. Nevertheless, the gain in terms of coverage of the human internal exposome in both range and depth makes these improvements in efficiency unmistakably worthwhile.

6. Conclusion

Twelve SPM were systematically compared for the HRMS-based non-targeted detection of low-abundant chemicals in complex blood-derived matrices using an innovative methodology based on a large and diverse spiking set at exposure-relevant concentrations. We demonstrated that SPM choice must be investigated with an application-appropriate design, as spiking levels and choice of spiking compounds may greatly affect the understanding of the SPM's impact on non-targeted assays results. The blood-derived matrix choice should also be investigated, as it may affect the observed chemical space. Based on the criteria used in this work, we showed that phospholipid and protein removal plate Phree and the classically used protein precipitation are both well suited to investigate the chemical exposome in serum or plasma samples. Moreover, they can both be used on the same samples, as their complementarity allow the broadening of the visible chemical space.

7. Associated content

7.1. Supporting Information

- “Supporting Information – Tables A“ : chemicals and reagents, detailed results of the SPM preselection, comparison of preselected SPM to protein precipitation, annotations obtained following the application of selected SPM to cohort samples, and methods detection limits (Excel).

- “Supporting Information – Figures S” : Solvents and chemicals, data acquisition parameters, quality control procedures, detailed sample preparation procedures, and quality control data for the cohort applications (Word).

8. Author information

Corresponding Author

*email: arthur.david@ehesp.fr

9. Acknowledgements

This research was supported by a research chair of excellence (2016-52/IdeX Université of Sorbonne Paris Cité) awarded to AD and a grant from the Brittany council (SAD). JC was funded by the Réseau Doctoral en Santé Publique.
The authors acknowledge Aude Dimeglio, Solène Giffard and Romain Letourneur for technical support, and Erwann Gilles for bioinformatics support.

10. References

1. Roth G.A., Mensah G.A., *et al.*, Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol* **2020,** *76* (25), 2982-3021.
2. Soriano J.B., Kendrick P.J., *et al.*, Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Respiratory Medicine* **2020,** *8* (6), 585-96.
3. Andra S.S., Austin C., *et al.*, 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. *Environ Int* **2017,** *100*, 32-61.
4. Chetwynd A.J.,David A., A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. *Talanta* **2018,** *182*, 380-90.
5. Jamin E.L., Bonvallot N., *et al.*, Untargeted profiling of pesticide metabolites by LC-HRMS: an exposomics tool for human exposure evaluation. *Anal Bioanal Chem* **2014,** *406* (4), 1149-61.
6. Panagopoulos Abrahamsson D., Wang A., *et al.*, A Comprehensive Non-targeted Analysis Study of the Prenatal Exposome. *Environ Sci Technol* **2021**.

7. Rafiei A.,Sleno L., Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Commun Mass Spectrom* **2015,** *29* (1), 119-27.
8. Chaker J., Gilles E.*, et al.*, From Metabolomics to HRMS-Based Exposomics: Adapting Peak Picking and Developing Scoring for MS1 Suspect Screening. *Anal Chem* **2021,** *93* (3), 1792-800.
9. Rappaport S.M., Barupal D.K.*, et al.*, The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* **2014,** *122* (8), 769-74.
10. David A., Chaker J.*, et al.*, Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **2021,** *156*, 106630.
11. Tulipani S., Llorach R.*, et al.*, Comparative analysis of sample preparation methods to handle the complexity of the blood fluid metabolome: when less is more. *Anal Chem* **2013,** *85* (1), 341-8.
12. Monteiro Bastos da Silva J., Chaker J.*, et al.*, Improving Exposure Assessment Using Non-Targeted and Suspect Screening: The ISO/IEC 17025: 2017 Quality Standard as a Guideline. *Journal of Xenobiotics* **2021,** *11* (1), 1-15.
13. Caballero-Casero N., Belova L.*, et al.*, Towards harmonised criteria in quality assurance and quality control of suspect and non-target LC-HRMS analytical workflows for screening of emerging contaminants in human biomonitoring. *TrAC Trends in Analytical Chemistry* **2021,** *136*.
14. Pourchet M., Debrauwer L.*, et al.*, Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. *Environ Int* **2020,** *139*, 105545.
15. Sitnikov D.G., Monnin C.S.*, et al.*, Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. *Sci Rep* **2016,** *6*, 38885.
16. Rico E., Gonzalez O.*, et al.*, Evaluation of human plasma sample preparation protocols for untargeted metabolic profiles analyzed by UHPLC-ESI-TOF-MS. *Anal Bioanal Chem* **2014,** *406* (29), 7641-52.
17. Yang Y., Cruickshank C.*, et al.*, New sample preparation approach for mass spectrometry-based profiling of plasma results in improved coverage of metabolome. *J Chromatogr A* **2013,** *1300*, 217-26.
18. Want E.J., O'Maille G.*, et al.*, Solvent-Dependent Metabolite Distribution, Clustering, and Protein Extraction for Serum Profiling with Mass Spectrometry. **2006**.
19. David A., Abdul-Sada A.*, et al.*, A new approach for plasma (xeno)metabolomics based on solid-phase extraction and nanoflow liquid chromatography-nanoelectrospray ionisation mass spectrometry. *J Chromatogr A* **2014,** *1365*, 72-85.
20. Tulipani S., Mora-Cubillos X.*, et al.*, New and vintage solutions to enhance the plasma metabolome coverage by LC-ESI-MS untargeted metabolomics: the not-so-simple process of method performance evaluation. *Anal Chem* **2015,** *87* (5), 2639-47.
21. Gika H.,Theodoridis G., Sample preparation prior to the LC–MS-based metabolomics/metabonomics of blood-derived samples. *Bioanalysis* **2011,** *3* (14), 1647-61.
22. Vuckovic D., Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Anal Bioanal Chem* **2012,** *403* (6), 1523-48.
23. Whiley L., Godzien J.*, et al.*, In-vial dual extraction for direct LC-MS analysis of plasma for comprehensive and highly reproducible metabolic fingerprinting. *Anal Chem* **2012,** *84* (14), 5992-9.
24. Ramesh B., Manjula N.*, et al.*, Comparison of conventional and supported liquid extraction methods for the determination of sitagliptin and simvastatin in rat plasma by LC-ESI-MS/MS. *J Pharm Anal* **2015,** *5* (3), 161-8.
25. Jiang C., Wang X.*, et al.*, Dynamic Human Environmental Exposome Revealed by Longitudinal Personal Monitoring. *Cell* **2018,** *175* (1), 277-91 e31.

26. Knolhoff A.M., Premo J.H.*, et al.*, A Proposed Quality Control Standard Mixture and Its Uses for Evaluating Nontargeted and Suspect Screening LC/HR-MS Method Performance. *Anal Chem* **2021,** *93* (3), 1596-603.
27. Binter A.C., Bannier E.*, et al.*, Exposure of pregnant women to organophosphate insecticides and child motor inhibition at the age of 10-12 years evaluated by fMRI. *Environ Res* **2020,** *188*, 109859.
28. Want E.J., Wilson I.D.*, et al.*, Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* **2010,** *5* (6), 1005-18.
29. Aalizadeh R., Thomaidis N.S.*, et al.*, Quantitative Structure-Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. *J Chem Inf Model* **2016,** *56* (7), 1384-98.
30. Bonini P., Kind T.*, et al.*, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020,** *92* (11), 7515-22.
31. Kim S., Chen J.*, et al.*, PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **2019,** *47* (D1), D1102-D9.
32. Wishart D.S., Tzur D.*, et al.*, HMDB: the Human Metabolome Database. *Nucleic Acids Res* **2007,** *35* (Database issue), D521-6.
33. Horai H., Arita M.*, et al.*, MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **2010,** *45* (7), 703-14.
34. Allen F., Pon A.*, et al.*, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* **2014,** *42* (Web Server issue), W94-9.
35. Ruttkies C., Schymanski E.L.*, et al.*, MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **2016,** *8*, 3.
36. Schymanski E.L., Jeon J.*, et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.
37. Ahmad S., Kalra H.*, et al.*, HybridSPE: A novel technique to reduce phospholipid-based matrix effect in LC-ESI-MS Bioanalysis. *J Pharm Bioallied Sci* **2012,** *4* (4), 267-75.
38. Barri T.,Dragsted L.O., UPLC-ESI-QTOF/MS and multivariate data analysis for blood plasma and serum metabolomics: Effect of experimental artefacts and anticoagulant. *Analytica Chimica Acta* **2013,** *768*, 118-28.
39. Denery J.R., Nunes A.A.*, et al.*, Characterization of differences between blood sample matrices in untargeted metabolomics. *Anal Chem* **2011,** *83* (3), 1040-7.

Chapter IV. Optimizing data processing for exposomics applications: uncovering the potential of low-abundant peaks and MS1 data

1. Context and summary

This chapter was published as an original paper as first author in the journal Analytical Chemistry: **Chaker, J.**, Gilles, E., Leger, T., Jegou, B., & David, A.* (2021). From metabolomics to HRMS-based exposomics: Adapting peak picking and developing scoring for MS1 suspect screening. *Anal Chem* **(IF=6.8)**, 93(3), 1792-1800.

Once an optimized analytical fingerprint of a sample is acquired, this data must be transformed to a list of features, each characterized by a m/z, an Rt, and an area. Features are then aligned for all samples, and annotation or suspect screening may be performed. While many software tools are available to process non-targeted data, most if not all were developed for metabolomics applications. In the case of exposomics, as compounds of interest are often lowly abundant, it is crucial to ensure that data processing tools are capable of accurately disentangling these signals from the noise. The aligned feature lists generated by the data processing step are then used for annotation. As for data processing, there are many available tools relying on various principles to achieve this step. The appropriate tool must thus be found and optimized (i.e. relying on the suitable parameters, implementing a relevant library, etc.) to improve efficiency and lower the number of false positive annotations. A key step of the exposomics workflow therefore consists in optimizing these tools to process non-targeted data. This chapter is the result of two separate optimization steps (i.e. data processing and suspect screening) condensed in one manuscript published in Analytical Chemistry.

The first objective of this chapter was to systematically optimize and evaluate four software tools for the processing of non-targeted exposomics data. This was performed by comparing the processing results of data obtained from plasma and serum samples spiked using a 45-compound set spiked at 10 ng/mL (see Chapter III, paragraph 4.2.2). Each tool was first optimized individually, manually and with automatized parametrization tools when available (i.e. IPO and Autotuner for XCMS), and the best datasets were compared among the tools. Evaluated parameters were detection frequency of spiked compounds, computing time, ease of implementation, area integration repeatability, and detection significance (i.e. significance of the difference in areas between spiked and non-spiked samples).

The second objective of this chapter was to describe the newly developed suspect screening tool . It relies on different chemical predictors (i.e. m/z, experimental and/or predicted retention time, as well as isotopic fit) to score the proximity between features and suspects, and therefore provides an easy-to-read indicator of each annotation's reliability. The modelling of these predictors is described, and their relevance and accuracy are illustrated through an application to non-spiked samples.

Results of these comparisons are described and discussed throughout this article.The inadequacy of existing automatized parametrization tools built for metabolomics applications is discussed. Moreover, the need for tailored and optimized tools for processing HRMS-based exposomics data is underlined. Furthermore, the usefulness of confidence indices for suspect screening implemented in the in-house tool is demonstrated as an efficient way to prioritize annotations for manual curation.

From metabolomics to HRMS-based exposomics: adapting peak picking and developing scoring for MS1 suspect screening

Jade Chaker, Erwann Gilles, Thibaut Léger, Bernard Jégou, Arthur David*

Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) – UMR_S 1085, F-35000 Rennes, France

To whom correspondence should be addressed:
Tel: +33 299022885
email: arthur.david@ehesp.fr

2. Abstract

The technological advances of cutting-edge high-resolution mass spectrometry (HRMS) has set the stage for a new paradigm for exposure assessment. However, some adjustments of the metabolomics workflow are needed before HRMS-based methods can detect the low-abundant xenobiotics in human matrices. It is also essential to provide tools to speed up marker identifications. Here, we first show that metabolomics software packages developed for automated optimization of XCMS parameters can lead to a false negative rate of up to 80% for chemicals spiked at low levels in blood. We then demonstrate that manual selection criteria in open source (XCMS, MZmine2) and vendor software (MarkerView™, Progenesis QI) allow to decrease the rates of false negative up to 2% for these spiked chemicals. We next report an MS1 automatized suspect screening workflow that allow for a rapid pre-annotation of HRMS datasets. The novelty of this suspect screening workflow is to combine several predictors based on m/z, retention time (Rt) prediction models and isotope ratio to generate intermediate and global scorings. Several Rt prediction models were tested and hierarchized (PredRet, Retip, RTI and a logP model), and a non-linear scoring was developed to account for Rt variations observed within individual runs. We then tested the efficiency of this suspect screening tool to detect spiked and non-spiked chemicals in human blood. Compared to other existing annotation tools, its main advantages include the use of Rt predictors using different models, its speed and the use of efficient scoring algorithms to prioritize pre-annotated markers and reduce false positives.

Key words: Exposomics, high-resolution mass spectrometry, exposure assessment, peak picking, suspect screening, annotation tool

Graphical abstract

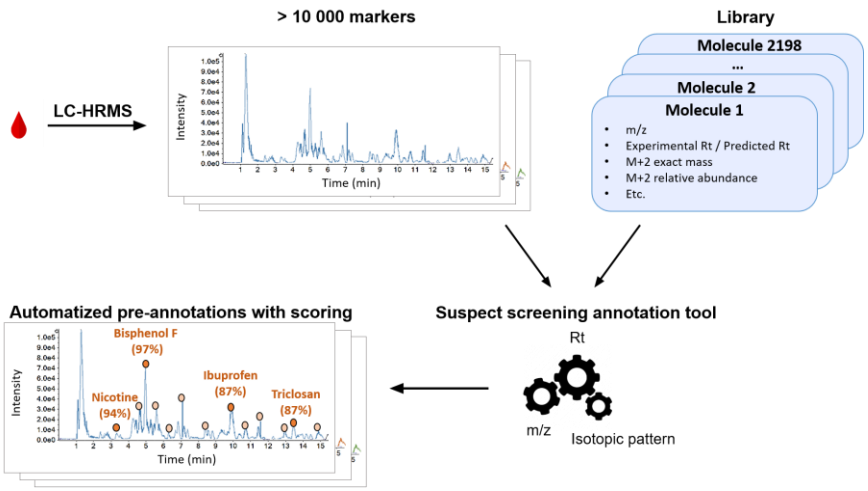

Figure IV.1 – Graphical abstract for the research paper titled "From metabolomics to HRMS-based exposomics: adapting peak-picking and developing scoring for MS1 suspect screening"

3. Introduction

Recently, the technological advances of cutting-edge high-resolution mass spectrometry (HRMS) has set the stage for a new paradigm to assess human exposure to complex mixtures of xenobiotics.[1] Using HRMS platforms coupled to liquid chromatography (LC), it is now possible to profile several thousands of small molecules (<1500 Da) in a biological sample during a single analysis, including both endogenous and exogenous molecules and their transformation products.[2, 3] The holistic characterization of exogenous chemical mixtures accumulating in human biological samples (i.e. the internal chemical exposome) using HRMS platforms would be a step forward to investigate the environmental aetiology of many multifactorial chronic diseases with an unprecedented precision.[1, 4, 5] It is therefore paramount to break down the remaining technological barriers and methodological issues to be able to perform large-scale exposomics studies using HRMS-based methods.

One of the obstacles to overcome is the analytical sensitivity issue which is currently preventing the detection of low-abundant exogenous chemicals in complex biological matrices.[6, 7] Concentrations of environmental contaminants can be on average 1,000 times lower than concentrations of endogenous chemicals and food chemicals in human blood.[8] Improving the analytical sensitivity of LC-HRMS platforms is therefore a necessary step to go from metabolomics-oriented studies toward exposomics studies.[2, 3] It is also critical to ensure that bioinformatics tools designed to process LC-HRMS data can disentangle chemicals' small signals from the noise. To this aim, optimization of adjustable parameters in software available for processing raw data is a key step to ensure that even the low abundant chemicals of interest will be picked up.[9, 10] Software packages such as the IPO[11] or Autotuner[12] have already been developed for automated optimization of XCMS parameters to improve the detection of reliable signals. Studies comparing automated optimization and manual selection criteria for metabolomics applications have already been performed.[13] However, these studies are missing for exposomics applications where the aim is also to include infrequent signals often close to the noise that can be used to identify unrecognized exposure.[14]

Besides sensitivity, another bottleneck preventing comprehensive characterization of exogenous chemical mixtures present in biological samples is the annotation of the thousands of signals present in HRMS datasets. Over the years, many annotation tools (e.g. CAMERA, ProbMetab, MolNetEnhancer and MetAssign) relying on analytical predictors (e.g. m/z, Rt, isotopes) and correlation/clustering methods have been developed for metabolite annotation.[15-18] More recently, annotation tools such as xMSannotator[19] have incorporated biological correlations in addition to analytical correlations while other tools are now integrating MS/MS[20-22] to improve metabolite annotation. Besides these annotation tools, the qualitative

suspect screening approach is also being increasingly used to prioritize relevant xenobiotics for human exposure assessment.[23, 24] Suspect screening uses exact mass of suspects from in-house libraries/database as a priori information.[25, 26] Compared to other annotation tools which often rely on large databases such as HMDB[27], KEGG[28] or ChemSpider,[29] the suspect screening strategy can be less time-consuming in particular if the list of suspects arises from a systematic prioritization strategy. However, predictors other than exact mass must be added in the suspect screening workflow to decrease the rate of false positives and therefore limit the number of putative annotations that need manual curation.

Here, we first compared the ability of metabolomics software packages developed for automated optimization of XCMS (IPO[11] and Autotuner[12]) and manual selection criteria to detect low-abundant spiked chemicals. Manual optimization was also extended to another open source software (MZmine2[30]) and 2 vendor tools (e.g. MarkerView™ and Progenesis QI) to compare their efficiency to detect low-abundant spiked chemicals. We demonstrate the importance of fine-tuning critical parameters for both open source and vendor software to dramatically decrease the rate of false negatives. We next report an MS1 automatized suspect screening workflow. The novelty of this suspect screening workflow is to combine several predictors based on m/z, Rt prediction models and isotope ratio checks to generate intermediate and global scorings using multi-criteria algorithms. Several Rt prediction models were tested and hierarchized, and a non-linear scoring was developed to account for Rt variations observed within individual runs. We show this suspect screening tool's high efficiency for the rapid annotation of low-abundant spiked and non-spiked exogenous chemicals in human plasma and serum (annotation confirmed with MS/MS data). Compared to other existing annotation tools (e.g. xMSannotator,[19] MS-DIAL,[20] msPurity[21]), its main advantages include the use of Rt predictors based on different models, its speed and the use of efficient scoring algorithm to prioritize pre-annotated markers and reduce false positives.

4. Experimental section

4.1. Spiking experiments and sample preparation

Spiking experiments were performed on human blood plasma and serum samples to optimize and compare the efficiency of data processing software to detect low-abundant signals in biological samples. Plasma and serum bags were acquired from the French blood agency (Etablissement Français du Sang, EFS). Homogenate plasma and serum samples (n=4, 100 µL each) were spiked with a mix of selected classes of contaminants (i.e. pesticides, pharmaceuticals and diet toxins) and metabolites (i.e. steroids, eicosanoids and neurotransmitters) (n=45, see Supporting Information Table A1 for suppliers) to give 10 ng/mL

in matrix. Non-spiked plasma and serum samples (n=4, 100 µL each) from the same homogenates were also used to check for any background contamination. A mix of 21 labeled internal standards (100 ng/mL) was used to monitor analytical variabilities during sample preparation and UHPLC-ESI-QTOF injections. Protein precipitation was performed using a 4:1 (v:v) ratio of cold methanol to matrix. To improve protein removal, samples were allowed to stand at -20°C for one hour prior to centrifugation. After centrifugation at 4°C and 17,000g for 20 min, supernatants were collected and evaporated to dryness under vacuum. Samples were recovered in 100 µL of 90:10 (v:v) ultrapure water to acetonitrile ratio.

4.2. Data acquisition and quality control

Samples were analyzed on an AB SCIEX X500R QTOF interfaced with an AB SCIEX ExionLC AD UHPLC. Compound chromatographic separation was achieved with an Acquity UHPLC HSS T3 C18 column (1.8µm, 1.0 x150mm) maintained at 40°C. Details regarding the injection, chromatographic separation and ESI source parameters can be found in the SI. Samples were analyzed in full scan experiment in both – and + ESI modes. MS/MS fragmentation data for chemical elucidation was obtained by analysis of selected samples in sequential window acquisition of theoretical mass spectrum (SWATH). Quality Control procedures are specified in the SI.

4.3. Peak picking optimization: data processing tools

Mass spectra acquired in full scan were processed (peak picking, deconvolution, alignment, gap filling) using four software programs: instrument-specific software MarkerView™1.3 (AB SCIEX), vendor software Progenesis QI for Metabolomics (Waters), and open-source solutions MZmine2[30] (v2.51) and XCMS[31] (v3.6.1). Two R packages, IPO[11] and Autotuner[12], were used to test automatized parameter optimization of XCMS. For Progenesis, XCMS and MZmine2, raw data files (in wiff2 data format) were converted to 64 bit .mzML (full scan) using MSConvert from ProteoWizard.[32] Two pipelines were used within the MZmine2 solution: Automated Data Analysis Pipeline (ADAP) (with “ADAP Chromatogram Builder” and “Chromatogram Deconvolution – Wavelets (ADAP)” steps), and Continuous Wavelet Transformation (CWT) (with “Chromatogram Builder” and “Chromatogram deconvolution – Wavelets (XCMS)” steps). For all software, a set of default parameters and a set of optimized parameters were tested to ensure optimal detection of spiked compounds (Figure IV.2).

Five criteria were established to compare the four software tools and all possible parameter optimization algorithms. First, the detection frequency of spiked chemicals in blood plasma and serum samples was used to study the efficiency of parameters optimization. Then, mean areas for spiked and non-spiked samples, associated standard deviations, fold changes, and p-

values (unpaired t-tests) were computed to model the detection significance in spiked versus non-spiked samples (threshold set at p=0.05). The semi-quantification performances were the third parameter implemented; the percentage of spiked compounds with area coefficient of variation (CV) below 30% were compared for all software according to the criteria proposed by Want et al.[33] Independent peak integration of all spiked compounds were carried out using Sciex OS software (v1.2) to validate the accuracy of these 3 parameters. The fourth parameter was computing time (computer configuration available in SI Table A2) and the last one was ease of implementation (based on presence and user-friendliness of GUI, as well as number of customizable parameters).

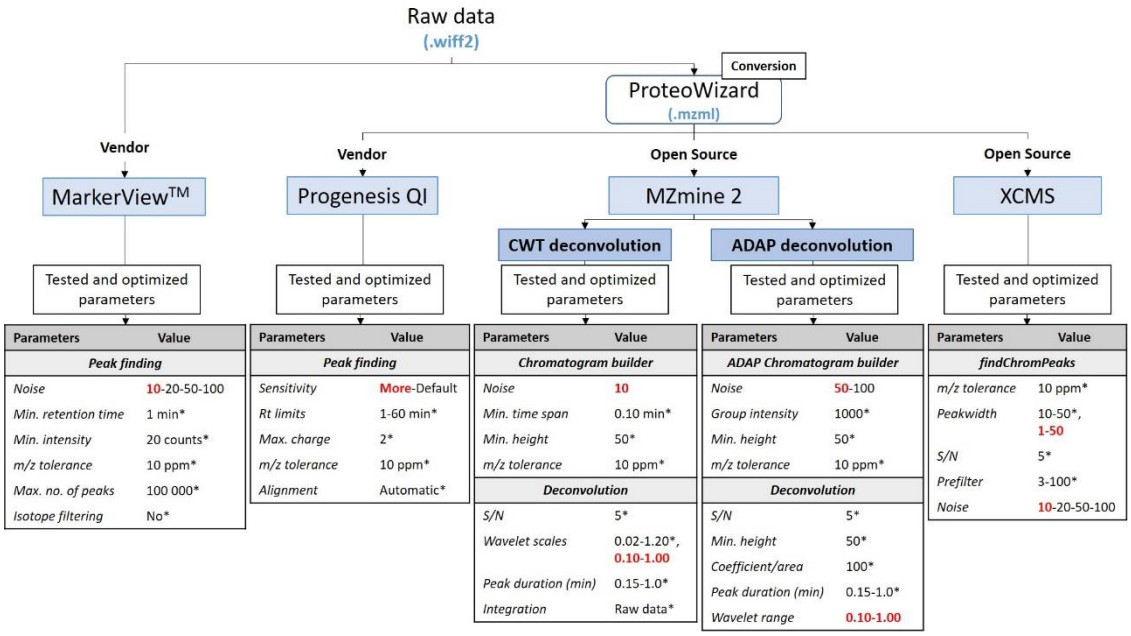

Figure IV.2 - Data preprocessing flowchart illustrating all tested parameters, including default parameters for the authors' system (*) and optimized parameters (in bold and red) for each data preprocessing software tool.

4.4. Suspect screening predictors

4.4.1. Mass-to-charge ratio (m/z)

Mass-to-charge ratios were calculated in-house using atomic monoisotopic masses obtained through the MIDAs C++ program (Molecular Isotopic Distribution Analysis)[34] with the Fast Fourier Transform (FFT)-based method (nucleon domain).

4.4.2. Retention time

Four tools were used to attempt modelling Rt. Two models were first constructed using a training set of 134 standards and then evaluated using a set of 30 standards (see SI Table

A3). Experimental Rt for these standards were acquired from repeated injections (n=4). Simple regression linear models were used; adjusted coefficients of determination R^2_{adj} were computed to describe correlation between variables and standard deviation of predictions. Models were considered validated if a R^2 value greater than 0.7 was reached.

The first Rt prediction model was constructed using octanol-water partition coefficients (logP). Although compounds may be ionized in the considered experimental conditions, logP was preferred to logD since experimental logD values are hardly available. LogP values were extracted from PubChem.[35] Compounds of the training set were only used for model construction if an experimental logP was available (see SI Fig.B2). Experimental Rts were regressed on experimental logP values from compounds of the training set for which this parameter was available (n=101). The resulting equation was used to predict Rts for validation set compounds.

A Quantitative Structure-Retention Relationship-based tool, available on the online Retention Time Indices (RTI) platform, was used to construct the second model through correlation of Rt and chemical structure of a compound,[36] and is calibrated using two sets of nineteen compounds (see SI Table A4). Compounds from the training and evaluation sets were submitted through the “Batch mode” pipe, using the “Chemical Space Boundary” uncertainty measurement. Experimental Rt were regressed on RTI values of compounds of the training set for which a RTI value was generated (n=99), as some compounds were out of the model’s applicability domain. Rt values for the validation set compounds were predicted using the resulting equation to perform model evaluation.

The third and fourth Rt prediction tools did not require the construction of a model, as a predicted Rt value was directly available. Retip[37] relies on five machine learning algorithms, and requires previously acquired experimental Rt values and InChI identifiers. PredRet[38] uses a user-driven database of compounds Rt to return a prediction of a compound’s Rt if it has been determined in a similar chromatographic system. To implement this last tool, the in-house chromatographic system was described: column type, column, eluents and additives were specified. Compounds from the training and validation set as well as their InChI identifiers were inputted.

4.4.3. Isotopic pattern

Theoretical isotopologue probabilities P_0, P_1, and P_2 (i.e. first, second, and third isotopologue of masses M_0, M_1 and M_2) for all compounds from the training and validation set were computed using the MIDAs[34] software with the FFT-based method. Experimental isotopologue abundances A_0, A_1 and A_2 were determined through targeted data processing. P_2/P_0 was

regressed on A_2/A_0 for all standards for which an experimental A_2 was detected (n=103). Prediction bands (99%) were determined to estimate confidence in A_2/A_0 ratio value.

4.5. Suspect screening annotation tools

A two-part Visual Basic program was used to automatize part of the suspect screening annotation step. The two parts of the suspect screening program were created to respectively generate the predictors for a suspect list database and then test the correlations between suspects and markers from HRMS datasets. The database includes 2198 compounds commonly detected in human blood referenced in databases such as HMDB[39] or the Blood Exposome Database.[40] In this database, three predictors (m/z, Rt, and isotopic fit) were generated for each suspect: suspect compounds were associated to a formula, M_0, M_2, P_0 and P_2 values, and experimental or predicted Rt values if available. The library computes monoisotopic mass, as well as common adducts masses ($[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$ for positively charged adducts, and $[M-H]^-$, $[M-H_2O-H]^-$, $[M+Cl]^-$, $[M+FA-H]^-$ for negatively charged).

The second part of the program, which performs the pre-annotation, scans individual markers (Mass × Rt), evaluates their proximity to the suspects using Confidence Indices (CI), and prioritizes the best candidate, if any. CI were built for each predictor as shown in Equation IV.1. A global confidence index (GCI) was also built as the mean of the three CI.

$$\mathrm{CI_i} = 1 - \frac{\left| \frac{\mathrm{i_{feature}} - \mathrm{i_{suspect}}}{\mathrm{i_{suspect}}} \right|}{\Delta_\mathrm{i}}$$

Equation IV.1 – Expression of Confidence Indices (CI) for all predictors (i= m/z, Rt, or M_2/M_0 ratio). Δ_i is a confidence interval and is specifically defined for each predictor as the maximal acceptable deviation from the reference value.

The maximal acceptable deviation for mass $\Delta_{m/z}$ was defined based on instrumental uncertainty, and can take two values: 15 ppm for masses strictly lower than 200 Da, and 10 ppm for masses over 200 Da.

The Δ_{Rt} value was determined based on analytical Rt variability. This variability was estimated by computing Rt standard deviation (SD) for spiking standards in all analyzed spiked sample (8 spiked plasmas and sera) and for all isotopically-labeled compounds spiked in all analyzed sample (16 spiked and unspiked plasmas and sera). As analytical variability in Rt is heterogeneous along the chromatogram, run time was divided in sections based on observable different variability levels. The maximal expected Rt deviation Δ_{Rt} was constructed by selecting the highest compound Rt SD for each section, matrix and mode, and multiplying by three

(assuming normal distribution and applying statistics' empirical rule to encompass 99.7% of values). Highest SD was selected in order to avoid excessive stringency and account for untested factors such as long-term analytical drift.

The $\Delta_{A2/A0}$ value was also set by using the empirical rule: RMSE of A_2/A_0 error (ratio of integration of experimental signals vs. ratio of theoretical abundances) as presented in 6.3 was multiplied by three. The calculation of the CI for this last predictor is based on a step-wise approach. Indeed, the software tool first provides the likeliness of presence of the M_2 isotopologue in the feature table, and then proceeds with the abundance A_2/A_0 ratio computing. A more detailed representation of this tool's workflow is available in SI Fig.B1.

This in-house suspect screening tool was compared to four already available annotation and/or suspect screening software tools: xMSannotator,[19] MS-DIAL,[20] msPurity[21] and MZmine2.[30] The following criteria was used for comparison: possible use of in-house libraries or existing databases, possible use of experimental and/or predicted Rt for annotation (as opposed to clustering), use of MS2 predictor, scoring, and prioritization of annotations.

4.6. Data availability

The data files and associated metadata are available as .mzML in the MetaboLights repository[41] under the identification number: MTBLS1785.

5. Results and discussion

5.1. Optimization of HRMS data processing tools for exposomics studies

Independent peak integration of all spiked compounds ensured they provided reliable signals. In plasma (serum), signal/noise values ranged from 38 to 1.3E+7 (23 to 1.4E+6), median m/z and Rt shifts were 1ppm and 0.1mn (1ppm and 0.1mn), peak asymmetry factors were averaging at 1.47 (1.41) and all below 1.86 (1.76), and area values were above 3.6E+3 (2.7E+2).

5.1.1. XCMS: automated optimization versus manual selection criteria

The ability of two automatized optimization tools IPO and Autotuner, which were both developed for metabolomics applications, were tested for R-implemented open source XCMS (Fig.2). IPO-optimized parameters allowed detection of only 29% of spiked compounds in plasma (20% in serum), but with a maximal semi-quantification score. Since IPO optimization

parameters rely on "reliable peaks" which are defined based on the identification of ^{13}C peaks using 3 criteria relative to the ^{12}C peak,[11] we can only assume that these criteria were too stringent for many spiked compounds although they produced relevant analytical criteria for both detection (see above) and annotation (including relevant MS/MS spectra). Since low abundant peaks did not necessarily answer the algorithm's criteria, parameter optimization such as "max peakwidth" were too high (30.7s) for most of these signals. Autotuner, on the other hand, allowed the detection of 73% of spiked compounds in both matrices, but less than 20% of them had an area CV lower than 30% on four replicates. The "max peakwidth" parameter is tuned to a low value (less than 10s), which is not coherent with the width of the considered compounds, leading to a splitting of peaks and thus a less reliable integration value. These results highlight the necessity to adapt tools built for metabolomics to the needs of chemical exposomics, and underlines the already described efficiency of manual tuning when dealing with less optimal peaks.[13]

XCMS was then tested using four sets of parameters (Figure IV.3). Firstly, the set further referred to as "default parameters" was determined by a priori adaptation of suggested parameters for detection of low-abundant chemicals in complex matrices. Secondly, through visual examination of the data (data not depicted), the "peakwidth" parameter from the "centWave" function was determined to be sensitive and was optimized: the minimal time for a peak identification was set at 1 second to account for narrow signals. This allowed to increase the detection percentage of spiked compounds of 18 points in plasma (64 to 82%) and 11 points in serum (60 to 71%).

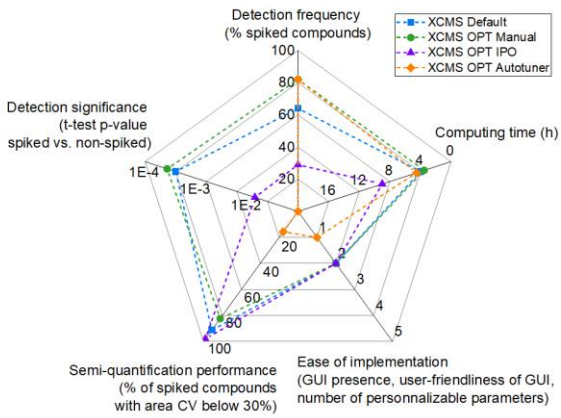

Figure IV.3 - Data processing (i.e. peak picking, deconvolution, alignment, gap filling) evaluation using XCMS for detection and semi-quantification of low-level spiked compounds in plasma samples (n=4 each). Four sets of parameters were used: Default (blue squares), manual (green rounds), IPO (purple triangles), and Autotuner (orange diamonds) optimization. Outer edges identify best performances.

5.1.2. Manual optimization of MZmine 2 and vendor software

As for XCMS, individual optimization of each tool was mandatory to decrease the rate of false negatives and determine relevant parameters to detect low-abundant compounds.

For MZmine2, the CWT and the ADAP pipelines were both optimized and compared. Default CWT parameters refer to values used by Myers et al. (2017) for plasma samples.[10] Within this pipeline, the "wavelet scales" parameter was identified as critical through GUI data visualization. The optimized bracket (0.10-1.00) showed a 9-point increase in detection frequency of spiked compounds in plasma and a 6-point increase in serum compared with the default bracket (0.02-1.20). This comes at the cost of computing time, which almost doubles for plasma samples and increases about 15% for serum samples.

As for the ADAP pipeline for MZmine2, parameters were optimized through data previsualization. Wavelet range parameter was identified as critical, and the bracket 0.10-1.00 was determined to be most appropriate. The two optimized pipelines were compared at the lowest attainable noise level with the available hardware (10 for CWT, 50 for ADAP). ADAP presented better results in terms of spiked compounds detection percentage (82% to 96% in plasma and 84% to 89% in serum).

For MarkerView™ and Progenesis QI vendor software, only few parameters can be modified and the most critical one is the noise threshold. For MarkerView™, three lower values (e.g. 50, 20 and 10) were tested in addition to the default (100). Intensity threshold value of 10 was determined to be optimal, with a detection of 89% of spiked compounds in plasma and 82% in serum. For Progenesis QI, the automatic sensitivity method was used and sensitivity values "default" and "more" were tested. The "more" value was selected as optimal, as compared to the default, detection of spiked compounds increased 18 points for plasma (62 to 80%) and 11 points for serum (67 to 78%).

5.1.3. Comparison of optimized data processing tools to detect low abundant compounds

The ability to detect low-abundant chemicals in plasma and serum was assessed for the 4 software tools (Figure IV.4). Detailed results for each spiking compound and all tested parameters is available in SI (Table A5 and Figure B4). For both matrices, MZmine2 offers the best detection frequency of spiked compounds. As for detection significance between spiked samples and non-spiked samples, all tools allowed to reach the 0.05 p-value threshold to describe a significant difference in areas of detected spiking compounds between spiked and non-spiked samples. Median detection significance (t-test p-value) in plasma is lower (i.e.

higher p-value) for Progenesis QI compared to the other three tools. In serum, XCMS gives the lowest detection significance.

Regarding semi-quantification performance, all tools allowed to pass the repeatability criteria from Want et al.[33] (i.e. feature integration such as more than 80% of detected spiked compounds had an area CV lower than 30%). In serum, similar values are achieved for all software programs (between 80% for MZmine2 and 86% for MarkerView™). In plasma, value for this parameter was significantly better using MZmine2 compared to the other three tools.

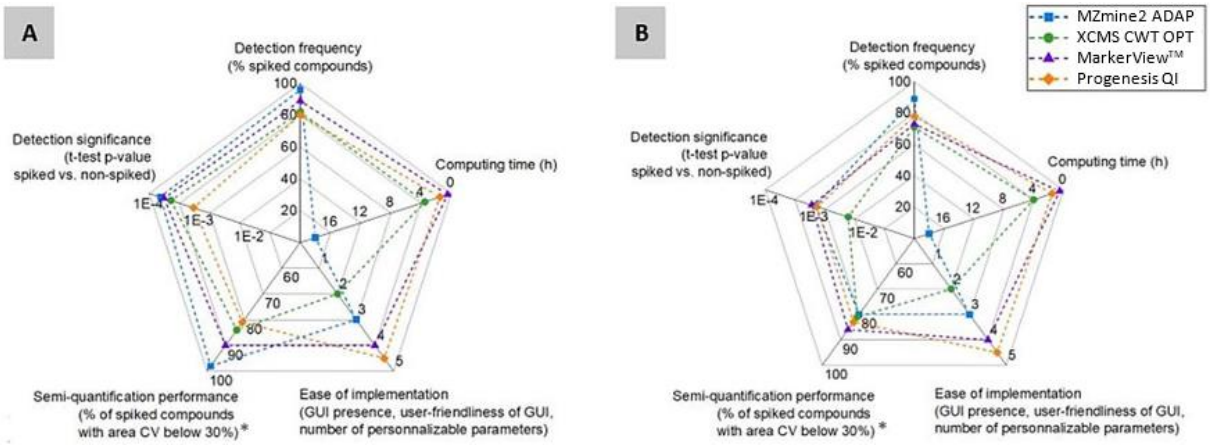

Figure IV.4 - Data processing (i.e. peak picking, deconvolution, alignment, gap filling) evaluation for detection and semi-quantification of low-level spiked compounds in (A) plasma and (B) serum samples (n=4 each). Four optimized software tools were used: MZmine 2 (blue squares), XCMS (green rounds), MarkerView™ (purple triangles), and Progenesis QI (orange diamonds). Outer edges identify best performances.

As for computing time and ease of implementation, vendor software tools have the best performances. These tools are the fastest and easiest to implement, as they have user-friendly GUIs and require little to no building of the processing pipeline. Progenesis QI also offers visual reviewing of the data which allows the user added control. MZmine2 is the most time-consuming data processing tool (averaging at 18 hours). XCMS is the most flexible and is constantly evolving but is less user friendly as it uses command-line interface.

In conclusion, the four investigated data processing tools, when optimized, presented acceptable performance regarding detection frequency, detection significance in spiked versus non-spiked samples and semi-quantitative performance. Vendor tools made a significant difference regarding computing time. MarkerView™ is particularly interesting since it was proven to be effective over the five indicators. Its main disadvantage is its “black-box”-like functioning, with little user input or overview and only accepting wiff2 format. MZmine2 had the

best performance for spiking standard detection, and offers both a GUI and control on the data processing workflow, but its comparatively long computing time can stifle its systematized use. Improvements are nevertheless still required to improve the detection of low-abundant signals close to the baseline for all software to decrease the remaining false negatives.

5.2. Modelling suspect screening predictors

We next developed a suspect screening workflow that incorporate for the first time several Rt prediction models in addition to m/z and isotope ratio checks. Multi-criteria algorithms were then developed to generate intermediate CI for each predictor as well as a global CI built as the mean of the three CI.

5.2.1. Retention time prediction models

Four tools were used to attempt Rt modelling: an in-house model based on logP, Retip, RTI, and PredRet. However, PredRet could not be retained for further comparison with the other models since predicted Rt were returned for 16 compounds out of 134 submitted (12% response rate) which is significantly lower than what was obtained for RTI (74% response rate). Lack of data regarding previous injections of those standards on similar chromatographic systems could explain these results. This highlights the need for community participation to such tools, for a more thorough coverage of chromatographic systems and compounds.

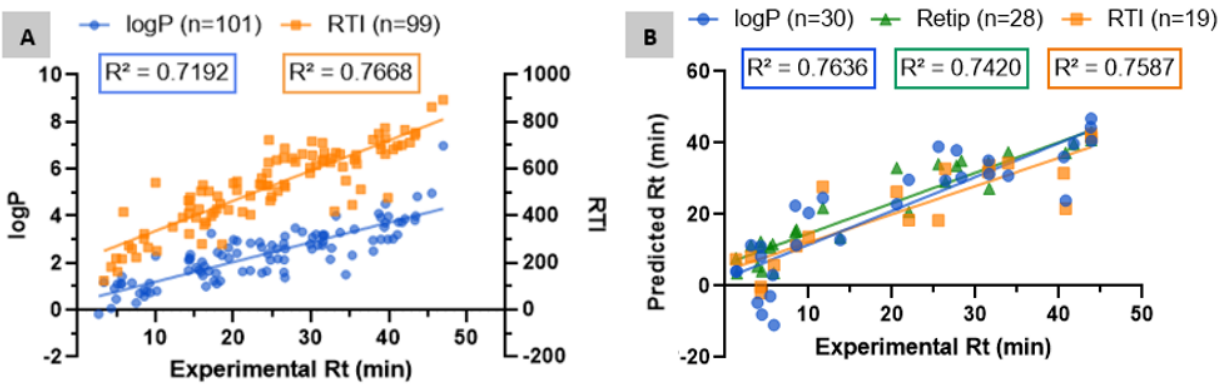

Figure IV.5 - Construction (A) of two Rt prediction models using simple linear regression models and validation (B) of all usable Rt prediction models. The logP model uses experimental octanol-water partition coefficients as predictors and the RTI model uses Retention Time Indices (RTI) as predictors. PredRet, a fourth Rt prediction tool, was also tested. PredRet predictions are not depicted as the number of responses were significantly lower (n=16), rendering it not statistically comparable.

Linear regression models construction for logP and RTI models are presented in Figure IV.5A, and results on the validation set for the logP, Retip and RTI models are presented in Figure IV.5B. A coefficient of determination of 0.72 was obtained for the logP model. R^2 value was higher compared to other models constructed similarly, such as the ones described by McEachran et al. (2018)[42] (R^2=0.66 on 78 compounds) and Bade et al. (2015)[43] (R^2=0.67 on 595 compounds). This was expected as experimental logP values were exclusively used to build this model to avoid accumulating error from logP modelling and Rt modelling. The model constructed using RTI values presented a R^2 value of 0.77. This model's performance is coherent with the RTI developers' model description[36] namely a R^2 value around 0.84.

Both models as well as the Retip model were then validated using a 30-compound validation set; both R^2 values were similar, although with 28 and 19 compounds for Retip and RTI, respectively, since some compounds were not covered by the models. RMSE values were found to be of 13.7%, 12.6% and 11.5% of run time for the logP, Retip, and RTI models, respectively, suggesting a more precise prediction of Rt using the RTI model, then Retip, then the logP model. Based on these results, the four possible Rt values for a given compound were hierarchized for determination of CI as follows: experimental Rt if available, followed by the RTI predicted, Retip predicted, then logP predicted. In addition to evaluation of Rt prediction tools, analytical Rt variability was investigated to avoid excessive stringency in the CI calculation by accounting for fluctuations in analytical variability caused by the matrix or conditions of elution over the course of the analysis (see SI Fig.B3). Computed SD for compounds were plotted against run time and allowed the creation of four sections based on visual inspection of the data: 0-5 min, 5-15 min, 15-30 min, and 30-60 min. The third section (15-30 min) showed maximal Rt variability for all matrix×mode combination, whereas the second section (5-15 min) presented lowest Rt variability in all cases except for compounds in serum in ESI (-) mode (where variability was lower in first section).

Overall, similar values were found within each sector for all four tested conditions (matrix × ionization mode), as lowest SD was always less than 15s from highest SD in a given sector. Therefore, to avoid multiplication of conditions, highest SD was selected for each sector, and multiplied by three to define maximal acceptable deviation for Rt depending on absolute Rt. The obtained variable is referred to as Δ_{Rt} and takes the value of 0.28, 0.21, 1.29 or 0.93 min if the compounds has a Rt of 0-5, 5-15, 15-30, or 30-60 min respectively.

5.2.2. Isotopic pattern

Isotopic pattern distribution was described using the ratio of third to first isotopologue A_2/A_0. The linear regression correlating theoretical P_2/P_0 and experimental A_2/A_0 ratios (n=98) is

represented in Figure IV.6. A R^2 value of 0.996 and a RMSE of 0.02 were achieved, suggesting a high similarity between these two ratios and thus confirms a practical feasibility of using this ratio for suspect screening with the applied conditions.

The investigated compounds were separated into eleven groups based on contents in Br, Cl and S atoms (and combinations). Compounds constituting these groups formed varyingly distant clusters. Five main clusters are formed based on contents in halogens Br and Cl (no

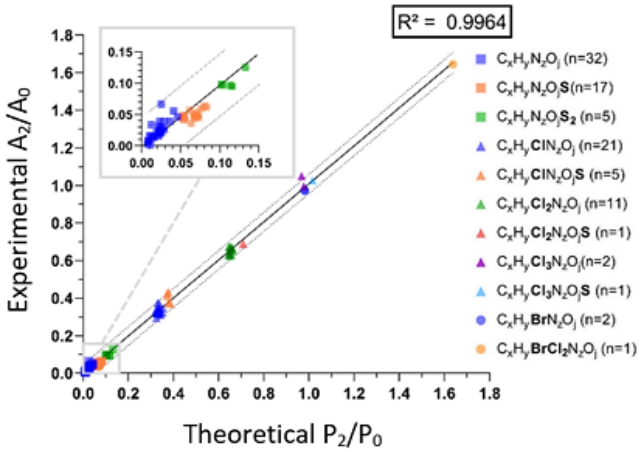

Figure IV.6 - Linear regression analysis of A_2/A_0 according to P_2/P_0. Prediction bands placed at 3 RMSE (99%) are depicted in dotted lines. Compounds are separated into 11 groups based on contents in Br, Cl, and S atoms (and combinations)

halogens, one Cl, two Cl, three Cl or one Br, and combination of one Br and two Cl), which largely influence A_2/A_0 value. It is also observed that compounds' content in S atoms dictates their placement within each of these five main clusters, which is coherent with the $^{34}S/^{32}S$ ratio value of 0.05.

Prediction bands were placed at 3 RMSE to establish a limit where more than 99% of future points are expected to be placed. A value for maximal acceptable deviation between P_2/P_0 and A_2/A_0 ratio of 0.1 was determined from the width of prediction bands. Given this value, it would be possible to discriminate compounds from different major clusters (i.e. based on Br or Cl content), but not compounds from groups with equal contents in halogens and different contents in sulfur. The maximal acceptable deviation value of 0.1 is identified as $\Delta_{isotopic\ fit}$ and is used to assist suspect screening approaches by determination of the CI for isotopic pattern (or $CI_{isotopic\ fit}$). This CI is implemented in the suspect screening annotation tool.

5.3. Efficiency of the suspect screening tool and comparison with other annotation tools

The in-house automatized suspect screening annotation tool was applied to the eight spiked plasma and serum samples to assess its performance. In addition, four other annotation and suspect screening tools were used for comparison (Figure IV.7).

Results for the in-house suspect screening annotation tool individual compounds are available in SI Table A6, and SI Fig.B5. Overall, 100% of spiked compounds that were picked up could be pre-annotated in plasma and serum samples, with an average of 1.1 suggested markers per compound in both plasma and serum when filtering on $CI_{m/z} > 0.7$ and $CI_{Rt} > 0.5$. A $CI_{isotopic\ fit}$ was computed for 31% of detected spiked compounds in plasma, and 36% in serum. Mean $CI_{m/z}$, CI_{Rt} and $CI_{isotopic\ fit}$ values for detected spiked compounds in plasma (serum) were 0.82 (0.83), 0.98 (0.97), and 0.76 (0.71). Overall, all three mean CI were found to be over 0.70 for spiked compounds, which highlights the relevance of these indicators for pre-annotation. Using our library of 2198 chemicals, the time needed to generate this pre-annotation after data acquisition was less than 2h (50 min for MarkerView data processing and 1 h for the pre-annotation VBA-based program).

It is important to mention that it is quite difficult to compare all annotation tools since they do not work the same way and have different purposes. Indeed, some tools use specific analytical predictor such as the MS2 for the annotation (MS-DIAL, msPurity, MZmine2) while our in-house tool is the only one to rely on Rt prediction models. Considering these limitations, we observed that frequency of detection in plasma (and serum) were, respectively, of 100% (100%) for MZmine2, 79% (79%) for MS-DIAL, 79% (79%) for MS-DIAL, 77% (73%) for msPurity, and 66% (66%) for xMSannotator. Since different factors inherent to the tool could be involved in the difference of frequency of detection of this selected list of compounds, we mainly based our comparison on their ability to score and prioritize successful annotations made. MZmine2 is the only tool which does not provide scoring of the suggested annotation, although it offers some parameters such as detection frequency and whether peaks are detected or estimated which can help prioritization. MS-DIAL uses a score as a cutoff, even though it is not displayed to the user. xMSannotator bases its scoring on m/z feature matching with different adducts/isotopes of a candidate, and in-set correlation between features. msPurity scores precursor purity to establish reliability of spectral matching for all features. Our in-house annotation tool displays four scores based on the three previously described predictors and global fit to pre-annotation. Scores from msPurity and xMSannotator can also be used for prioritization, although they offer mild visibility on the fit between feature and pre-

annotation. MS-DIAL offers an efficient form of ranking with an indication of whether the pre-annotation considers MS/MS or not, and allows a visualization of spectral matching. The individual score for each predictor accompanied by the global confidence index offered for the in-house tool allows a particularly efficient way to cutoff and prioritize pre-annotations. It is important to mention that some of these annotation tools offer specificities that could not be considered (e.g. biological correlations for xMSannotator) in the context of this study but that are definitively worth of interest.

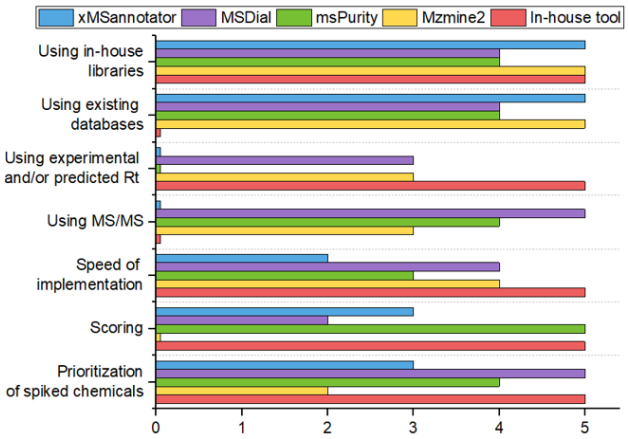

Figure IV.7 - Comparison of five suspect screening tools: xMSannotator (blue), MS-DIAL (purple), msPurity (green), MZmine2 (yellow) and in-house tool (red). Comparison was made on use of in-house databases, use of predicted or experimental Rt and MS/MS, speed of implementation, scoring and prioritization. Details are available in SI Fig.B6.

This suspect screening tool was then used on data generated from the four non-spiked plasma and serum samples to evaluate its applicability and relevance when investigating the internal chemical exposome. MS/MS data was used to manually confirm pre-annotations according to Schymanski et al.[44] Both over-the-counter medication such as ibuprofen (level 1, both matrices) or paracetamol (level 1, plasma), and prescription drugs such as the diuretic medication hydrochlorothiazide (level 2a, serum) were annotated. Markers indicative of lifestyle were confirmed in plasma, such as nicotine metabolites cotinine (level 1) and 3-hydroxycotinine (level 2a), or tetrahydrocannabinol (level 2a) and cannabidiol (level 2a). Other exposition markers were annotated, such as plasticizer bisphenol F (level 2a, serum), mono(2-ethylhexyl) phthalate (level 2a, plasma), organophosphate flame-retardant tris(1-chloro-2-propyl) phosphate (level 2a, serum) or antifungals ethyl- and butyl- paraben and metabolite 4-hydroxybenzoic acid (level 2a, serum). Dietary biomarkers were found in both plasma and serum, such as α-tocopherol (level 2a) or caffeine (level 1) and its three metabolites

paraxanthine, theobromine and theophylline (level 2a). The calorie-free sweetener acesulfame (level 1), was also found in both matrices.

6. Conclusion

HRMS-based methods have a great potential to help characterizing the human internal exposome. We demonstrated here that adjustments of the metabolomics workflow is nevertheless required for exposomics applications to detect low-abundant xenobiotics. Optimization of specific criteria for open source and vendor software can decrease dramatically the false negative rate. Nevertheless, this false negative rate can still reach up to 29% for some software, highlighting the need for further improvements. Besides detection frequency, automatic suspect screening workflow could help to speed up the annotation of the internal chemical exposome as this approach relies on suspect lists that can be prioritized. We report here an innovative workflow that incorporates for the first time several Rt prediction models. We also provide a comparison of several recent annotation tools that use specific different analytical criteria for the annotation process. One of the main advantages of this in-house suspect screening tool lie in the development of individual scores for each predictor accompanied by the global confidence index allowing a particularly efficient way to cutoff and prioritize pre-annotations.

7. Associated content

7.1. Supporting information

- Standard compounds supplier, hardware specifications, RTI calibrant sets, results of data processing and annotation (Excel)

- Chemicals and reagents, data acquisition, quality controls, modelling of retention time, optimization of individual data processing tools, and CI distribution for spiking compounds (Word).

8. Author information

Corresponding Author

*email: arthur.david@ehesp.fr

Author Contributions

9. Acknowledgements

This research was supported by a research chair of excellence (2016-52/IdeX Université of Sorbonne Paris Cité) awarded to AD and a grant from the Brittany council (SAD). JC was funded by the Réseau Doctoral en Santé Publique.

10. References

1. Vermeulen R., Schymanski E.L.*, et al.*, The exposome and health: Where chemistry meets biology. *Science* **2020,** *367* (6476), 392-6.
2. David A., Abdul-Sada A.*, et al.*, A new approach for plasma (xeno)metabolomics based on solid-phase extraction and nanoflow liquid chromatography-nanoelectrospray ionisation mass spectrometry. *J Chromatogr A* **2014,** *1365*, 72-85.
3. David A., Lange A.*, et al.*, Disruption of the Prostaglandin Metabolome and Characterization of the Pharmaceutical Exposome in Fish Exposed to Wastewater Treatment Works Effluent As Revealed by Nanoflow-Nanospray Mass Spectrometry-Based Metabolomics. *Environ Sci Technol* **2017,** *51* (1), 616-24.
4. Wild C.P., The exposome: from concept to utility. *Int J Epidemiol* **2012,** *41* (1), 24-32.
5. Rappaport S.M., Implications of the exposome for exposure science. *J Expo Sci Environ Epidemiol* **2011,** *21* (1), 5-9.
6. Chetwynd A.J.,David A., A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. *Talanta* **2018,** *182*, 380-90.
7. Chetwynd A.J., David A.*, et al.*, Evaluation of analytical performance and reliability of direct nanoLC-nanoESI-high resolution mass spectrometry for profiling the (xeno)metabolome. *J Mass Spectrom* **2014,** *49* (10), 1063-9.
8. Rappaport S.M., Barupal D.K.*, et al.*, The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* **2014,** *122* (8), 769-74.
9. Li Z., Lu Y.*, et al.*, Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Anal Chim Acta* **2018,** *1029*, 50-7.
10. Myers O.D., Sumner S.J.*, et al.*, Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal Chem* **2017,** *89* (17), 8689-95.
11. Libiseller G., Dvorzak M.*, et al.*, IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **2015,** *16*, 118.
12. McLean C.,Kujawinski E.B., AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. *Anal Chem* **2020,** *92* (8), 5724-32.
13. Alboniga O.E., Gonzalez O.*, et al.*, Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* **2020,** *16* (1), 14.
14. Uppal K., Walker D.I.*, et al.*, Computational Metabolomics: A Framework for the Million Metabolome. *Chem Res Toxicol* **2016,** *29* (12), 1956-75.
15. Kuhl C., Tautenhahn R.*, et al.*, CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* **2012,** *84* (1), 283-9.
16. Silva R.R., Jourdan F.*, et al.*, ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics* **2014,** *30* (9), 1336-7.
17. Daly R., Rogers S.*, et al.*, MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **2014,** *30* (19), 2764-71.

18. Ernst M., Kang K.B., *et al.*, MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* **2019,** *9* (7).
19. Uppal K., Walker D.I., *et al.*, xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal Chem* **2017,** *89* (2), 1063-7.
20. Tsugawa H., Cajka T., *et al.*, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.
21. Lawson T.N., Weber R.J., *et al.*, msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Anal Chem* **2017,** *89* (4), 2432-9.
22. Yin Y., Wang R., *et al.*, DecoMetDIA: Deconvolution of Multiplexed MS/MS Spectra for Metabolite Identification in SWATH-MS-Based Untargeted Metabolomics. *Anal Chem* **2019,** *91* (18), 11897-904.
23. Wang A., Gerona R.R., *et al.*, A Suspect Screening Method for Characterizing Multiple Chemical Exposures among a Demographically Diverse Population of Pregnant Women in San Francisco. *Environ Health Perspect* **2018,** *126* (7), 077009.
24. Grashow R., Bessonneau V., *et al.*, Integrating Exposure Knowledge and Serum Suspect Screening as a New Approach to Biomonitoring: An Application in Firefighters and Office Workers. *Environ Sci Technol* **2020,** *54* (7), 4344-55.
25. Moschet C., Piazzoli A., *et al.*, Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. *Anal Chem* **2013,** *85* (21), 10312-20.
26. Pourchet M., Debrauwer L., *et al.*, Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. *Environ Int* **2020,** *139*, 105545.
27. Wishart D.S., Feunang Y.D., *et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **2018,** *46* (D1), D608-D17.
28. Kanehisa M., The KEGG database. *Novartis Found Symp* **2002,** *247*, 91-101; discussion -3, 19-28, 244-52.
29. Pence H.E.,Williams A., ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **2010,** *87* (11), 1123-4.
30. Pluskal T., Castillo S., *et al.*, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**.
31. Smith C.A., Want E.J., *et al.*, XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006,** (78), 779-87.
32. Adusumilli R.,Mallick P., Data Conversion with ProteoWizard msConvert. *Methods Mol Biol* **2017,** *1550*, 339-68.
33. Want E.J., Wilson I.D., *et al.*, Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* **2010,** *5* (6), 1005-18.
34. Alves G., Ogurtsov A.Y., *et al.*, Molecular Isotopic Distribution Analysis (MIDAs) with Adjustable Mass Accuracy. *Journal of The American Society for Mass Spectrometry* **2013,** *25* (1), 57-70.
35. Kim S., Chen J., *et al.*, PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **2019,** *47* (D1), D1102-D9.
36. Aalizadeh R., Thomaidis N.S., *et al.*, Quantitative Structure-Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. *J Chem Inf Model* **2016,** *56* (7), 1384-98.
37. Bonini P., Kind T., *et al.*, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020,** *92* (11), 7515-22.
38. Stanstrup J., Neumann S., *et al.*, PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. *Anal Chem* **2015,** *87* (18), 9421-8.
39. Wishart D.S., Feunang Y.D., *et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **2018,** *46* (D1), D608-d17.

40. Barupal D.K.,Fiehn O., Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. *Environ Health Perspect* **2019,** *127* (9), 97008.
41. Haug K., Cochrane K.*, et al.*, MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* **2020,** *48* (D1), D440-D4.
42. McEachran A.D., Mansouri K.*, et al.*, A comparison of three liquid chromatography (LC) retention time prediction models. *Talanta* **2018,** *182*, 371-9.
43. Bade R., Bijlsma L.*, et al.*, Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. *Talanta* **2015,** *139*, 143-9.
44. Schymanski E.L., Jeon J.*, et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.

Chapter V. Implementing a large-scale suspect screening approach to characterize the human chemical exposome

In the recent years, the growing interest in investigating the links between environmental exposures and health has led to the development of new methodologies to study the exposome[1-3]. Following the technological advancements based on HRMS, the rise of non-targeted approaches, in particular, hold great promises to expand knowledge on the human chemical exposome[2, 4]. However, these approaches require the optimization of each step of the workflow (notably sample preparation, data processing, annotation) to achieve the sensitivity and robustness ideally needed to limit biases in the visible chemical space[5-8]. The previous chapters presented the optimization steps undertaken to improve the efficiency of the aforementioned steps. Briefly, a dual sample preparation process involving PPT and the Phree PLR plate was recommended based on the complementarity of the image of the chemical exposome they provide. Regarding data processing, several software tools (including both vendor and open source tools) were optimized to detect low-abundant chemicals in blood-derived matrices, and correctly optimized vendor software was found to adequately perform this task with low implementation times. Lastly, an annotation tool adapted to exposomics application was developed. MS1 chemical predictors were chosen and optimized to compare suspects and features, and significantly lower the rate of false positive annotations. These developments allowed constructing a workflow suited to detect low-abundant compounds in plasma and serum samples.

While the presented optimizations allow achieving an adequate sensitivity performance, the workflow's robustness must still be evaluated. To this end, the workflow may be implemented at a larger scale, i.e. move beyond the scale of one batch. Large-scale applications come with specific challenges, mainly revolving around insufficient system stability over the course of multiple batches, sometimes injected over several weeks or months[9, 10]. This may be translated by a low repeatability in Rt, and/or in signal, leading to poor comparability between samples.

The optimized workflow was applied to analyze blood serum samples from 125 12-year-old boys issued from the Breton mother-child cohort Pélagie. Given the large amount of data collected and generated for this cohort[11-13], it will offer a rare opportunity to study the links between the chemical exposome and health. The first step in establishing these links is to accurately describe the chemical exposome of this population through the non-targeted analysis of 125 serum samples.

In this chapter, 125 serum samples were analyzed and processed using the non-targeted optimized workflow developed in the context of this PhD work. Quality control criteria based on feature area and Rt repeatability in QC samples and internal standards were established to ensure comparability of the samples. The processed data was then annotated assisted by the in-house tool and MS-DIAL[14].

Hence, the objectives of this chapter were:

i. To study the robustness of the analytical and bio-informatics workflow implemented during this PhD.
ii. To study the relevance of using the in-house software through the comparison of MS1 and MS2 predictors' accuracy (MS-DIAL) for the annotated compounds.
iii. To characterize chemical exposures in Pélagie through the categorization of annotated compounds. Exposure profiles combining various chemicals of interest were described.
iv. To study the complementarity of the two used SPM at larger scale (as described in Chapter III).

1. Outgrowing the scale of a batch: quality control

The 125 samples were prepared with two SPM using the pipeline based on PPT and Phree, as proposed in Chapter III (Figure V.1). For each SPM, the 125 samples were separated in five 25-sample batches and injected to acquire data in both ESI modes (i.e. 500 injections in total for the samples), and 20% of randomly selected samples were re-injected for MS2 acquisitions in both ESI modes. In total, 20 batches were to be injected (n= 960 injections in total including QCs and MS2). However, due to technical difficulties mainly revolving around the instability of the LC, only the first three batches of Phree samples (as opposed to 5 for PPT) were further processed, i.e. 75 samples. The comparison of the two SPMs (robustness, annotation) were then only performed on the first three batches.

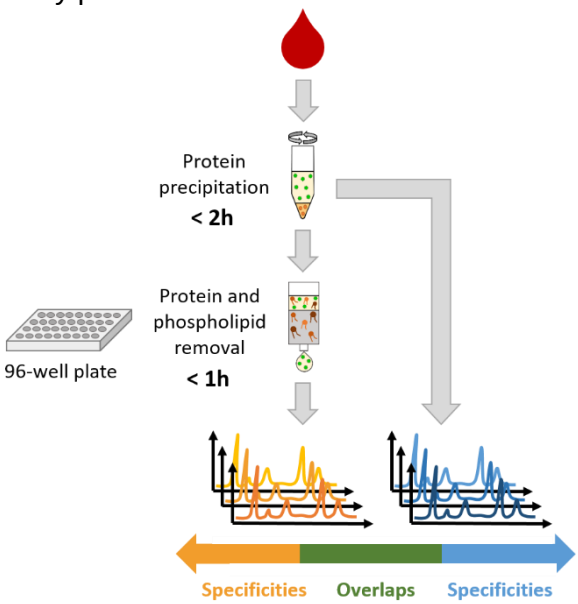

Figure V.1 – Schematized representation of a dual sample preparation process, where half of the supernatant from protein precipitation is injected as is (after reconstitution), and the other half is used for further protein and phospholipid removal before injection on the UPLC-ESI-QTOF. In total, 960 samples were injected including QCs and MS2 acquisitions.

Sample analysis was performed over the course of 9 weeks (7.5 weeks of non-stop analysis, and 1.5 weeks of cumulated preventive and curative maintenances between batches). The same composite QC sample was injected throughout all batches for interbatch correction. One large composite QC sample by SPM was prepared (800 µL per SPM), and was injected 11 times per batch (first 5 injections for system equilibration, last 6 to assess analytical drift). To ensure the comparability of data acquired over this extended period, quality control was performed on the injected batches on select analytes (i.e. spiked internal standards) and at the non-targeted scale. This quality control step was performed at the targeted level, using internal standards spiked in all samples (including QC samples), and at the non-targeted level on all features obtained from QC samples. Results of the quality control process are presented in Figure V.2.

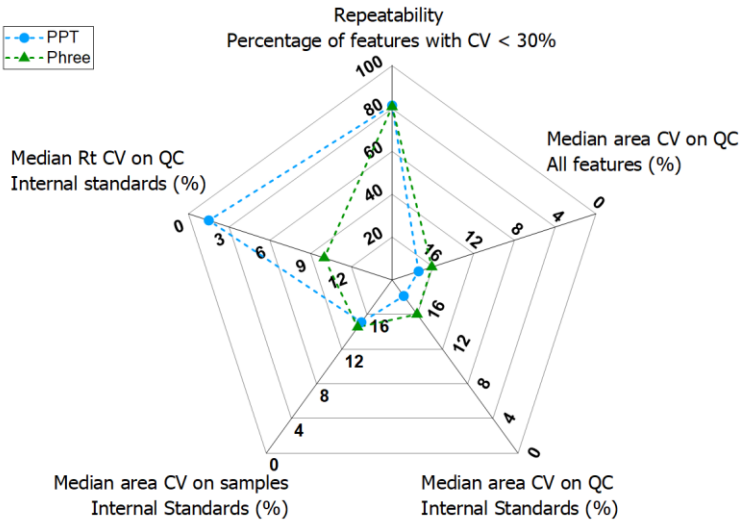

Figure V.2 – Quality control parameters for the application of two sample preparation methods to cohort samples (n=75 samples) before correction. Outer edges identify best performances.

Firstly, the repeatability of the analytical sensitivity of the QTOF was evaluated on all batches at the non-targeted scale (i.e. on all features of all injected QCs) using the criteria proposed by Want et al. (2010)[15], by verifying that over 80% of QC features common to at least 80% of QC samples (i.e. 5 out of the 6 QCs injected between samples at the batch level) presented area CV values under 30%. This parameter was assessed at 82% and 83% for Phree and PPT, respectively, which indicates a satisfactory repeatability in both cases. Furthermore, the median area CV was computed on all batches for all features, for internal standards in QC samples, and for internal standards in all samples. Median are CV values at the non-targeted scale were of 16.1% and 17.4% in samples prepared by Phree and PPT respectively. When focusing on internal standards, median area CV were of 16.0% and 14.5% in QC samples and in all samples when prepared by Phree, and of 18.2% and 15.1% in QC samples and in all

samples when prepared by PPT. These values, while all under the 20% threshold, were indicative of a general tendency for batch-dependent variability, especially towards the last batches. This was remedied through the implementation of a total ion current normalization on all samples (including QC samples), i.e. a division of each feature's area by the sample's total ion current. This normalization was chosen for its already demonstrated efficiency in other omics approaches[16], and was performed on all SPM×ESI mode combination (i.e. PPT in ESI (−) mode, PPT in ESI (+) mode, Phree in ESI (−) mode, Phree in ESI (+) mode). The results of this normalization on the mean feature area in the case of PPT samples injected in ESI (+) mode is illustrated in Figure V.3. Results for other SPM×ESI mode combinations are available in Appendix 4.1.

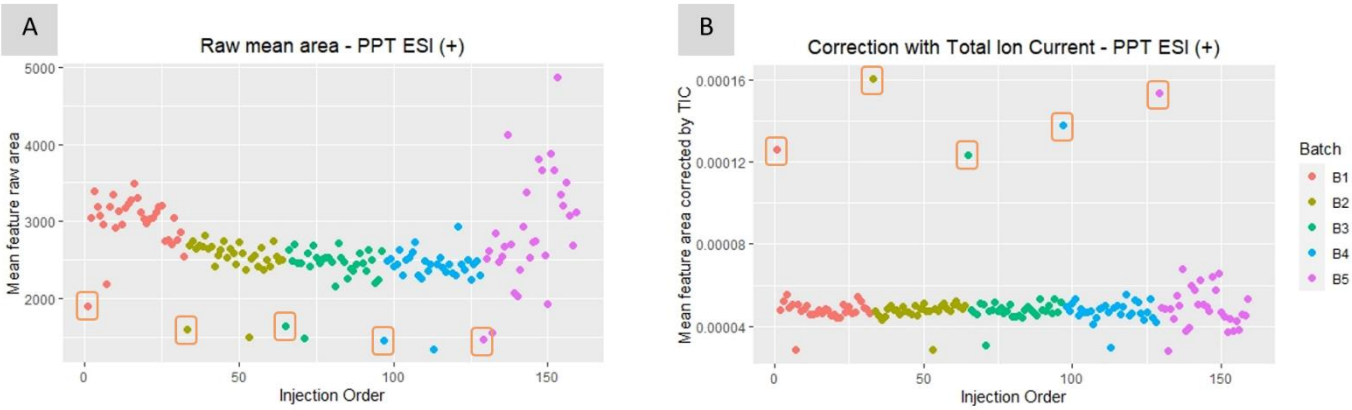

Figure V.3 – Mean feature raw area (A) and mean feature area after total ion current correction (B), shown on samples (including the composite quality control samples) prepared by protein precipitation (PPT) injected in ESI (+) mode on the UHPLC-ESI-QTOF. Blank samples for each batch are identified by orange squares.

Besides analytical sensitivity, the Rt CV on internal standards in QC samples was computed and determined to be satisfactory, i.e. under the 10% threshold for both SPM on all batches (1.5% and 9.6% in PPT and Phree samples respectively). After this normalization, the mean feature area is comparable in samples across all batches. This normalization step was performed to the analytical variations between batches. The effect of normalization on a large scale was verified by performing a PCA before and after normalization. Results on PPT samples injected in ESI (+) mode are presented in Figure V.4. Results for other SPM×ESI mode combinations are available in Appendix 4.1. As expected, the normalization step allowed reducing the dispersion of samples initially observed in batch 5 (and at a lesser scale in batch 1) in this case.

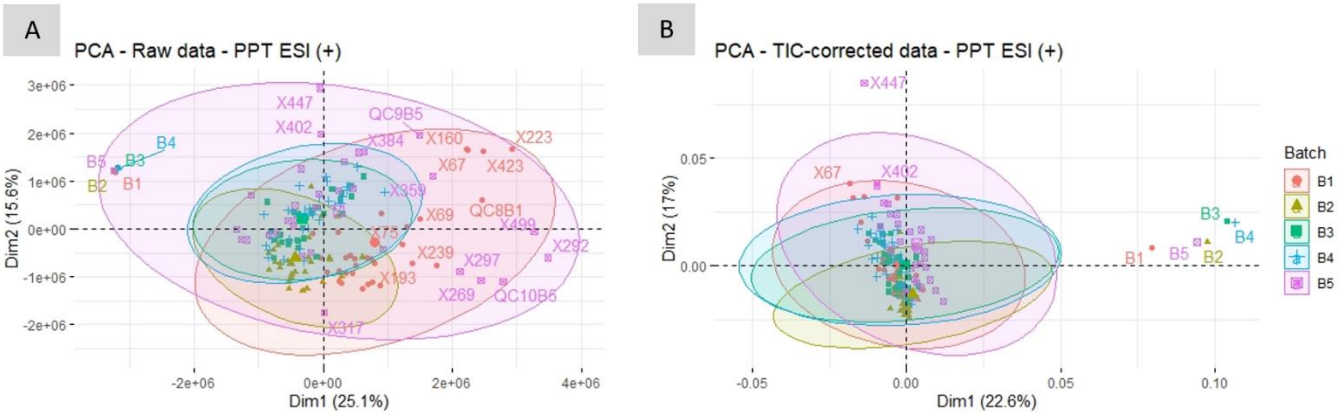

Figure V.4 – PCA using raw area (A) and PCA using area after total ion current correction (B), shown on samples prepared by protein precipitation (PPT) injected in ESI (+) mode on the UHPLC-ESI-QTOF.

Quality control parameters were computed again using normalized data. Results are presented in Figure V.5. The normalization step resulted in a decrease of median area CV values at the non-targeted scale 35% for both SPM (CV values of 10.5% and 11.2% for Phree and PPT respectively). Similarly, when focusing on internal standards in QC samples and in all samples for both SPM, a 31-48% decrease in median area CV was observed. This significant reduction in area variability underlines the relevance of the total ion current normalization. Overall, median area CV values were always under 12%, which is a satisfactory value in regards to the scale of this application. It should also be noted that there is no observed difference in median area CV between both SPM, thus confirming their equal adequacy for large-scale applications.

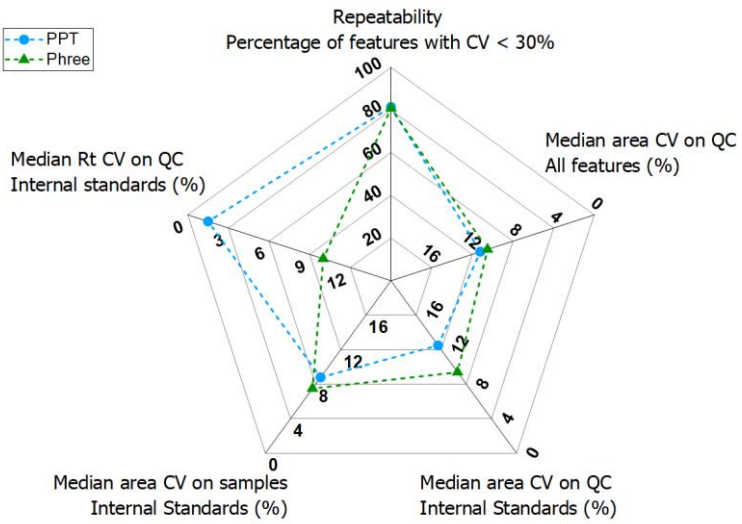

Figure V.5 – Quality control parameters for the application of two sample preparation methods to cohort samples (n=75 samples) after correction. Outer edges identify best performances.

2. Implementing a suspect screening approach at a large scale

3.1. Comparing the use of MS1 and MS2 predictors for annotation in an exposomics context

Two suspect screening approaches were implemented to perform the annotations. Firstly, raw data obtained from the chemical analysis was processed using the optimized MarkerView data processing tool as described in Chapter IV, paragraph 5.1.3. The resulting feature list was then processed by the in-house tool, resulting in pre-annotations prioritized using MS1 predictors. Secondly, raw MS2 IDA data was processed using MS-DIAL's All public spectral database, resulting in annotations prioritized through an MS2 matching. Manual curation was performed on results from both tools. In the case of the in-house tool, MS2 spectra for suggested pre-annotations were compared to reference spectra, isotopic patterns were verified, and plausibility was checked. Reference spectra could be spectra acquired in-house (highest confidence), obtained from shared online databases such as MassBank[17] (high confidence), or obtained from in-silico prediction tools such as CFM-ID[18] or MetFrag[19] (medium confidence). In the case of MS-DIAL, the visual representation of the matching feature and reference MS2 spectra (from online MS2 spectra database) was checked, along with isotopic patterns and plausibility. Results from the manual curation are available in Appendices 4.2 and 4.3. Generating pre-annotation data was faster with the in-house tool while manual curation was overall faster using MS-DIAL, as spectral data is made available to the user. Table V.1 provides an overview of the data generated by both tools and the results of manual curation. Annotated compounds are available in Appendix 4.2, and MS2 data is available in Appendix 4.3.

	In-house software	**MS-DIAL**
Number of suspects	5,898 (ESI +) 5,898 (ESI −)	13,303 (ESI +) 12,879 (ESI −)
Median number of unique suggested annotations	2,422	418
Median number of total suggested annotations	8,354 (raw) 1,928 (global CI > 0.70)	730
Cumulated number of confirmed annotations	81	68
Estimated manual curation time (effective days)	~ 35	~ 15

Table V.1 – Overview of the data generated by two suspect screening tools based on either MS1 or MS2 predictors (In-house software and MS-DIAL respectively). Median and cumulated values are determined based on the four sample preparation method × ionization mode possible combinations (i.e. protein precipitation and Phree phospholipid removal plate in positive and negative ionization modes).

As expected, the in-house tool generated more suggested annotations. This can be explained by the joint effect of the high selectivity of fragmentation patterns, leading to more elimination of false positive suggested annotations, and the fact that only a limited number of compounds were fragmented during the MS2 analysis, which may lead to some false negatives (i.e. compounds present in the sample, detected during the analytical step, but not annotated). Moreover, the total number of annotations suggested by the in-house software could be reduced using threshold values on the implemented confidence indices (CI). A cutoff value of 0.70 was choosen based on previous observations to reduce this number by 74-82%. This also allows prioritizing features that deserved more attention for manual curation. On the other hand, the MS1 annotations suggested by MS-DIAL (without MS2 data) average at around 11,000 per SPM×ESI mode combination, and can hardly be further prioritized due to the lack of additional reliable information such as scoring.

Manual curation allowed to confirm the annotation of 92 compounds with a level of 4 (with global CI ≥ 0.85) or better according to Schymanski et al. (2014)[20], with the overlap of 57 compounds between the two suspect screening tools. MS-DIAL did not suggest 24 of the total annotated compounds. Firstly, three compounds (i.e. 4-chlorophenol, pentachlorophenol and triclosan glucuronide) were only prioritized by the in-house software since no associated MS2 data was acquired, and attributed a level 4. However, this level does not accurately reflect the confidence that can be put in these annotations. Indeed, it does not account for the verification of the very particular isotopic patterns linked to the presence of one, five and three chlorine atoms respectively, which is a highly discriminating characteristic when looking at the M2/M0 ratio. Moreover, predicted Rt values strongly support these annotations. The current confidence level system also does not account for the annotation of another metabolite of triclosan (i.e. triclosan sulfate, level 1). A visualization of MS1 evidence supporting the pentachlorophenol and triclosan glucuronide annotations is presented in Figure V.6.

Thirteen additional compounds were attributed a 2b level since there is no available MS2 experimental (in-house or from shared databases) reference spectra for these structures (e.g. 1,3,5-tris(2,2-dimethylpropionylamino)benzene or propylparaben sulfate), requiring the use of a fragmentation prediction model such as MetFrag[19] or CFM-ID[18, 19]. The remaining eight compounds not annotated by MS-DIAL were not listed in their database and were confirmed with analytical standards available in-house (e.g. triclosan sulfate, acetaminophen glucuronide, etc.). On the other hand, eleven compounds were not prioritized by the in-house software because they were not in the used suspect list (e.g. 10,11-dihydroxy-10,11-dihydrocarbamazepine, auraptene, lenticin, etc.). This underlines the need for sustaining the data collection effort in the community to continue expanding suspect lists with relevant compounds.

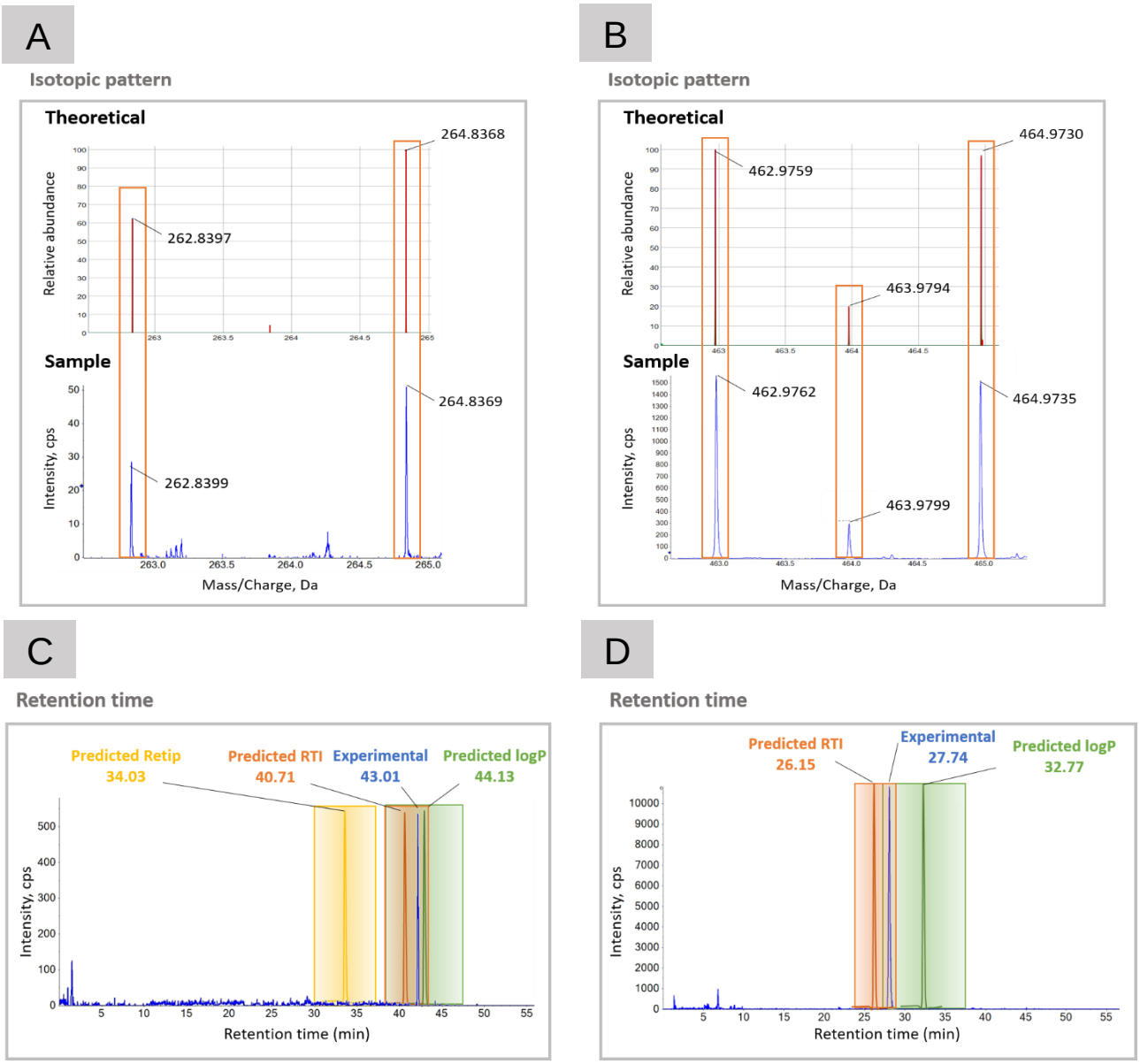

Figure V.6 – MS1 predictors supporting the pentachlorophenol (A- isotopic pattern, C- retention time) and the triclosan glucuronide (B- isotopic pattern, D- retention time) annotations. Theoretical and experimental isotopic patterns are compared based on coherence between mass/charge ratios and isotopic area ratios. Experimental retention times are compared to values predicted using RTI (orange), Retip (yellow) or a polarity-based linear regression (logP) (green) and their respective confidence intervals (represented by the color gradients). This data was acquired in negative ionization mode on the UHPLC-ESI-QTOF.

Overall, the use of MS2 predictors is extremely powerful but can encounter some critical obstacles in exposomics applications, notably the lack of MS2 acquisition for the feature of interest, and the lack of reference spectra for the hundreds of thousands possible suspects[21]. When these issues arise, it is crucial to have other predictors based on MS1 to efficiently prioritize the massive number of suggested annotations for manual curation. Predictors such

as predicted Rt values and isotopic pattern, associated to confidence indices scoring the proximity between suspect and feature, allow avoiding false negatives and prioritizing features of interest. For instance, the pentachlorophenol and triclosan glucuronide annotations, both standing at a level 4, cumulate distinctive isotopic pattern (confirmed with comparison of therotical M2/M0) and coherence with multiple predicted retention times, as shown in Figure V.6. This underlines the lack of accounting for some important discriminating predictors in the current confidence level system. In these particular cases, intermediary levels could be considered to distinguish between compounds with no MS2 data but different amounts of MS1 evidence. Other parameters not used with this analytical system, such as the collision cross section (CCS) used in ion mobility systems, may also be considered in this updated classification, as presented in Figure V.7. In this case, using the in-house software was critical in significantly expanding the number of annotated compounds (+26%).The confidence levels attributed to annotated compounds using both classifications are available in Appendix 4.4.

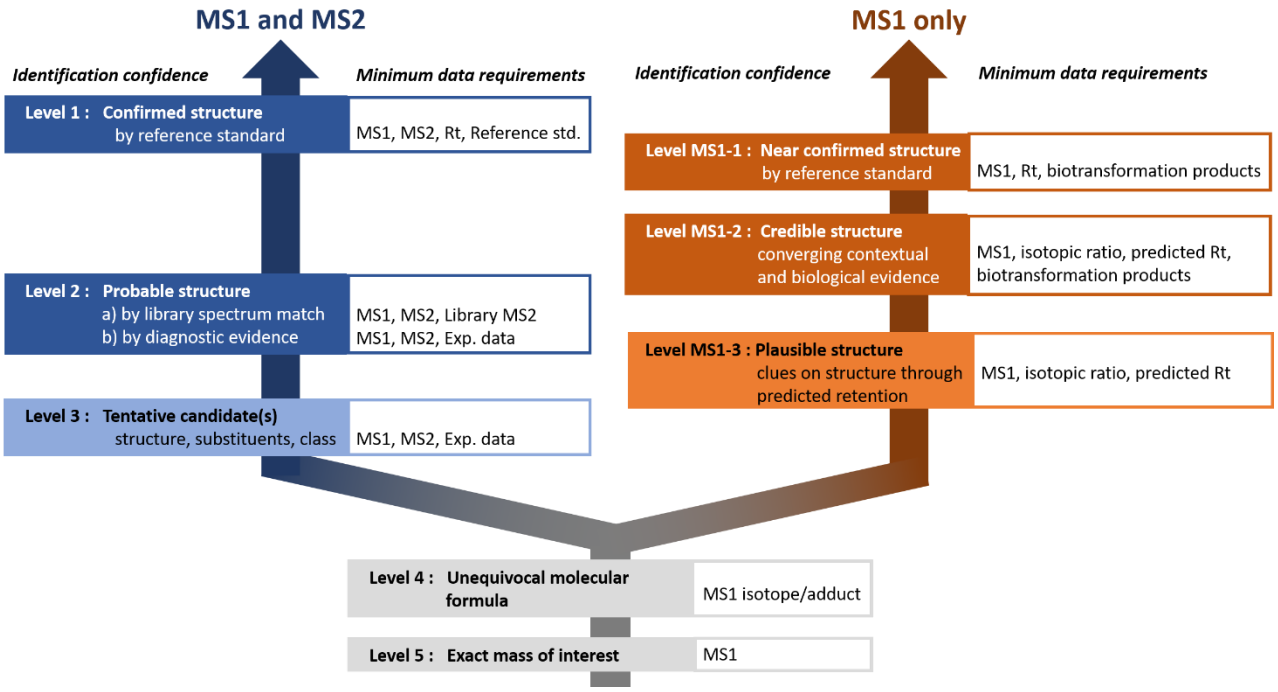

Figure V.7 – Updated identification confidence levels accounting for new methodological tools, such as prediction models for retention time (Rt) and biotransformation products. MS2 refers to any form of fragmentation.

3.2. Describing the environmental chemical exposures in the Pélagie cohort

The data collected on PPT samples and Phree samples injected in both ESI (−) and ESI (+) modes allowed annotating 92 compounds from the internal chemical exposome with a level of 4 or higher according to Schymanski et al. (2014)[20] (level MS1-3 or higher according to the suggested updated classification). Exposure to most of these compounds can occur through multiple sources (e.g. 2-hydroxybenzoic acid, or salicylic acid, primarily used as a preservative in industrial foods, but that can also be used as a medication or as a synthesis intermediate). A non-exhaustive classification of sources for annotated compounds is available in Appendix 4.5. However, for illustrating purposes, only primary uses were considered in the following descriptions. The repartition of the 92 annotated compounds by primary use is presented in Appendix 4.5. The repartition of primary uses is presented in Figure V.8.

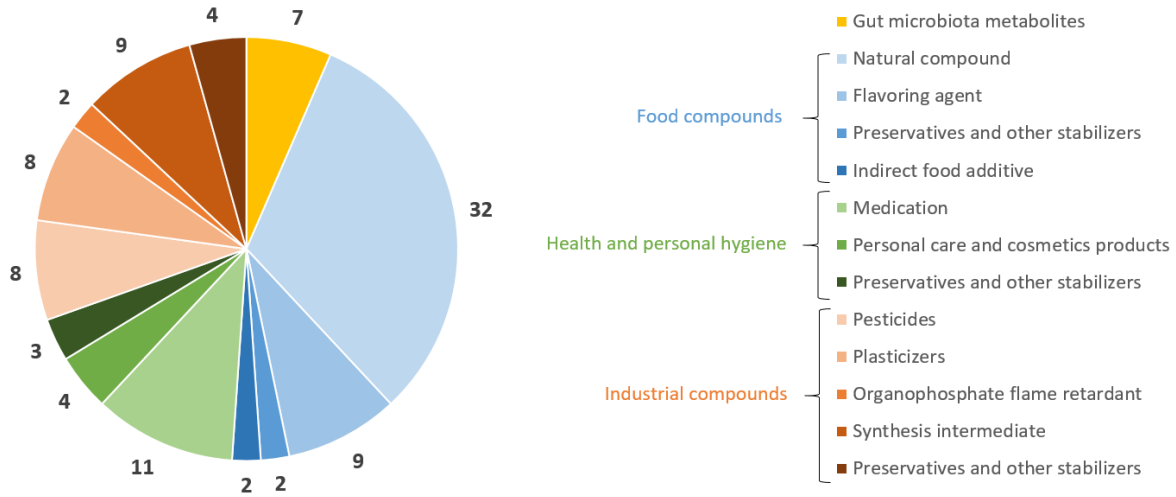

Figure V.8 – Classification of the major source of annotated compounds (n=92), expressed in percentages. Gut microbiota metabolites are shown in yellow, compounds obtained from food in blues, compounds obtained from health and personal hygiene products in greens, and industrial compounds in oranges.

Four main categories were identified: gut microbiota metabolites, compounds originating from food, compounds used for health and hygiene purposes (e.g. pain management, antiepileptic medication, surfactants used in shower gels), and industrial compounds (e.g. synthesis intermediates used in the manufacturing of dyes, rubbers or pesticides). These categories represented respectively 7%, 45%, 18% and 30% of annotated compounds. Gut microbiota metabolites are included in the internal chemical exposome, as the microbiome operates as an interface between external exposures and the individual; therefore, gut microbiota

metabolites may reflect the external exposome and constitute their own category of substances from the human internal chemical exposome.

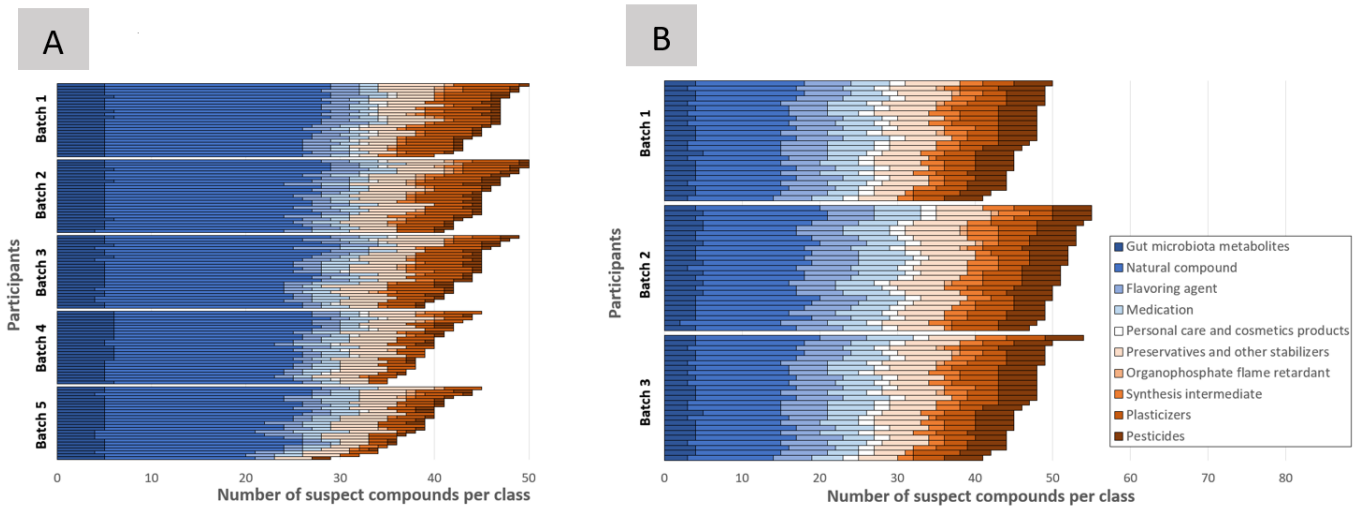

Figure V.9 – Detection of suspect compounds per class in each participant (separated by batch) in protein precipitated samples (A) and Phree samples (B). Preservatives and other stabilizers found in processed foods, health and personal hygiene products and industrial compounds were combined in a single category for clarity.

The highest contributing subcategory was natural compounds obtained from food and their metabolites (e.g. caffeine and paraxanthine, piperine, flavins, etc.) representing almost a third of annotated compounds. Representing a significant 11%, the medication subcategory (health and personal hygiene category) includes, for instance, non-steroidal anti-inflammatory ibuprofen, as well as antiepileptic carbamazepine and metabolites 10,11-dihydroxy-10,11-dihydrocarbamazepine and 2-hydroxycarbamazepine. Food compounds and pharmaceutical products (i.e. medication) represent more than half of annotated compounds (56%). This was expected, as they can be concentrated up to 10^6 times more than some industrial pollutants (e.g. pesticides) in blood[2]. The pesticides subcategory, contributing 9% of all annotated compounds, includes parent compounds such as bromoxynil or tritosulfuron, and metabolites such as chlorothalonil metabolite 4-hydroxy-2,5,6-trichloroisophthalonitrile and bromoxynil metabolite 3,5-dibromo-4-hydroxybenzoic acid. Usual suspects were annotated in the plasticizer subcategory (8%) such as phthalates and perfluoroalkyl substances[4, 22]. The detection of annotated compounds was assessed in each sample. A representation of the detected compounds in each participant is presented in Figure V.9. Proportionately, the most represented chemical class is natural food compounds (49% of annotated compounds in PPT samples, 27% in Phree samples), and the least represented is organophosphate flame retardants (0.6% and 0.4%).

For both SPM, gut microbiota metabolites and natural food compounds showed little variation in proportions among participants. CV were computed on proportions of these two classes were under 15% for both SPM (14% and 7% respectively in PPT samples, and 14% and 8% in Phree samples). The highest inter-individual variability was observed for the proportions of organophosphate flame retardants (CV values of 165% and 210% in PPT and Phree samples respectively), synthesis intermediates (CV values of 115% and 27% in PPT and Phree samples respectively) and pesticides and their metabolites (CV values of 65% and 9% in PPT and Phree samples respectively). This is coherent with the fact that most individuals would be exposed to ubiquitous food compounds and well-known gut microbiota metabolites, but their exposure to industrial compounds are more susceptible to vary depending on their lifestyle (e.g. living in an urban or rural area, dietary habits, etc.). It should be noted that some combinations of compounds may be indicative of a given individual's lifestyle. For instance, as the case of 48 participants, co-exposure to pesticides ioxynil, bromoxynil and transformation product 3,5-dibromo-4-hydroxybenzoic acid, which are mostly used in agriculture, may indicate living in a rural area. Similarly, co-exposure to artificial sweeteners acesulfame, aspartame and sucralose, as is the case for 12 participants, may be an indication of a more industrial processed diet. However, given that the number of annotated compounds is high compared to the number of participants, establishing such profiles is challenging in terms of statistical power. Moreover, finding the determinants of those exposures would require additional indirect measurements (i.e. analyzing environmental samples) and/or the use of questionnaires.

Detection frequencies for all compounds and both SPM were computed. Detailed results are available in Appendix 4.4. Overall, out of 92 annotated compounds, 54 have a detection frequency over 80% in either or both SPM. As shown in Appendix 4.6, almost 15% of those ubiquitous compounds (i.e. 8 compounds) are not documented in the NORMAN Network's extensive SUSDAT list, which combines more than 111,000 structures from 94 community-shared suspect lists[21]. These compounds include 3 phase I and II metabolites (hydroxylated and sulfated forms), which highlights the need to include known or predicted metabolites in suspect lists. It should also be noted that other metabolization pathways should be taken into account when predicting metabolite structures, as they could allow integrating a temporal aspect to the exposure evaluation[23]. Moreover, 10 compounds have a detection frequency over 80% with at least one SPM, and do not have any available toxicological data according to the CompTox dashboard[24]. One of those compounds (Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate, found in 86% of PPT samples) is a phthalate, some of which are classified as endocrine and metabolic disruptors[25]. This underlines the potential of suspect screening approaches to uncover previously poorly documented exposures to chemical compounds of concern.

Pesticide and endocrine disruptor bromoxynil, detected in 61% of samples, had previously been reported in the urine of 22% of pregnant women from this cohort[26]. This may suggest a repeated or chronic exposure to this compound for some individuals of this cohort. Moreover, previously reported levels of bromoxynil in plasma samples from rural teenage residents varied from trace levels to 140 ng/mL[27]. Similarly, pesticide metabolite fipronil sulfone, detected in 29% of samples, was previously reported in human blood (general population) at concentration comprised between 0.1 and 4 ng/mL[28]. These documented low levels are a preliminary indication that the implemented workflow presents adequate sensitivity performances, although targeted assays should be performed on the investigated samples to confirm bromoxynil levels.

Previous studies on the Pélagie cohort did not investigate bromoxynil metabolite 3,5-dibromo-4-hydroxybenzoic acid. However, this compound is detected in 97% of samples with higher area values (factor 3-8 depending on sample). Further review of the literature indicated that this metabolite was not reported in HBM studies in blood or urine before. This underlines the potential of using suspect screening approaches to uncover new relevant biotransformation products to better evaluate human exposure to chemicals of concern. Although bromoxynil was banned in France in 2021, identifying this new biomarker of exposure may be useful for retrospective analysis, in the case of persistence in the environment, or in countries where it is not banned.

Overall, a set of compounds with very diverse physical-chemical characteristics (i.e. $-2.7 \leq \log P \leq 16$, and $100.0754 \leq [M+H]^+ \leq 811.4913$) was annotated in these samples. These compounds also include various chemical functions, and have diverse sources. The most inter-individual variability was observed on compounds usually referred to as pollutants, as opposed to food compounds and gut microbiota metabolites, which appears coherent. Lastly, the visible exposure profiles on PPT and Phree samples seem to present differences both in the proportions and variability of chemical classes, which raises a question regarding the relevance of using two SPM in light of the performed annotations.

3.3. Exploring the potential of dual sample preparation

The two SPM used to prepare the serum samples were compared according to the methodology described in Chapter III, paragraph 4.2.3. Briefly, area fold changes (FC) were computed between both SPM on compounds annotated in the first three batches (i.e. 89 compounds out of 92 annotated in total). Median fold changes are represented in Figure V.10.

Xenobiotics presenting FC values below 0.5 and over 2 (i.e. favored by one of the SPM) represented 94% of the total annotated compounds. This is coherent to the results from the

pilot study presented in Chapter III, for which this condition represented 93% of annotated compounds. Moreover, more than 74% of annotated compounds were only visible using one SPM, which confirms the critical need for orthogonal methods to widen the visible chemical space. This tendency was further explored at a larger scale by computing FC values on quality control samples. Results are presented in Table V.2.

Fold change (FC) values	**Features (%)**
0 (only in PPT)	43.0
$0 < FC \leq 0.5$	5.3
$0.5 < FC \leq 2$	7.2
$2 < FC < \infty$	7.3
∞ (only in Phree)	37.2

Table V.2 – Percentage of features of quality control samples injected in positive and negative ionization modes on the UHPLC-ESI-QTOF, categorized by fold change (FC) values (i.e. area ratio of features in Phree and protein precipitation).

At this scale, 80% of features are visible with only one SPM, and an additional 13% of features are favored by one SPM. Overall, FC values are oriented towards extreme values. This is coherent with what was observed in the serum samples in the pilot study. This was tentatively attributed in part to the observation of abundant and often multiple charged peptide peaks in serum samples prepared with PPT only. This observation was replicated in this assay, which supports this hypothesis.

There was no visible bias towards either SPM in terms of proportion of favored features, despite the fact that Phree samples were two times more concentrated than PPT samples. This might be explained by the fact that the sensitivity gain through the concentration factor in Phree samples is compensated by the higher selectivity of this SPM (i.e. loss of signal for phospholipids, etc.).

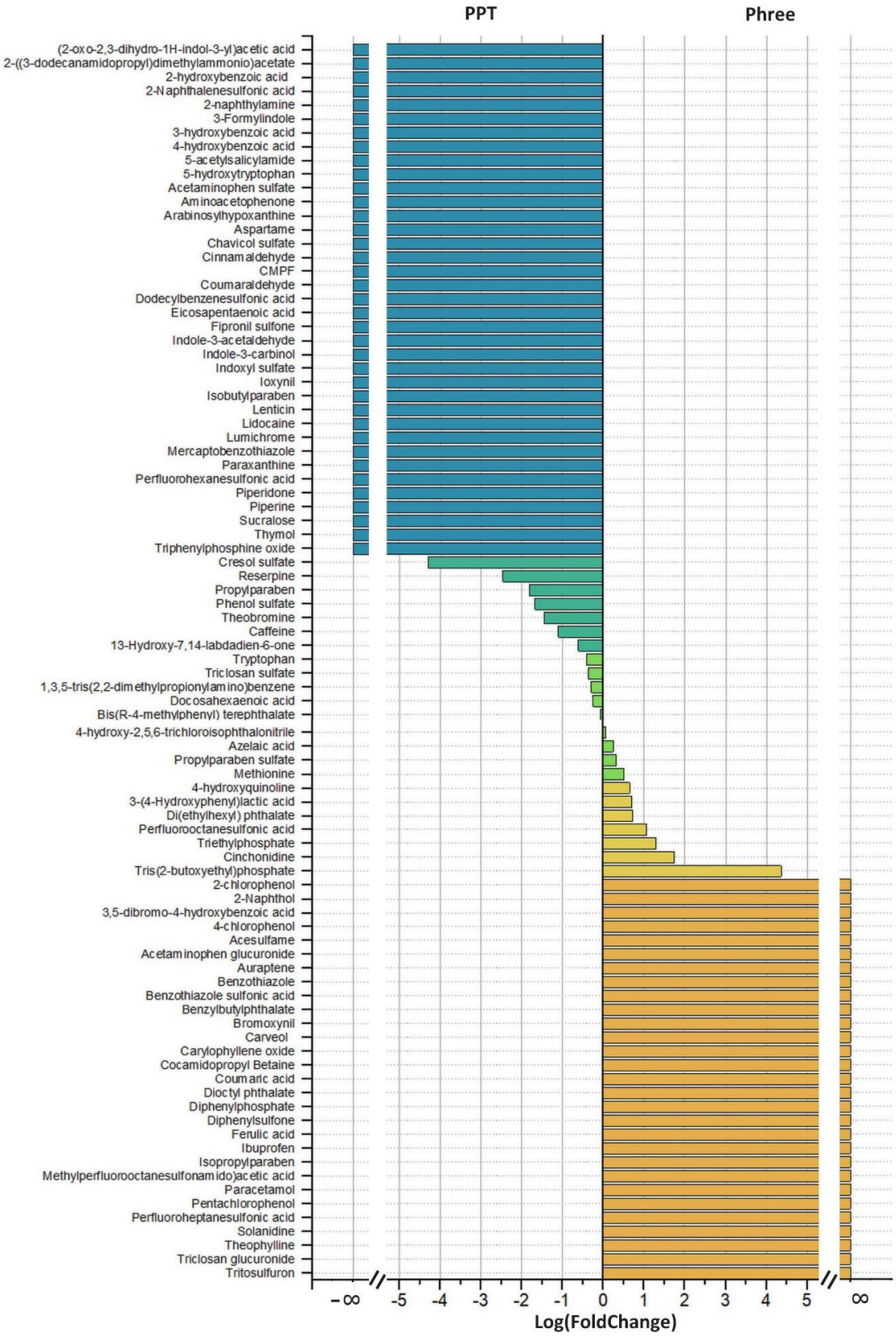

Figure V.10 -Comparison of annotated xenobiotics' areas in samples prepared with protein precipitation (PPT) and protein removal plate Phree in Pelagie serum samples. Logged values of fold changes (i.e. area ratio between Phree and PPT) are presented on the x-axis, where –∞ and +∞ values represent the absence of compounds in samples prepared with Phree and PPT, respectively. Bars on the left of the central vertical axis represent compounds presenting higher areas in PPT samples and vice-versa.

Given the relatively small number of annotations, identifying with certainty the driving factors of the enhanced detection of compounds with one SPM or the other is challenging. However, some tendencies were identified. Overall, polar compounds (i.e. low Rt) seemed favored by PPT. For instance, 2/3 of compounds with Rt values under 10 minutes had FC values under 1. This might be explained by the hypothesized capability of the Phree plate to retain highly polar compounds (i.e. polar heads of phospholipids)[29], thus leading to the lower detection of these compounds when using this SPM. On the contrary, 2/3 of compounds with Rt values over 40 minutes had FC values over 1 (i.e. favored by Phree). This was expected, since samples prepared by PPT presumably contain more phospholipids notably, which are usually detected between 40 and 45 minutes. Phenomena such as ion suppression may therefore explain why other compounds eluting at this time are proportionately less ionized, and therefore less detected.

Regarding compounds favored by PPT, gut microbiota metabolites seem to be more readily detectable when using this SPM, with 5 compounds out of 6 presenting a FC value lower than 0.021, i.e. detected more than 476 times more in samples prepared with PPT than with Phree, and the 3 indole derivatives out of those 5 were only detected with PPT. Similarly, phase II sulfate metabolites seem to be more detectable in PPT samples, with 6 out of 7 compounds (including 3 also classified at gut microbiota metabolites) presenting a FC value between 0.428 and 0 (i.e. over 2.3 times in PPT samples compared to Phree to only detected in PPT samples).

On the other hand, both organophosphate flame retardants are detected almost 20 times and 2300 times more in Phree samples compared to PPT samples (Triethylphosphate and Tris(2-butoxyethyl)phosphate respectively). The only other compound with a phosphate group (i.e. diphenylphosphate) was also favored by Phree (not detectable at all in PPT samples). This was rather unexpected, as PLR plates are hypothesized to retain phospholipids through a Lewis acid-base interaction between the stationary phase and the esterified phosphate group found in phospholipids[29]. However, it is possible that only highly polar phosphate groups such as those found in phospholipids are retained by the plate, since the considered compounds are mid-polar (logP values ranging from 0.8 to 2.8). Lastly, 3 out of 4 annotated phthalates are better detected in samples prepared by the Phree PLR plate (FC varying from 5.28 to +∞). The remaining phthalate is, more precisely, a terephthalate (i.e. substituents are in the para-position instead of the ortho- position), and is very mildly favored by PPT (FC = 0.86). Alongside the fact that the terephthalate substituents are larger than the substituents on the annotated phthalates, it could be hypothesized that sterically hindered phthalates are less likely to pass through the PLR plate, thus being less favored by this SPM compared to less hindered ones.

Overall, using a dual sample preparation process to prepare complex samples such as serum samples allows significantly increasing the width of the observable chemical space, in this case almost twofold. The relevance of using complementary SPM is very probably exacerbated by the overall low abundance of xenobiotics in the samples. Indeed, low-abundant compounds have an increased probability of being lost to either SPM, and therefore generating an extreme FC value. Therefore, using a dual sample preparation process is a major advantage to increase the accuracy of the characterization of the chemical exposome. However, initial sample volume should account for this fact, which might be limiting in the case of valuable biological samples.

In this chapter, the large-scale application of the previously optimized workflow using 125 samples from the Breton Pélagie cohort was presented. This scaling up process has necessitated using total ion current area normalization to account for the analytical variability that occurred over the course of the multiple-week analysis campaign. The use of a suspect screening strategy involving MS1 and MS2 predictors has led to the annotation of 92 environmental chemical compounds with various uses including pesticides, medication, preservatives and synthesis intermediates. Comparing the detection of these annotated compounds in samples prepared with PPT and the Phree PLR plate demonstrated the relevance of combining SPM to expand the visible chemical space. Indeed, close to 75% of annotated compounds were only visible with one SPM. This comparison also allowed identifying some factors, such as polarity or steric hindrance, that might determine whether a compound is more readily detectable with either SPM. For instance, polar compounds seem to be better detected in samples prepared with PPT, whereas organophosphate flame-retardants are favored in samples prepared with Phree PLR plates. This large-scale application is therefore a successful application of the optimized suspect screening workflow developed in this PhD work. Its implementation has allowed expanding knowledge about the chemical exposome of the considered population. As one of the Pélagie cohort's objectives is to investigate the role of the urban-rural context on human health, the chemical fingerprints could be further used in association to this contextual data. This could be useful to prioritize more features for annotation and continue documenting the chemical exposome of Breton 12-year-olds.

References

1. Huhn S., Escher B.I.*, et al.*, Unravelling the chemical exposome in cohort studies: routes explored and steps to become comprehensive. *Environmental Sciences Europe* **2021,** *33* (1).
2. David A., Chaker J.*, et al.*, Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int* **2021,** *156*, 106630.
3. Vermeulen R., Schymanski E.L.*, et al.*, The exposome and health: Where chemistry meets biology. *Science* **2020,** (367), 392-6.
4. Panagopoulos Abrahamsson D., Wang A.*, et al.*, A Comprehensive Non-targeted Analysis Study of the Prenatal Exposome. *Environ Sci Technol* **2021**.
5. Milman B.L.,Zhurkovich I.K., The chemical space for non-target analysis. *TrAC Trends in Analytical Chemistry* **2017,** *97*, 179-87.
6. Athersuch T., Metabolome analyses in exposome studies: Profiling methods for a vast chemical space. *Arch Biochem Biophys* **2016,** *589*, 177-86.
7. Jones D.P., Sequencing the exposome: A call to action. *Toxicol Rep* **2016,** *3*, 29-45.
8. Monteiro Bastos da Silva J., Chaker J.*, et al.*, Improving Exposure Assessment Using Non-Targeted and Suspect Screening: The ISO/IEC 17025: 2017 Quality Standard as a Guideline. *Journal of Xenobiotics* **2021,** *11* (1), 1-15.
9. Liu K.H., Nellis M.*, et al.*, Reference Standardization for Quantification and Harmonization of Large-Scale Metabolomics. *Anal Chem* **2020**.
10. Brunius C., Shi L.*, et al.*, Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* **2016,** *12* (11), 173.
11. Binter A.C., Bannier E.*, et al.*, Exposure of pregnant women to organophosphate insecticides and child motor inhibition at the age of 10-12 years evaluated by fMRI. *Environ Res* **2020,** *188*, 109859.
12. Cartier C., Warembourg C.*, et al.*, Organophosphate Insecticide Metabolites in Prenatal and Childhood Urine Samples and Intelligence Scores at 6 Years of Age: Results from the Mother-Child PELAGIE Cohort (France). *Environ Health Perspect* **2016,** *124* (5), 674-80.
13. Chevrier C., Serrano T.*, et al.*, Environmental determinants of the urinary concentrations of herbicides during pregnancy: the PELAGIE mother-child cohort (France). *Environ Int* **2014,** *63*, 11-8.
14. Tsugawa H., Cajka T.*, et al.*, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.
15. Want E.J., Wilson I.D.*, et al.*, Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* **2010,** *5* (6), 1005-18.
16. Mizuno H., Ueda K.*, et al.*, The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics. *Biomed Chromatogr* **2017,** *31* (1).
17. Horai H., Arita M.*, et al.*, MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **2010,** *45* (7), 703-14.
18. Allen F., Pon A.*, et al.*, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* **2014,** *42* (Web Server issue), W94-9.
19. Ruttkies C., Schymanski E.L.*, et al.*, MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **2016,** *8*, 3.
20. Schymanski E.L., Jeon J.*, et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.
21. Aalizadeh R., Alygizakis N.*, et al.*, Merged NORMAN Suspect List: SusDat. v0.4.1 ed.; 2022.
22. Wang A., Gerona R.R.*, et al.*, A Suspect Screening Method for Characterizing Multiple Chemical Exposures among a Demographically Diverse Population of Pregnant Women in San Francisco. *Environ Health Perspect* **2018,** *126* (7), 077009.

23. David A., Chaker J.*, et al.*, Acetaminophen metabolism revisited using non-targeted analyses: Implications for human biomonitoring. *Environ Int* **2021,** *149*, 106388.
24. Williams A.J., Grulke C.M.*, et al.*, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* **2017,** *9* (1), 61.
25. Desvergne B., Feige J.N.*, et al.*, PPAR-mediated activity of phthalates: A link to the obesity epidemic? *Mol Cell Endocrinol* **2009,** *304* (1-2), 43-8.
26. Bonvallot N., Jamin E.L.*, et al.*, Suspect screening and targeted analyses: Two complementary approaches to characterize human exposure to pesticides. *Science of The Total Environment* **2021,** *786*.
27. Semchuk K., McDuffie H.*, et al.*, Body mass index and bromoxynil exposure in a sample of rural residents during spring herbicide application. *J Toxicol Environ Health A* **2004,** *67* (17), 1321-52.
28. McMahen R.L., Strynar M.J.*, et al.*, Identification of fipronil metabolites by time-of-flight mass spectrometry for application in a human exposure study. *Environ Int* **2015,** *78*, 16-23.
29. Ahmad S., Kalra H.*, et al.*, HybridSPE: A novel technique to reduce phospholipid-based matrix effect in LC-ESI-MS Bioanalysis. *J Pharm Bioallied Sci* **2012,** *4* (4), 267-75.

Conclusion and perspectives

Characterizing the human internal chemical exposome using non-targeted approaches presents several methodological and technological challenges. Indeed, existing workflows classically used in metabolomics should be adapted at every step to allow the detection of low-abundant chemicals in complex biological matrices. To address these challenges, the optimization of the most critical steps of an HRMS-based exposomics workflow was performed in this PhD project. The developed HRMS-based non-targeted workflow was then implemented in a larger scale application to assess human exposure to complex chemical mixtures.

Three steps of the non-targeted and suspect screening workflow were investigated, namely sample preparation, data processing, and annotation. Firstly, the preparation of serum and plasma samples with twelve sample preparation methods was investigated. Two SPM, namely protein precipitation and the Phree phospholipid removal plate, presented adequate performance for quantitative (e.g. recovery, repeatability, etc.) and qualitative (e.g. ease of implementation, etc.) criteria. Their application on cohort plasma and serum samples allowed demonstrating their complementarity, as more than 60% of features were at least significantly favored by either SPM, and 40% of features was only visible in with one SPM. As they provided different pictures on the chemical exposome, their combined use is relevant in the context of characterizing a diverse set of compounds. A single sample preparation workflow involving both sample preparation methods was proposed as a way to widen the visible chemical space. This work demonstrated the necessity to systematically delineate the impact of sample preparation on the perimeter of the observable chemical space.

Data processing in non-targeted exposomics applications is a particularly complex task, as the compounds of interest often present as low-abundant signals that should be properly disentangled from the noise. As the many available data processing software tools were mostly built for metabolomics, they should be optimized and evaluated for exposomics applications. Four software tools, including vendor (i.e. MarkerView and Progenesis QI) and open source (i.e. MZMine2 and XCMS) software, were therefore optimized and compared for the processing non-targeted exposomics data. This systematic evaluation highlighted the need for manual optimization of non-targeted data processing software for exposomics applications. This optimization is necessary, as it allowed increasing the detection of spiked samples by as much as 18%.

Lastly, the need for efficient annotation strategies is still salient in HRMS-based exposomics applications. The developed software aimed to partly automatize a suspect screening approach based on three MS1 chemical predictors: m/z, experimental and/or predicted Rt and isotopic fit. Confidence indices were built to score the likeness of suspects and features, and allow the efficient prioritization of suggested pre-annotations. A global confidence index

combines all computed CI to score the overall resemblance between suspect and feature, and can used as a cut-off criterion to limit false positive annotations. This tool was compared to other tools available to assist suspect screening approaches (i.e. xMSannotator, MS-DIAL, msPurity and MZMine2). The use of experimental and predicted Rt as well as the scoring system were major advantages of the in-house software for compound prioritization. However, it does not yet allow the use of MS2 fragmentation patterns, which is a highly discriminant criterion allowing to significantly limit false positive annotation when it is available. The first implementation of the in-house software allowed the annotation of diverse compounds of the internal chemical exposome with high confidence indices, which highlighted the relevance of the scoring system for prioritizing suggested annotations.

The optimized workflow was implemented on a large-scale proof-of-concept application. This study on 125 serum samples from 12-year-old Bretons allowed demonstrating the applicability of this workflow on a multi-batch scale to characterize the human internal chemical exposome. Indeed, the use of the previously described strategies for sample preparation, data processing and annotation has allowed identifying 92 highly diverse compounds in terms of mass (i.e. $100.0754 \leq [M+H]^+ \leq 811.4913$), polarity (i.e. $-2.7 \leq \log P \leq 16$) and sources (e.g. dietary, medication, industrial, etc.). This application provided valuable information on the chemical exposome in general, and on the impact of different workflow steps on the results of such studies. In particular, the use of MS1 predictors for annotation allowed prioritizing metabolites of known toxicants, which would have otherwise been missed. The generated data will allow to better apprehend the perimeter of the chosen workflow, and to identify the gaps needing additional investigating efforts. Additionally, the chemical fingerprints generated in this large-scale application could be reused with different data processing and annotation strategies, such as integrating other types of data collected according to the epidemiological experimental design (i.e. data from targeted assays, clinical data, lifestyle data, etc.), and establishing associations to further investigate.

Overall, non-targeted and suspect screening approaches are highly promising to investigate the environment-health links. However, several challenges remain to be addressed to implement these approaches to their full potential, such as the need for multi-systems approaches when aiming for a wider characterization of the chemical exposome. Indeed, no single analytical platform will allow capturing the wide range of compounds currently in use in our environment. Therefore, combining different technologies, such as LC-HRMS and GC-HRMS, would be helpful in expanding the visible chemical space. Adding a separation to the chromatographic separation (i.e. LC×LC or GC×GC) or using ion mobility spectrometry may also present a valuable addition to characterize the human internal chemical exposome. The

two main limitations to these approaches are the financial burden induced by purchasing and maintaining several pieces of equipment, and the limited availability of software tools and/or databases to process the generated data. Collaborations at the national, European and/or international level may greatly help in overcoming these limitations.

Another challenge that should be addressed is the ongoing need to improve the annotation process. Indeed, despite the many efforts undertaken in the last few years to expand suspect and MS2 libraries, they remain incomplete and/or non-interchangeable between tools. Pursuing the existing efforts in terms of both data collection and harmonization will be beneficial to the scientific community. Moreover, specific efforts should be dedicated to including known or predicted metabolites of exposome compounds, as they may only be detectable under metabolized forms. Acquiring MS2 data for these compounds is also challenging, as many are not commercialized, and their custom-made synthesis represents a financial burden. At the scale of the laboratory, further developments will be carried out regarding the in-house software, such as adding MS2 predictors to further reduce false positive annotations.

Non-targeted and suspect screening approaches should be used to generate lists of compounds of interest that should be further investigated. Particularly, they should be followed up by large-scale targeted HBM studies, to confirm these compounds' prevalence in the population of interest and to generate quantitative data. This would be crucial in evaluating the need for risk assessment, and regulatory action further down the line. These HBM programs should also go through a harmonization process to ensure inter-comparability of data acquired over several countries and/or continents, as is done in the HBM4EU initiative.

Lastly, these chemicals of interest should be further investigated through toxicological approaches to improve knowledge on their mechanism of action. Regulatory action may be taken in accordance with the results of the risk assessment process. It should also be noted that the developed toxicological approaches should be high-throughput, and ideally consider mixture effects.

To conclude, the workflow optimized in the context of this PhD was demonstrated as efficient for the non-targeted characterization of the human internal chemical exposome. These approaches are highly valuable tools to investigate the effects of environmental chemical exposures on health, and generate a rapidly increasing interest at the European and international scale, as demonstrated by the setting up of the EIRENE infrastructure for instance. Large-scale collaborations at these levels will allow generating robust and inter-comparable data to both describe the human chemical exposome and hopefully better

understand the etiology of chronic disease. However, developments and harmonization efforts are still required to reach the full potential of non-targeted and suspect screening approaches, and offer operational solutions to limit the presence of harmful chemicals in our environment.

Appendices

1. Appendix 1. Chapter II

1.1. Detailed list of the optimization mix and internal standards

Table A1 – Detailed list of the optimization mix and internal standards, with SMILES identifiers, chemical formulas, monoisotopic mass, observed ion, retention time, octanol-water partition coefficient (logP), and CAS number when available.

	Compound name	SMILES	Chemical formula	Monoisotopic mass (Da)	Observed ion	Retention time (min)	logP	CAS
Standard compounds	2-Phenylphenol	C1=CC=C(C=C1)C2=CC=CC=C2O	C12H10O	170.0732	$[M-H]^-$	30.19	3.28	90-43-7
	Acetochlor	CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C	C14H20ClNO2	269.1183	$[M-H]^-$	40.57	4.14	123113-74-6
	Acetylsalicylic acid	CC(=O)OC1=CC=CC=C1C(=O)O	C9H8O4	180.0423	$[M-H]^-$	8.65	1.24	50-78-2
	Aflatoxin B1	COC1=C2C3=C(C(=O)CC3)C(=O)OC2=C4C5C=COC5OC4=C1	C17H12O6	312.0634	$[M+H]^+$	17.52	1.73	27261-02-5
	Aminobenzimidazole	C1=CC=C2C(=C1)NC(=N2)N	C7H7N3	133.0640	$[M+H]^+$	4.74	0.91	934-32-7
	Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	C19H26O2	286.1933	$[M+H]^+$	31.50	2.75	63-05-8
	Arachidonic Acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	C20H32O2	304.2402	$[M-H]^-$	47.00	6.99	93444-49-6
	Azoxystrobin	COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC	C22H17N3O5	403.1168	$[M+H]^+$	38.03	2.64	215934-32-0
	Boscalid	C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl	C18H12Cl2N2O	342.0327	$[M+H]^+$	38.00	2.96	188425-85-6
	Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	C15H12N2O	236.0950	$[M+H]^+$	18.01	2.45	298-46-4
	Carbendazim	COC(=O)NC1=NC2=CC=CC=C2N1	C9H9N3O2	191.0695	$[M+H]^+$	5.69	1.52	63278-70-6
	Chlorpyrifos	CCOP(=S)(OCC)OC1=NC(=C(C=C1Cl)Cl)Cl	C9H11Cl3NO3PS	348.9263	$[M+H]^+$	45.53	4.70	39475-55-3
	Clothianidin	CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl	C6H8ClN5O2S	249.0087	$[M+H]^+$	7.99	0.73	205510-53-8
	Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	C18H21NO3	299.1521	$[M+H]^+$	5.12	1.39	76-57-3
	Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)CO)O)C	C21H28O5	360.1937	$[M+H]^+$	16.12	1.47	53-06-5
	Cotinine	CN1C(CCC1=O)C2=CN=CC=C2	C10H12N2O	176.0950	$[M+H]^+$	4.31	0.07	486-56-6
	Cyprodinil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C3CC3	C14H15N3	225.1266	$[M+H]^+$	33.22	4.00	121552-61-2
	Diazinon	CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C	C12H21N2O3PS	304.1011	$[M+H]^+$	43.38	3.81	30583-38-1
	Diclofenac	C1=CC=C(C(=C1)CC(=O)O)NC2=C(C=CC=C2Cl)Cl	C14H11Cl2NO2	295.0167	$[M-H]^-$	39.59	4.51	15307-86-5
	Dimethyldithiophosphate	COP(=S)(OC)S	C2H7O2PS2	157.9625	$[M-H]^-$	2.95	0.63	756-80-9

Table A1 – (continued) Detailed list of the optimization mix and internal standards, with SMILES identifiers, chemical formulas, monoisotopic mass, observed ion, retention time, octanol-water partition coefficient (logP), and CAS number when available.

	Compound name	SMILES	Chemical formula	Monoisotopic mass (Da)	Observed ion	Retention time (min)	logP	CAS
Standard compounds	Estrone	CC12CCC3C(C1CCC2=O)CCC4=C3C=CC(=C4)O	C18H22O2	270.1620	$[M+H]^+$	31.60	3.13	53-16-7
	Fluoxetine	CNCCC(C1=CC=CC=C1)OC2=CC=C(C=C2)C(F)(F)F	C17H18F3NO	309.1340	$[M+H]^+$	23.71	4.05	57226-07-0
	Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	C21H30O5	362.2093	$[M+H]^+$	15.86	1.61	50-23-7
	Hydroxyindoleacetic acid	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	C10H9NO3	191.0582	$[M-H]^-$	5.71	1.41	113303-91-6
	Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	C13H18O2	206.1307	$[M-H]^-$	39.94	3.97	58560-75-1
	Imidacloprid	C1CN(C(=N[N+](=O)[O-])N1)CC2=CN=C(C=C2)Cl	C9H10ClN5O2	255.0523	[M+H]+	8.57	0.57	138261-41-3
	Ketoprofen	CC(C1=CC(=CC=C1)C(=O)C2=CC=CC=C2)C(=O)O	C16H14O3	254.0943	$[M+H]^+$	28.13	3.12	22071-15-4
	Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	C20H32O4	336.2301	$[M-H]^-$	39.52	4.10	71160-24-2
	Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N	C25H40N2O6S	496.2607	$[M-H]^-$	33.04	1.40	73836-78-9
	Malathion	CCOC(=O)CC(C(=O)OCC)SP(=S)(OC)OC	C10H19O6PS2	330.0361	$[M+H]^+$	40.81	2.89	121-75-5
	Nicotine	CN1CCCC1C2=CN=CC=C2	C10H14N2	162.1157	$[M+H]^+$	3.37	1.17	551-13-3
	Paracetamol	CC(=O)NC1=CC=C(C=C1)O	C8H9NO2	151.0633	$[M+H]^+$	4.98	0.31	8055-08-1
	Paroxetine	C1CNCC(C1C2=CC=C(C=C2)F)COC3=CC4=C(C=C3)OCO4	C19H20FNO3	329.1427	$[M+H]^+$	18.34	1.23	63952-24-9
	Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	C17H19NO3	285.1365	$[M+H]^+$	36.42	2.78	147030-08-8
	Pravastatin	CCC(C)C(=O)OC1CC(C=C2C1C(C(C=C2)C)CCC(CC(CC(=O)O)O)O)O	C23H36O7	424.2461	$[M+H]^+$	20.50	1.65	81093-37-0
	Prochloraz	CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=CN=C2	C15H16Cl3N3O2	375.0308	$[M+H]^+$	38.74	3.78	67747-09-5
	Progesterone	CC(=O)C1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	C21H30O2	314.2246	$[M+H]^+$	42.10	3.87	257630-50-5
	Propiconazole	CCCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl	C15H17Cl2N3O2	341.0698	$[M+H]^+$	41.73	3.72	75881-82-2
	Prostaglandin D2	CCCCCC(C=CC1C(C(CC1=O)O)CC=CCCCC(=O)O)O	C20H32O5	352.2250	$[M-H]^-$	27.60	3.23	41598-07-6
	Prostaglandin E2	CCCCCC(C=CC1C(CC(=O)C1CC=CCCCC(=O)O)O)O	C20H32O5	352.2250	$[M-H]^-$	26.50	2.82	363-24-6
	Prostaglandin F2a	CCCCCC(C=CC1C(CC(C1CC=CCCCC(=O)O)O)O)O	C20H34O5	354.2406	$[M-H]^-$	25.60	2.61	13535-33-6
	Prostaglandin J2	CCCCCC(C=CC1C(C=CC1=O)CC=CCCCC(=O)O)O	C20H30O4	334.2144	$[M-H]^-$	26.54	3.60	60203-57-8

Table A1 – (continued) Detailed list of the optimization mix and internal standards, with SMILES identifiers, chemical formulas, monoisotopic mass, observed ion, retention time, octanol-water partition coefficient (logP), and CAS number when available.

	Compound name	SMILES	Chemical formula	Monoisotopic mass (Da)	Observed ion	Retention time (min)	logP	CAS
Standard compounds	Sertraline	CNC1CCC(C2=CC=CC=C12)C3=CC(=C(C=C3)Cl)Cl	C17H17Cl2N	305.0738	$[M+H]^+$	24.34	5.10	79559-97-0
	Solanidine	CC1CCC2C(C3C(N2C1)CC4C3(CCC5C4CC=C6C5(CCC(C6)O)C)C)C	C27H43NO	397.3345	$[M+H]^+$	24.54	4.88	80-78-4
	Tebuconazole	CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O	C16H22ClN3O	307.1451	$[M+H]^+$	39.36	3.70	80443-41-0
	Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	C19H28O2	288.2089	$[M+H]^+$	28.90	3.32	58-22-0
	Thiacloprid	C1CSC(=NC#N)N1CC2=CN=C(C=C2)Cl	C10H9ClN4S	252.0236	$[M+H]^+$	12.24	1.25	111988-49-9
	Thiamethoxam	CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl	C8H10ClN5O3S	291.0193	$[M+H]^+$	6.97	1.52	153719-23-4
	Triclosan	C1=CC(=C(C=C1Cl)O)OC2=C(C=C(C=C2)Cl)Cl	C12H7Cl3O2	287.9512	$[M-H]^-$	43.79	4.76	3380-34-5
	Venlafaxine	CN(C)CC(C1=CC=C(C=C1)OC)C2(CCCCC2)O	C17H27NO2	277.2042	$[M+H]^+$	9.84	0.43	93413-69-5
Internal standards	2-phenylphenol-13C6	n.a.	[13C]6C6H10O	176.0933	[M-H]-	30.19	n.a.	287389-48-4
	Acetochlor-d11	n.a.	C14D11H9ClNO2	280.1873	[M-H]-	40.57	n.a.	1189897-44-6
	Azoxystrobin-d4	n.a.	C22D4H13N3O5	407.1419	[M+H]+	38.03	n.a.	1346606-39-0
	Carbamazepine-13C6	n.a.	[13C]6C9H12N2O	242.1151	[M+H]+	18.01	n.a.	n.a.
	Carbendazim-d4	n.a.	C9D4H5N3O2	195.0946	$[M+H]^+$	5.69	n.a.	291765-95-2
	Chlorpyrifos-d10	n.a.	C9D10HCl3NO3PS	358.9891	[M+H]+	45.53	n.a.	285138-81-0
	Cotinine-d3	n.a.	C10D3H9N2O	179.1138	[M+H]+	4.31	n.a.	110952-70-0
	Diazinon-d10	n.a.	C12D10H11N2O3PS	314.1638	$[M+H]^+$	43.38	n.a.	100155-47-3
	Diclofenac-13C6	n.a.	[13C]2C12H11Cl2NO2	297.0234	$[M-H]^-$	39.59	n.a.	n.a.
	Dimethyldithiophosphate-13C2	n.a.	[13C]2H7O2PS2	159.9692	$[M-H]^-$	2.95	n.a.	1329610-82-3
	Estrone-d4	n.a.	C18D4H18O2	274.1871	$[M+H]^+$	31.60	n.a.	53866-34-5
	Fluoxetine-d6	n.a.	C17D6H12F3NO	315.1717	$[M+H]^+$	23.71	n.a.	n.a.
	Hydrocortisone-d4	n.a.	C21D4H26O5	366.2344	$[M+H]^+$	15.86	n.a.	73565-87-4
	Ibuprofen-d3	n.a.	C13D3H15O2	209.1495	$[M-H]^-$	39.94	n.a.	121662-14-4
	Imidacloprid-d4	n.a.	C9D4H6ClN5O2	259.0774	[M+H]+	8.57	n.a.	1015855-75-0
	Ketoprofen-d3	n.a.	C16D3H11O3	257.1131	$[M+H]^+$	28.13	n.a.	159490-55-8

Table A1 – (continued) Detailed list of the optimization mix and internal standards, with SMILES identifiers, chemical formulas, monoisotopic mass, observed ion, retention time, octanol-water partition coefficient (logP), and CAS number when available.

	Compound name	SMILES	Chemical formula	Monoisotopic mass (Da)	Observed ion	Retention time (min)	logP	CAS
Internal standards	Leukotriene B4-d4	n.a.	C20D4H28O4	340.2552	$[M-H]^-$	39.52	n.a.	93951-88-3
	Paracetamol-d4	n.a.	C8D4H5NO2	155.0884	$[M+H]^+$	4.98	n.a.	64315-36-2
	Prostaglandin E2-d4	n.a.	C20D4H28O5	356.2501	$[M-H]^-$	26.50	n.a.	34210-10-1
	Tebuconazole-d6	n.a.	C16D6H16ClN3O	313.1828	$[M+H]^+$	39.36	n.a.	1246818-83-6
	Testosterone-d3	n.a.	C19D3H25O2	291.2278	$[M+H]^+$	28.90	n.a.	77546-39-5
	Thiamethoxam-d4	n.a.	C8D4H6ClN5O3S	295.0444	$[M+H]^+$	6.97	n.a.	1331642-98-8

1.2. Column diameter and flow rate optimization

Table A2a – Detailed mean area and coefficient of variation (CV) on 3 replicates for column diameter and flow rate optimization with 50 standards spiked at 20 ng/mL

	20 pg							
	Ø 2.1 mm				Ø 1.0 mm			
	0.30 mL/min		0.15 mL/min		0.10 mL/min		0.05 mL/min	
Compound name	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**
2-Phenylphenol	0	0.0	0	0.0	849	4.4	1711	2.6
Acetochlor	4475	0.9	9755	3.6	7597	4.5	14321	1.1
Acetylsalicylic acid	3808	11.9	1175	20.7	0	0.0	1522	10.5
Aflatoxin B1	7860	2.1	14889	1.6	27507	4.4	48357	1.6
Aminobenzimidazole	42778	1.9	66388	1.3	106609	0.7	109914	2.6
Androstenedione	22545	1.0	43711	2.2	63325	3.9	116688	2.7
Arachidonic Acid	0	0.0	0	0.0	10608	1.5	2145	2.9
Azoxystrobin	27111	1.4	50413	1.9	74847	1.5	138630	1.2
Boscalid	9724	2.4	19987	2.7	30977	1.7	59087	1.7
Carbamazepine	40813	2.9	69130	1.6	56790	3.1	92419	1.5
Carbendazim	26851	2.8	50107	1.3	58931	1.7	98122	2.2
Chlorpyrifos	9160	2.9	8375	4.9	26006	4.2	17375	2.4
Clothianidin	3967	1.8	7293	2.1	3512	3.3	6277	1.9
Codeine	29568	2.3	51887	1.3	84592	1.0	125101	0.8
Cortisone	11580	1.8	18896	1.8	36473	2.3	61874	1.2
Cotinine	17447	1.8	24134	2.3	43291	0.6	44092	3.8
Cyprodinil	220019	3.5	304236	2.0	154113	3.3	271592	2.6
Diazinon	353679	1.8	670435	1.8	431259	1.1	646828	1.6
Diclofenac	5852	1.8	11670	2.4	14361	1.3	26631	1.2
Dimethyldithiophosphate	0	0.0	0	0.0	552	1.9	894	1.9
Estrone	9883	2.0	20255	1.9	27555	1.4	54243	1.8
Fluoxetine	34612	3.2	47031	1.8	48521	1.1	73356	0.9
Hydrocortisone	15075	1.7	24010	1.2	47430	2.3	79301	2.0

Table A2a – (continued) Detailed mean area and coefficient of variation (CV) on 3 replicates for column diameter and flow rate optimization with 50 standards spiked at 20 ng/mL

	20 pg							
	Ø 2.1 mm				**Ø 1.0 mm**			
	0.30 mL/min		**0.15 mL/min**		**0.10 mL/min**		**0.05 mL/min**	
Compound name	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**
Hydroxyindoleacetic acid	0	0.0	2694	4.2	1875	2.9	3388	2.8
Ibuprofen	0	0.0	0	0.0	214	1.7	660	1.8
Imidacloprid	11776	2.1	15929	1.4	9504	1.8	16735	1.2
Ketoprofen	598052	1.0	942528	1.2	109328	1.1	173304	1.2
Leukotriene B4	1917	3.3	2783	6.4	6625	1.3	9609	2.9
Leukotriene D4	2623	18.4	3548	12.3	8816	8.4	13882	0.5
Malathion	6006	2.5	11565	1.8	8529	3.4	15602	0.3
Nicotine	4192	3.2	6936	3.3	6392	4.9	11201	2.8
Paracetamol	6660	2.9	14481	3.6	15879	2.0	25567	2.5
Paroxetine	64379	2.5	80067	2.8	111047	2.2	186423	0.3
Piperine	22541	1.4	38849	3.4	49507	1.6	90501	1.3
Pravastatin	2006	2.7	2444	4.0	2201	2.7	2201	3.1
Prochloraz	11770	8.1	19850	5.7	16390	7.0	29451	2.2
Progesterone	28426	0.3	71350	3.0	98298	7.9	161313	1.5
Propiconazole	32615	3.4	80152	5.7	88708	4.6	137210	2.0
Prostaglandin D2	2493	7.9	3345	8.9	9217	3.0	11703	4.3
Prostaglandin E2	1808	8.2	2983	6.9	7654	1.3	13791	2.8
Prostaglandin F2a	2448	6.7	3355	2.2	8857	0.9	16552	2.2
Prostaglandin J2	2854	6.6	3149	2.1	7569	0.5	14228	2.2
Sertraline	13444	3.0	19287	1.5	13848	2.3	21028	0.7
Solanidine	64689	0.8	85483	1.8	100975	1.0	158675	2.5
Tebuconazole	49740	4.3	79713	4.3	98595	6.2	166400	1.9
Testosterone	27084	3.2	47005	1.9	72103	1.9	126546	1.3
Thiacloprid	19908	3.3	28586	1.6	27747	4.1	54234	2.5

Table A2a – (continued) Detailed mean area and coefficient of variation (CV) on 3 replicates for column diameter and flow rate optimization with 50 standards spiked at 20 ng/mL

	20 pg							
	Ø 2.1 mm				**Ø 1.0 mm**			
	0.30 mL/min		**0.15 mL/min**		**0.10 mL/min**		**0.05 mL/min**	
Compound name	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**
Thiamethoxam	3969	6.2	6867	2.7	8244	5.0	11818	2.2
Triclosan	17442	2.4	20114	3.0	27857	1.2	44355	2.0
Venlafaxine	118494	2.2	133888	1.4	143819	1.6	235916	2.4
Median	**11675**	**2.4**	**19569**	**2.1**	**27531**	**2.0**	**44224**	**2.0**

Table A2b – Detailed mean area and coefficient of variation (CV) on 3 replicates for column diameter and flow rate optimization with 50 standards spiked at 200 ng/mL

	200 pg							
	Ø 2.1 mm				**Ø 1.0 mm**			
	0.30 mL/min		**0.15 mL/min**		**0.10 mL/min**		**0.05 mL/min**	
Compound name	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**
2-Phenylphenol	1258	7.5	2523	0.7	6053	3.9	14740	2.6
Acetochlor	44920	0.7	101742	1.0	72798	0.3	150161	2.6
Acetylsalicylic acid	3511	5.4	1904	17.5	2541	6.0	5506	8.1
Aflatoxin B1	92085	7.3	175374	2.1	278845	2.8	530533	2.1
Aminobenzimidazole	482920	1.7	660167	2.2	1004025	0.9	1059186	1.2
Androstenedione	314943	0.9	569012	2.8	728528	2.1	1391952	2.2
Arachidonic Acid	11696	5.8	16799	3.7	76929	12.8	23384	9.4
Azoxystrobin	416255	2.1	700821	1.8	839750	2.8	1579550	2.8
Boscalid	136846	1.6	260687	1.9	365343	0.8	737820	0.5
Carbamazepine	486607	1.6	794147	1.2	562533	1.2	1009330	3.3

Table A2b – (continued) Detailed mean area and coefficient of variation (CV) on 3 replicates for column diameter and flow rate optimization with 50 standards spiked at 200 ng/mL

	200 pg							
	Ø 2.1 mm				Ø 1.0 mm			
	0.30 mL/min		0.15 mL/min		0.10 mL/min		0.05 mL/min	
Compound name	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**
Carbendazim	303001	1.4	529255	0.2	584976	2.3	1038643	0.9
Chlorpyrifos	143384	7.2	132435	7.0	323684	0.5	232045	4.5
Clothianidin	44093	0.8	77368	1.3	33979	1.7	61958	1.7
Codeine	374122	2.4	596108	1.7	845795	1.6	1342160	1.6
Cortisone	179839	0.9	251319	1.2	422111	1.6	717505	1.2
Cotinine	189159	2.8	249964	0.5	424143	0.9	180963	4.5
Cyprodinil	2408536	1.1	3459616	1.0	1716890	2.1	3191239	2.2
Diazinon	3445514	1.5	6974612	2.0	4164532	1.1	6849338	1.0
Diclofenac	66823	1.4	131355	1.6	158232	1.5	302371	2.6
Dimethyldithiophosphate	16549	2.7	18745	2.5	19885	2.2	24558	2.2
Estrone	123256	3.2	250416	2.2	307707	2.3	496308	3.1
Fluoxetine	499044	2.5	639329	3.4	566846	1.2	887694	3.1
Hydrocortisone	218973	5.1	321839	2.5	526415	0.8	910547	0.7
Hydroxyindoleacetic acid	5521	2.8	25046	3.5	14672	4.0	26215	3.0
Ibuprofen	22198	3.3	27854	3.1	27820	2.2	37854	2.5
Imidacloprid	163270	1.2	177924	1.7	308757	2.9	382121	1.6
Ketoprofen	601779	1.6	928294	0.5	114289	1.4	186576	2.4
Leukotriene B4	17460	1.8	27102	1.2	69337	1.5	102238	2.7
Leukotriene D4	26004	3.6	34823	2.8	117829	1.7	164339	2.0
Malathion	75799	1.1	149175	5.4	96160	2.5	179783	4.5
Nicotine	41136	2.3	60698	1.6	20386	9.6	48066	1.9
Paracetamol	71826	2.6	149189	1.6	147647	1.0	239774	1.5
Paroxetine	817603	2.5	1033107	1.1	1169024	6.1	2124497	1.2
Piperine	313665	1.0	504088	1.2	586547	1.2	1098772	1.7
Pravastatin	24687	2.3	27128	1.7	24320	4.1	22995	2.9

Table A2b – (continued) Detailed mean area and coefficient of variation (CV) on 3 replicates for column diameter and flow rate optimization with 50 standards spiked at 200 ng/mL

	200 pg							
	Ø 2.1 mm				**Ø 1.0 mm**			
	0.30 mL/min		**0.15 mL/min**		**0.10 mL/min**		**0.05 mL/min**	
Compound name	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**	**Mean**	**CV (%)**
Prochloraz	177023	1.5	238416	1.2	199840	2.6	380390	2.0
Progesterone	420824	2.1	971167	0.6	1057300	1.1	1883434	3.9
Propiconazole	488615	1.4	1099382	1.0	1046343	1.9	1614942	0.6
Prostaglandin D2	18030	2.9	30089	3.5	100645	2.2	129757	2.2
Prostaglandin E2	15201	2.5	26502	4.0	84631	1.0	149932	5.7
Prostaglandin F2a	12235	2.0	18597	1.7	19742	1.8	29545	2.0
Prostaglandin J2	19452	2.0	24560	2.0	22457	1.7	31247	1.8
Sertraline	189844	3.7	256606	1.1	150102	1.1	233249	1.7
Solanidine	860779	2.8	1092428	3.4	1114014	3.7	1851604	2.9
Tebuconazole	684637	2.0	1042733	2.2	1128292	2.1	1495029	4.1
Testosterone	375611	1.5	599287	1.5	796827	1.2	1454106	1.6
Thiacloprid	228117	1.3	338182	0.8	321764	0.9	380140	1.5
Thiamethoxam	41008	2.9	73229	3.4	71889	3.9	110050	4.3
Triclosan	178456	1.9	265478	1.4	345788	1.7	387750	1.5
Venlafaxine	1836079	2.3	1964486	2.6	1365043	1.4	2392643	1.6
Median	**170146**	**2.1**	**250190**	**1.7**	**308232**	**1.7**	**380265**	**2.2**

1.3. Detailed list of the retention time prediction set

Table A3 – Detailed list of compound from the retention time prediction set, with SMILES identifier, chemical formula, monoisotopic mass, and CAS number

Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
1-(3,4-Dichlorophenyl)-3-methylurea	CNC(=O)NC1=CC(=C(C=C1)Cl)Cl	C8H8Cl2N2O	218.0014	3567-62-2
1-(3,4-Dichlorophenyl)urea	C1=CC(=C(C=C1NC(=O)N)Cl)Cl	C7H6Cl2N2O	203.9857	2327-02-8
1-(4-Isopropylphenyl)urea	CC(C)C1=CC=C(C=C1)NC(=O)N	C10H14N2O	178.1106	56046-17-4
2,4-mcpa	CC1=C(C=CC(=C1)Cl)OCC(=O)O	C9H9ClO3	200.0240	94-74-6
2-chloro-4-methylbenzoic acid	CC1=CC(=C(C=C1)C(=O)O)Cl	C8H7ClO2	170.0135	7697-25-8
2-Phenylphenol	C1=CC=C(C=C1)C2=CC=CC=C2O	C12H10O	170.0732	90-43-7
Acetamiprid	CC(=NC#N)N(C)CC1=CN=C(C=C1)Cl	C10H11ClN4	222.0672	135410-20-7
Acetochlor	CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C	C14H20ClNO2	269.1183	123113-74-6
Aflatoxin B1	COC1=C2C3=C(C(=O)CC3)C(=O)OC2=C4C5C=COC5OC4=C1	C17H12O6	312.0634	27261-02-5
Alachlor	CCC1=C(C(=CC=C1)CC)N(COC)C(=O)CCl	C14H20ClNO2	269.1183	15972-60-8
Ametryn	CCNC1=NC(=NC(=N1)SC)NC(C)C	C9H17N5S	227.1205	834-12-8
Amidosulfuron	CN(S(=O)(=O)C)S(=O)(=O)NC(=O)NC1=NC(=CC(=N1)OC)OC	C9H15N5O7S2	369.0412	120923-37-7
Aminobenzimidazole	C1=CC=C2C(=C1)NC(=N2)N	C7H7N3	133.0640	934-32-7
Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	C19H26O2	286.1933	63-05-8
Arachidonic Acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	C20H32O2	304.2402	93444-49-6
Atrazine	CCNC1=NC(=NC(=N1)Cl)NC(C)C	C8H14ClN5	215.0938	1912-24-9
Atrazine-2-hydroxy	CCNC1=NC(=O)NC(=N1)NC(C)C	C8H15N5O	197.1277	2163-68-0
Atrazine-deisopropyl	CCNC1=NC(=NC(=N1)N)Cl	C5H8ClN5	173.0468	1007-28-9
Azoxystrobin	COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC	C22H17N3O5	403.1168	215934-32-0
Beflubutamid	CCC(C(=O)NCC1=CC=CC=C1)OC2=CC(=C(C=C2)F)C(F)(F)F	C18H17F4NO2	355.1195	113614-08-7
Bixafen	CN1C=C(C(=N1)C(F)F)C(=O)NC2=C(C=C(C=C2)F)C3=CC(=C(C=C3)Cl)Cl	C18H12Cl2F3N3O	413.0310	581809-46-3
Boscalid	C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl	C18H12Cl2N2O	342.0327	188425-85-6
Bromacil	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	C9H13BrN2O2	260.0160	314-40-9
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	C15H12N2O	236.0950	298-46-4

Table A3 – (continued) Detailed list of compound from the retention time prediction set, with SMILES identifier, chemical formula, monoisotopic mass, and CAS number

Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Carbaryl	CNC(=O)OC1=CC=CC2=CC=CC=C21	C12H11NO2	201.0790	51274-03-4
Carbendazim	COC(=O)NC1=NC2=CC=CC=C2N1	C9H9N3O2	191.0695	63278-70-6
Carbetamide	CCNC(=O)C(C)OC(=O)NC1=CC=CC=C1	C12H16N2O3	236.1161	16118-49-3
Carbofuran	CC1(CC2=C(O1)C(=CC=C2)OC(=O)NC)C	C12H15NO3	221.1052	1563-66-2
Chlorantraniliprole	CC1=CC(=CC(=C1NC(=O)C2=CC(=NN2C3=C(C=CC=N3)Cl)Br)C(=O)NC)Cl	C18H14BrCl2N5O2	480.9708	500008-45-7
Chloridazon	C1=CC=C(C=C1)N2C(=O)C(=C(C=N2)N)Cl	C10H8ClN3O	221.0356	1698-60-8
Chlorpyrifos	CCOP(=S)(OCC)OC1=NC(=C(C=C1Cl)Cl)Cl	C9H11Cl3NO3PS	348.9263	39475-55-3
Chlortoluron	CC1=C(C=C(C=C1)NC(=O)N(C)C)Cl	C10H13ClN2O	212.0716	15545-48-9
Clothianidin	CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl	C6H8ClN5O2S	249.0087	205510-53-8
Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	C18H21NO3	299.1521	76-57-3
Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)CO)O)C	C21H28O5	360.1937	53-06-5
Cotinine	CN1C(CCC1=O)C2=CN=CC=C2	C10H12N2O	176.0950	486-56-6
Cyprodinil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C3CC3	C14H15N3	225.1266	121552-61-2
Diazinon	CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C	C12H21N2O3PS	304.1011	30583-38-1
Dichlorprop	CC(C(=O)O)OC1=C(C=C(C=C1)Cl)Cl	C9H8Cl2O3	233.9851	120-36-5
Diclofenac	C1=CC=C(C(=C1)CC(=O)O)NC2=C(C=CC=C2Cl)Cl	C14H11Cl2NO2	295.0167	15307-86-5
Dimethenamid	CC1=CSC(=C1N(C(C)COC)C(=O)CCl)C	C12H18ClNO2S	275.0747	87674-68-8
Dimethomorph	COC1=C(C=C(C=C1)C(=CC(=O)N2CCOCC2)C3=CC=C(C=C3)Cl)OC	C21H22ClNO4	387.1237	110488-70-5
Dimethyldithiophosphate	COP(=S)(OC)S	C2H7O2PS2	157.9625	756-80-9
Diuron	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl	C9H10Cl2N2O	232.0170	102962-29-8
Estradiol-2-hydroxy	CC12CCC3C(C1CCC2O)CCC4=CC(=C(C=C34)O)O	C18H24O3	288.1725	362-05-0
Estrone	CC12CCC3C(C1CCC2=O)CCC4=C3C=CC(=C4)O	C18H22O2	270.1620	53-16-7
Estrone-2-hydroxy	CC12CCC3C(C1CCC2=O)CCC4=CC(=C(C=C34)O)O	C18H22O3	286.1569	362-06-1
Ethidimuron	CCS(=O)(=O)C1=NN=C(S1)N(C)C(=O)NC	C7H12N4O3S2	264.0351	30043-49-3
Fenamidone	CC1(C(=O)N(C(=N1)SC)NC2=CC=CC=C2)C3=CC=CC=C3	C17H17N3OS	311.1092	161326-34-7
Fenpropidine	CC(CC1=CC=C(C=C1)C(C)(C)C)CN2CCCCC2	C19H31N	273.2456	67306-00-7

Table A3 – (continued) Detailed list of compound from the retention time prediction set, with SMILES identifier, chemical formula, monoisotopic mass, and CAS number

Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Fenpropimorph	CC1CN(CC(O1)C)CC(C)CC2=CC=C(C=C2)C(C)(C)C	C20H33NO	273.2456	67564-91-4
Flonicamid	C1=CN=CC(=C1C(F)(F)F)C(=O)NCC#N	C9H6F3N3O	229.0463	158062-67-0
Flufenacet	CC(C)N(C1=CC=C(C=C1)F)C(=O)COC2=NN=C(S2)C(F)(F)F	C14H13F4N3O2S	363.0665	142459-58-3
Fluoxetine	CNCCC(C1=CC=CC=C1)OC2=CC=C(C=C2)C(F)(F)F	C17H18F3NO	309.1340	57226-07-0
Fluroxypyr	C(C(=O)O)OC1=NC(=C(C(=C1Cl)N)Cl)F	C7H5Cl2FN2O3	253.9661	69377-81-7
Flurtamone	CNC1=C(C(=O)C(O1)C2=CC=CC=C2)C3=CC(=CC=C3)C(F)(F)F	C18H14F3NO2	333.0977	96525-23-4
Foramsulfuron	CN(C)C(=O)C1=C(C=C(C=C1)NC=O)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC	C17H20N6O7S	452.1114	173159-57-4
Fosthiazate	CCO[P](=O)(SC(C)CC)N1CCSC1=O	C9H18NO3PS2	283.0466	98886-44-3
Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	C21H30O5	362.2093	50-23-7
Hydroxyindoleacetic acid	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	C10H9NO3	191.0582	113303-91-6
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	C13H18O2	206.1307	58560-75-1
Imazamethabenz-methyl	CC1=CC(=C(C=C1)C(=O)OC)C2=NC(C(=O)N2)(C)C(C)C	C16H20N2O3	288.1474	81405-85-8
Imazamox	CC(C)C1(C(=O)NC(=N1)C2=C(C=C(C=N2)COC)C(=O)O)C	C15H19N3O4	305.1376	114311-32-9
Imazaquin	CC(C)C1(C(=O)NC(=N1)C2=NC3=CC=CC=C3C=C2C(=O)O)C	C17H17N3O3	311.1270	81335-37-7
Imidacloprid	C1CN(C(=N[N+](=O)[O-])N1)CC2=CN=C(C=C2)Cl	C9H10ClN5O2	255.0523	138261-41-3
Iodosulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=C(C=CC(=C2)I)C(=O)OC	C14H13IN5NaO6S	528.9529	144550-36-7
Iprodione	CC(C)NC(=O)N1CC(=O)N(C1=O)C2=CC(=CC(=C2)Cl)Cl	C13H13Cl2N3O3	329.0334	36734-19-7
Irgarol	CC(C)(C)NC1=NC(=NC(=N1)NC2CC2)SC	C11H19N5S	253.1361	28159-98-0
Isoproturon	CC(C)C1=CC=C(C=C1)NC(=O)N(C)C	C12H18N2O	206.1419	34123-59-6
Isoproturon-didemethyl	CC(C)C1=CC=C(C=C1)NC(=O)N	C10H14N2O	178.1106	56046-17-4
Isoxaben	CCC(C)(CC)C1=NOC(=C1)NC(=O)C2=C(C=CC=C2OC)OC	C18H24N2O4	332.1736	82558-50-7
Isoxaflutole	CS(=O)(=O)C1=C(C=CC(=C1)C(F)(F)F)C(=O)C2=C(ON=C2)C3CC3	C15H12F3NO4S	359.0439	141112-29-0
Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	C20H32O4	336.2301	71160-24-2
Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N	C25H40N2O6S	496.2607	73836-78-9
Linuron	CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC	C9H10Cl2N2O2	248.0119	56645-87-5
Mesosulfuron-methyl	COC1=CC(=NC(=N1)NC(=O)NS(=O)(=O)C2=C(C=CC(=C2)CNS(=O)(=O)C)C(=O)OC)OC	C17H21N5O9S2	503.0781	208465-21-8

Table A3 – (continued) Detailed list of compound from the retention time prediction set, with SMILES identifier, chemical formula, monoisotopic mass, and CAS number

Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Mesotrione	CS(=O)(=O)C1=CC(=C(C=C1)C(=O)C2C(=O)CCCC2=O)[N+](=O)[O-]	C14H13NO7S	339.0413	104206-82-8
Metalaxyl	CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)COC	C15H21NO4	279.1471	57837-19-1
Metamitron	CC1=NN=C(C(=O)N1N)C2=CC=CC=C2	C10H10N4O	202.0855	41394-05-2
Metazachlor	CC1=C(C(=CC=C1)C)N(CN2C=CC=N2)C(=O)CCl	C14H16ClN3O	277.0982	67129-08-2
Methabenzthiazuron	CNC(=O)N(C)C1=NC2=CC=CC=C2S1	C10H11N3OS	221.0623	18691-97-9
Metobromuron	CN(C(=O)NC1=CC=C(C=C1)Br)OC	C9H11BrN2O2	258.0004	3060-89-7
Metolachlor	CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C	C15H22ClNO2	283.1339	55762-76-0
Metosulam	CC1=C(C(=C(C=C1)Cl)NS(=O)(=O)C2=NN3C(=CC(=NC3=N2)OC)OC)Cl	C14H13Cl2N5O4S	417.0065	139528-85-1
Metribuzine	CSC1=NN=C(C(=O)N1N)C(C)(C)C	C8H14N4OS	214.0888	21087-64-9
Metsulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(=O)OC	C14H15N5O6S	381.0743	74223-64-6
Nicosulfuron	CN(C)C(=O)C1=C(N=CC=C1)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC	C15H18N6O6S	410.1009	111991-09-4
Nicotine	CN1CCCC1C2=CN=CC=C2	C10H14N2	162.1157	551-13-3
Oryzalin	CCCN(CCC)C1=C(C=C(C=C1[N+](=O)[O-])S(=O)(=O)N)[N+](=O)[O-]	C12H18N4O6S	346.0947	19044-88-3
Paclobutrazol	CC(C)(C)C(C(CC1=CC=C(C=C1)Cl)N2C=NC=N2)O	C30H40Cl2N6O2	586.2590	76738-62-0
Paracetamol	CC(=O)NC1=CC=C(C=C1)O	C8H9NO2	151.0633	8055-08-1
Paroxetine	C1CNCC(C1C2=CC=C(C=C2)F)COC3=CC4=C(C=C3)OCO4	C19H20FNO3	329.1427	63952-24-9
Pencycuron	C1CCC(C1)N(CC2=CC=C(C=C2)Cl)C(=O)NC3=CC=CC=C3	C19H21ClN2O	328.1342	66063-05-6
Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	C17H19NO3	285.1365	147030-08-8
Pirimicarb	CC1=C(N=C(N=C1OC(=O)N(C)C)N(C)C)C	C11H18N4O2	238.1430	23103-98-2
Pravastatin	CCC(C)C(=O)OC1CC(C=C2C1C(C(C=C2)C)CCC(CC(CC(=O)O)O)O)O	C23H36O7	424.2461	81093-37-0
Prochloraz	CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=CN=C2	C15H16Cl3N3O2	375.0308	67747-09-5
Progesterone	CC(=O)C1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	C21H30O2	314.2246	257630-50-5
Propachlor	CC(C)N(C1=CC=CC=C1)C(=O)CCl	C11H14ClNO	211.0764	1918-16-7
Propamocarb	CCCOC(=O)NCCCN(C)C	C9H20N2O2	188.1525	24579-73-5
Propiconazole	CCCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl	C15H17Cl2N3O2	341.0698	75881-82-2
Propoxycarbazone	CCCOC1=NN(C(=O)N1C)C(=O)NS(=O)(=O)C2=CC=CC=C2C(=O)OC	C15H17N4NaO7S	420.0716	181274-15-7

Table A3 – (continued) Detailed list of compound from the retention time prediction set, with SMILES identifier, chemical formula, monoisotopic mass, and CAS number

Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Propyzamide	CC(C)(C#C)NC(=O)C1=CC(=CC(=C1)Cl)Cl	C12H11Cl2NO	255.0218	11097-11-3
Prostaglandin D2	CCCCCC(C=CC1C(C(CC1=O)O)CC=CCCCC(=O)O)O	C20H32O5	352.2250	41598-07-6
Prostaglandin E2	CCCCCC(C=CC1C(CC(=O)C1CC=CCCCC(=O)O)O)O	C20H32O5	352.2250	363-24-6
Prostaglandin F2a	CCCCCC(C=CC1C(CC(C1CC=CCCCC(=O)O)O)O)O	C20H34O5	354.2406	13535-33-6
Prosulfuron	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2CCC(F)(F)F	C15H16F3N5O4S	419.0875	94125-34-5
Pymetrozine	CC1=NNC(=O)N(C1)N=CC2=CN=CC=C2	C10H11N5O	217.0964	123312-89-0
Pyraclostrobin	COC(=O)N(C1=CC=CC=C1COC2=NN(C=C2)C3=CC=C(C=C3)Cl)OC	C19H18ClN3O4	387.0986	175013-18-0
Pyrimethanil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C	C12H13N3	199.1109	53112-28-0
Pyroxsulam	COC1=CC(=NC2=NC(=NN12)NS(=O)(=O)C3=C(C=CN=C3OC)C(F)(F)F)OC	C14H13F3N6O5S	434.0620	422556-08-9
Quinmerac	CC1=CC2=C(C(=C(C=C2)Cl)C(=O)O)N=C1	C11H8ClNO2	221.0244	90717-03-6
Sertraline	CNC1CCC(C2=CC=CC=C12)C3=CC(=C(C=C3)Cl)Cl	C17H17Cl2N	305.0738	79559-97-0
Simazine	CCNC1=NC(=NC(=N1)Cl)NCC	C7H12ClN5	201.0781	119603-94-0
Solanidine	CC1CCC2C(C3C(N2C1)CC4C3(CCC5C4CC=C6C5(CCC(C6)O)C)C)C	C27H43NO	397.3345	80-78-4
Spiroxamine	CCCN(CC)CC1COC2(CCC(CC2)C(C)(C)C)O1	C18H35NO2	297.2668	118134-30-8
Sulcotrione	CS(=O)(=O)C1=CC(=C(C=C1)C(=O)C2C(=O)CCCC2=O)Cl	C14H13ClO5S	328.0172	99105-77-8
Tebuconazole	CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O	C16H22ClN3O	307.1451	80443-41-0
Tebutame	CC(C)N(CC1=CC=CC=C1)C(=O)C(C)(C)C	C15H23NO	233.1780	35256-85-0
Terbuthylazine	CCNC1=NC(=NC(=N1)Cl)NC(C)(C)C	C9H16ClN5	229.1094	5915-41-3
Terbutryne	CCNC1=NC(=NC(=N1)SC)NC(C)(C)C	C10H19N5S	241.1361	886-50-0
Tertbutylazine-2-hydroxy	CCNC1=NC(=O)NC(=N1)NC(C)(C)C	C9H17N5O	211.1433	66753-07-9
Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	C19H28O2	288.2089	58-22-0
Thiacloprid	C1CSC(=NC#N)N1CC2=CN=C(C=C2)Cl	C10H9ClN4S	252.0236	111988-49-9
Thiamethoxam	CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl	C8H10ClN5O3S	291.0193	153719-23-4
Thifensulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=C(SC=C2)C(=O)OC	C12H13N5O6S2	387.0307	79277-27-3
Triadimenol	CC(C)(C)C(C(N1C=NC=N1)OC2=CC=C(C=C2)Cl)O	C14H18ClN3O2	295.1088	55219-65-3
Triazoxide	C1=CC2=C(C=C1Cl)[N+](=NC(=N2)N3C=CN=C3)[O-]	C10H6ClN5O	247.0261	72459-58-6

Table A3 – (continued) Detailed list of compound from the retention time prediction set, with SMILES identifier, chemical formula, monoisotopic mass, and CAS number

Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Triclopyr	C1=C(C(=NC(=C1Cl)Cl)OCC(=O)O)Cl	C7H4Cl3NO3	254.9257	55335-06-3
Triflusulfuron-methyl	CC1=C(C(=CC=C1)C(=O)OC)S(=O)(=O)NC(=O)NC2=NC(=NC(=N2)OCC(F)(F)F)N(C)C	C17H19F3N6O6S	492.1039	126535-15-7
Trinexapac-ethyl	CCOC(=O)C1CC(=O)C(=C(C2CC2)O)C(=O)C1	C13H16O5	252.0998	95266-40-3
Triticonazole	CC1(CCC(=CC2=CC=C(C=C2)Cl)C1(CN3C=NC=N3)O)C	C17H20ClN3O	317.1295	131983-72-7
Tritosulfuron	COC1=NC(=NC(=N1)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(F)(F)F)C(F)(F)F	C13H9F6N5O4S	445.0279	142469-14-5
Venlafaxine	CN(C)CC(C1=CC=C(C=C1)OC)C2(CCCCC2)O	C17H27NO2	277.2042	93413-69-5

2. Appendix 2. Supporting information – Chapter III

2.1. Table A1 – Standard compounds form and suppliers

Table A1 – Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Arachidonic Acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	Bertin	Powder
Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	Bertin	Powder
Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N	Bertin	Powder
Prostaglandin D2	CCCCCC(C=CC1C(C(CC1=O)O)CC=CCCCC(=O)O)O	Bertin	Powder
Prostaglandin E2	CCCCCC(C=CC1C(CC(=O)C1CC=CCCCC(=O)O)O)O	Bertin	Powder
Prostaglandin F2a	CCCCCC(C=CC1C(CC(C1CC=CCCCC(=O)O)O)O)O	Bertin	Powder
Prostaglandin J2	CCCCCC(C=CC1C(C=CC1=O)CC=CCCCC(=O)O)O	Bertin	Powder
Acetochlor	CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C	LGC	Powder
Acetylsalicylic acid	CC(=O)OC1=CC=CC=C1C(=O)O	LGC	Powder
Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	LGC	Powder
Carbendazim	COC(=O)NC1=NC2=CC=CC=C2N1	LGC	Powder
Clothianidin	CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl	LGC	Powder
Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)COO)O)C	LGC	Powder
Dimethyldithiophosphate	COP(=S)(OC)S	LGC	Powder
Estrone	CC12CCC3C(C1CCC2=O)CCC4=C3C=CC(=C4)O	LGC	Powder
Fluoxetine	CNCCC(C1=CC=CC=C1)OC2=CC=C(C=C2)C(F)(F)F	LGC	1.0 mg/mL in MeOH
Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	LGC	Powder
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	LGC	Powder
Paracetamol	CC(=O)NC1=CC=C(C=C1)O	LGC	Powder
Paroxetine	C1CNCC(C1C2=CC=C(C=C2)F)COC3=CC4=C(C=C3)OCO4	LGC	1.0 mg/mL in MeOH
Progesterone	CC(=O)C1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	LGC	Powder
Sertraline	CNC1CCC(C2=CC=CC=C12)C3=CC(=C(C=C3)Cl)Cl	LGC	1.0 mg/mL in MeOH
Tebuconazole	CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O	LGC	Powder
Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	LGC	Powder
Thiacloprid	C1CSC(=NC#N)N1CC2=CN=C(C=C2)Cl	LGC	Powder
Venlafaxine	CN(C)CC(C1=CC=C(C=C1)OC)C2(CCCCC2)O	LGC	Powder
Aflatoxin B1	COC1=C2C3=C(C(=O)CC3)C(=O)OC2=C4C5C=COC5OC4=C1	Sigma Aldrich	Powder
Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	Sigma Aldrich	Powder
Hydroxyindoleacetic acid	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	Sigma Aldrich	Powder
Ketoprofen	CC(C1=CC(=CC=C1)C(=O)C2=CC=CC=C2)C(=O)O	Sigma Aldrich	Powder
Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	Sigma Aldrich	Powder
Pravastatin	CCC(C)C(=O)OC1CC(C=C2C1C(C(C=C2)C)CCC(CC(CC(=O)O)O)O)O	Sigma Aldrich	Powder
Solanidine	CC1CCC2C(C3C(N2C1)CC4C3(CCC5C4CC=C6C5(CCC(C6)O)C)C)C	Sigma Aldrich	Powder
2-Phenylphenol	C1=CC=C(C=C1)C2=CC=CC=C2O	VWR	Powder
Aminobenzimidazole	C1=CC=C2C(=C1)NC(=N2)N	VWR	Powder

Table A1 – (continued) Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Azoxystrobin	COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC	VWR	Powder
Boscalid	C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl	VWR	Powder
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	VWR	Powder
Chlorpyrifos	CCOP(=S)(OCC)OC1=NC(=C(C=C1Cl)Cl)Cl	VWR	Powder
Cotinine	CN1C(CCC1=O)C2=CN=CC=C2	VWR	Powder
Cyprodinil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C3CC3	VWR	Powder
Diazinon	CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C	VWR	Powder
Diclofenac	C1=CC=C(C(=C1)CC(=O)O)NC2=C(C=CC=C2Cl)Cl	VWR	Powder
Imidacloprid	C1CN(C(=N[N+](=O)[O-])N1)CC2=CN=C(C=C2)Cl	VWR	Powder
Malathion	CCOC(=O)CC(C(=O)OCC)SP(=S)(OC)OC	VWR	Powder
Nicotine	CN1CCCC1C2=CN=CC=C2	VWR	Powder
Prochloraz	CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=CN=C2	VWR	Powder
Propiconazole	CCCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl	VWR	Powder
Thiamethoxam	CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl	VWR	Powder
Triclosan	C1=CC(=C(C=C1Cl)O)OC2=C(C=C(C=C2)Cl)Cl	VWR	Powder

2.2. Table A2 – Standard compounds physical-chemical characteristics

Table A2 – Standard compounds identifiers and physical-chemical characteristics (monoisotopic mass, retention time (Rt), octanol-water partition coefficient (logP))

Compound name	Chemical formula	Monoisotopic mass (Da)	Observed ion	Rt (min)	logP	CAS
2-Phenylphenol	C12H10O	170.0732	$[M-H]^-$	30.19	3.28	90-43-7
Acetochlor	C14H20ClNO2	269.1183	$[M-H]^-$	40.57	4.14	123113-74-6
Acetylsalicylic acid	C9H8O4	180.0423	$[M-H]^-$	8.65	1.24	50-78-2
Aflatoxin B1	C17H12O6	312.0634	$[M+H]^+$	17.52	1.73	27261-02-5
Aminobenzimidazole	C7H7N3	133.0640	$[M+H]^+$	4.74	0.91	934-32-7
Androstenedione	C19H26O2	286.1933	$[M+H]^+$	31.50	2.75	63-05-8
Arachidonic Acid	C20H32O2	304.2402	$[M-H]^-$	47.00	6.99	93444-49-6
Azoxystrobin	C22H17N3O5	403.1168	$[M+H]^+$	38.03	2.64	215934-32-0
Boscalid	C18H12Cl2N2O	342.0327	$[M+H]^+$	38.00	2.96	188425-85-6
Carbamazepine	C15H12N2O	236.0950	$[M+H]^+$	18.01	2.45	298-46-4
Carbendazim	C9H9N3O2	191.0695	$[M+H]^+$	5.69	1.52	63278-70-6
Chlorpyrifos	C9H11Cl3NO3PS	348.9263	$[M+H]^+$	45.53	4.70	39475-55-3
Clothianidin	C6H8ClN5O2S	249.0087	$[M+H]^+$	7.99	0.73	205510-53-8
Codeine	C18H21NO3	299.1521	$[M+H]^+$	5.12	1.39	76-57-3
Cortisone	C21H28O5	360.1937	$[M+H]^+$	16.12	1.47	53-06-5
Cotinine	C10H12N2O	176.0950	$[M+H]^+$	4.31	0.07	486-56-6
Cyprodinil	C14H15N3	225.1266	$[M+H]^+$	33.22	4.00	121552-61-2
Diazinon	C12H21N2O3PS	304.1011	$[M+H]^+$	43.38	3.81	30583-38-1
Diclofenac	C14H11Cl2NO2	295.0167	$[M-H]^-$	39.59	4.51	15307-86-5
Dimethyldithiophosphate	C2H7O2PS2	157.9625	$[M-H]^-$	2.95	0.63	756-80-9
Estrone	C18H22O2	270.1620	$[M+H]^+$	31.60	3.13	53-16-7
Fluoxetine	C17H18F3NO	309.1340	$[M+H]^+$	23.71	4.05	57226-07-0
Hydrocortisone	C21H30O5	362.2093	$[M+H]^+$	15.86	1.61	50-23-7
Hydroxyindoleacetic acid	C10H9NO3	191.0582	$[M-H]^-$	5.71	1.41	113303-91-6
Ibuprofen	C13H18O2	206.1307	$[M+H]^+$	39.94	3.97	58560-75-1
Imidacloprid	C9H10ClN5O2	255.0523	[M+H]+	8.57	0.57	138261-41-3
Ketoprofen	C16H14O3	254.0943	$[M+H]^+$	28.13	3.12	22071-15-4
Leukotriene B4	C20H32O4	336.2301	$[M-H]^-$	39.52	4.10	71160-24-2
Leukotriene D4	C25H40N2O6S	496.2607	$[M-H]^-$	33.04	1.40	73836-78-9
Malathion	C10H19O6PS2	330.0361	$[M+H]^+$	40.81	2.89	121-75-5
Nicotine	C10H14N2	162.1157	$[M+H]^+$	3.37	1.17	551-13-3
Paracetamol	C8H9NO2	151.0633	$[M+H]^+$	4.98	0.31	8055-08-1
Paroxetine	C19H20FNO3	329.1427	$[M+H]^+$	18.34	1.23	63952-24-9
Piperine	C17H19NO3	285.1365	$[M+H]^+$	36.42	2.78	147030-08-8
Pravastatin	C23H36O7	424.2461	$[M+H]^+$	20.50	1.65	81093-37-0
Prochloraz	C15H16Cl3N3O2	375.0308	$[M+H]^+$	38.74	3.78	67747-09-5
Progesterone	C21H30O2	314.2246	$[M+H]^+$	42.10	3.87	257630-50-5
Propiconazole	C15H17Cl2N3O2	341.0698	$[M+H]^+$	41.73	3.72	75881-82-2
Prostaglandin D2	C20H32O5	352.2250	$[M-H]^-$	27.60	3.23	41598-07-6
Prostaglandin E2	C20H32O5	352.2250	$[M-H]^-$	26.50	2.82	363-24-6
Prostaglandin F2a	C20H34O5	354.2406	$[M-H]^-$	25.60	2.61	13535-33-6
Prostaglandin J2	C20H30O4	334.2144	$[M-H]^-$	26.54	3.60	60203-57-8
Sertraline	C17H17Cl2N	305.0738	$[M+H]^+$	24.34	5.10	79559-97-0
Solanidine	C27H43NO	397.3345	$[M+H]^+$	24.54	4.88	80-78-4
Tebuconazole	C16H22ClN3O	307.1451	$[M+H]^+$	39.36	3.70	80443-41-0
Testosterone	C19H28O2	288.2089	$[M+H]^+$	28.90	3.32	58-22-0
Thiacloprid	C10H9ClN4S	252.0236	$[M+H]^+$	12.24	1.25	111988-49-9
Thiamethoxam	C8H10ClN5O3S	291.0193	$[M+H]^+$	6.97	1.52	153719-23-4
Triclosan	C12H7Cl3O2	287.9512	$[M-H]^-$	43.79	4.76	3380-34-5
Venlafaxine	C17H27NO2	277.2042	$[M+H]^+$	9.84	0.43	93413-69-5

2.3. Table A3 – Preselection: Recovery, repeatability and matrix effect of all sample preparation methods on individual compounds

Table A3 – Preselection: Recovery, repeatability (recovery coefficient of variation CV) and matrix effect (ME) at 20 and 150 ng/mL of all sample preparation methods on individual compounds

		PROTEIN PRECIPITATION				PHOSPHOLIPID AND PROTEIN REMOVAL PLATES							
		PPT				Phree ACN				Phree MeOH			
	Component	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)
POSITIVE IONISATION	2-Aminobenzimidazole	103.6	13.2	73.6	77.0	113.1	6.3	3.2	-95.3	115.4	3.0	41.6	-32.0
	4-Androstene-3,17-dione	107.1	6.0	60.8	58.9	123.7	2.6	31.9	26.2	106.8	3.8	29.0	41.2
	Acetochlor	106.8	2.1	78.1	75.7	69.3	3.1	0.8	12.2	45.1	7.9	76.9	32.2
	Aflatoxin B1	55.5	5.5	92.4	95.9	111.4	4.5	16.5	28.0	95.4	9.3	72.6	45.3
	Azoxystrobin	108.2	2.8	75.0	72.5	106.1	3.9	-25.3	14.7	118.0	2.9	82.2	34.8
	Boscalid	105.7	4.6	61.0	61.4	102.9	3.7	25.0	30.7	111.0	5.5	78.2	45.3
	Carbamazepine	107.5	2.5	37.8	47.0	106.7	2.2	18.5	-56.6	112.2	5.1	51.0	-13.2
	Carbendazim	110.4	1.8	58.9	74.1	112.4	1.3	-1.4	-48.4	111.6	5.4	31.3	-18.5
	Chlorpyrifos	109.0	2.7	92.9	92.4	63.6	3.8	91.8	76.6	42.8	3.6	96.5	89.8
	Clothianidin	100.8	0.6	55.8	61.6	107.0	3.0	49.3	-16.1	121.1	6.7	67.0	18.2
	Codeine	129.0	10.9	57.8	57.8	87.8	16.0	53.6	-13.8	112.5	18.2	70.6	4.6
	Cortisone	105.1	3.6	86.5	85.0	107.6	3.0	85.1	21.6	101.4	7.0	92.1	46.3
	Cotinine	110.4	9.9	45.2	42.3	85.0	0.9	19.4	13.0	94.1	24.5	65.4	28.0
	Cyprodinil	106.0	3.2	38.7	34.2	139.7	41.3	75.3	2.3	56.6	7.3	63.7	-31.1
	Diazinon	106.2	11.8	87.2	80.8	49.8	10.8	38.7	44.8	51.4	3.3	92.0	74.9
	Estrone	112.5	4.9	61.8	59.6	119.5	3.7	45.8	32.6	119.2	6.5	80.4	58.6
	Fluoxetine	104.8	2.7	94.3	87.7	117.0	11.8	24.4	19.0	75.2	10.4	80.3	28.4
	Hydrocortisone	107.2	8.1	90.8	90.7	104.5	3.4	80.8	55.3	93.4	16.7	53.4	25.2
	Imidacloprid	107.8	2.2	49.1	53.0	109.5	2.2	10.8	-9.1	115.8	6.2	67.3	18.7
	Ketoprofen	107.0	2.7	94.7	95.7	118.9	3.2	22.3	-61.3	107.5	6.3	65.4	-14.2
	Malathion	93.7	42.5	96.2	95.0	79.4	0.6	14.6	0.8	53.8	8.4	74.9	9.4
	Nicotine					84.7	3.0	80.5	55.9	61.6	45.9	94.9	89.4
	Paracetamol	100.7	8.6	80.5	81.3	131.3	2.7	83.4	33.9	107.0	8.4	81.6	56.8
	Paroxetine	132.4	22.1	93.6	85.8	131.8	34.8	25.5	23.9	90.6	6.7	82.0	34.8
	Piperine	156.2	12.4	93.4	91.1	114.9	20.0	55.1	67.4	107.5	15.0	88.7	82.0
	Pravastatin					83.9	16.1	23.5	29.9	110.3	18.9	76.1	41.0
	Prochloraz	100.5	2.2	65.5	66.9	115.2	13.0	55.6	37.6	100.1	16.1	62.7	37.8
	Progesterone	132.5	33.5	61.1	57.1	125.4	2.6	63.0	40.5	97.8	9.7	50.7	55.5
	Propiconazole	107.3	5.1	62.4	60.2	113.0	4.7	7.5	42.7	108.7	5.1	80.7	50.2
	Sertraline	79.5	15.8	89.5	83.4	119.9	5.9	45.7	29.4	84.7	9.8	4.4	20.5
	Solanidine	195.0	30.3	86.6	80.0	112.6	7.6	45.5	42.1	104.2	10.2	15.1	39.8
	Tebuconazole	108.5	1.8	60.4	59.8	103.0	1.1	16.9	21.7	110.0	5.3	42.8	46.3
	Testosterone	95.4	47.7	65.1	60.3	115.3	4.7	49.1	17.1	100.7	3.0	72.4	27.2
	Thiacloprid	115.6	5.0	66.0	69.4	108.6	9.2	27.8	29.4	111.5	7.7	37.1	40.2
	Thiamethoxam	111.8	2.2	77.6	77.8	109.5	0.5	60.6	34.1	114.0	6.6	83.5	56.6
	Venlafaxine	110.9	5.8	60.8	55.0	104.7	9.6	21.5	20.9	84.1	9.1	81.2	31.0
NEGATIVE IONISATION	2-Phenylphenol							13.1	28.3			95.3	44.4
	5-Hydroxyindole-3-acetic acid			100.0	98.9	110.3	19.6	36.1	47.9	96.1	29.7	-0.5	46.4
	Acetylsalicylic acid							16.8	27.7			39.9	11.5
	Arachidonic acid			73.2	59.9								
	Diclofenac	102.8	5.4	96.3	97.1	109.7	2.8	39.9	26.6	114.8	5.5	47.2	10.2
	Dimethyldithiophosphate	118.3	3.5	-2.6	-3.2	73.2	12.6	43.8	27.3	85.9	11.6	15.1	40.3
	Ibuprofen			100.0	98.9	89.4	19.6	21.3	5.3			-4.8	24.9
	Leukotriene B4			96.6	98.1	137.9	19.8	31.4	20.8	160.2	5.3	69.7	76.2
	Leukotriene D4			99.2	99.7	101.9	4.9	76.0	85.4			65.2	82.2
	Prostaglandin D2					81.7	16.7	65.8	62.4			42.9	54.5
	Prostaglandin E2	160.6	67.5	99.0	99.4	87.8	6.4	78.8	74.0	101.9	17.0	70.9	71.6
	Prostaglandin F2a	57.5	45.5	99.0	99.5	109.8	4.8	79.8	79.1	83.8	12.5	71.1	79.8
	Prostaglandin J2					81.5	11.2	19.1	32.8	71.1	9.5	-78.0	-7.7
	Triclosan			-9.7	81.7	137.3	32.8	27.7	50.5	87.5	19.2	61.7	68.1

Table A3 – (continued) Preselection: Recovery, repeatability (recovery coefficient of variation CV) and matrix effect (ME) at 20 and 150 ng/mL of all sample preparation methods on individual compounds

		PHOSPHOLIPID AND PROTEIN REMOVAL PLATES											
		PLD				Ostro				Prime HLB			
	Component	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)
POSITIVE IONISATION	2-Aminobenzimidazole	104.0	4.7	35.4	-0.3	95.5	14.2	66.1	70.6	121.3	4.7	-32.9	-4.3
	4-Androstene-3,17-dione	114.5	3.9	58.7	28.1	93.6	12.0	48.4	29.1	97.4	6.7	69.0	25.8
	Acetochlor	104.4	3.5	72.6	39.0	100.7	8.3	78.3	58.4	70.8	7.0	46.0	31.4
	Aflatoxin B1	137.4	2.8	50.8	38.8	97.0	14.6	64.9	75.6	88.8	7.4	20.9	33.0
	Azoxystrobin	127.8	2.3	63.8	8.2	100.9	8.7	54.6	29.5	103.9	4.6	29.4	13.1
	Boscalid	136.7	1.1	65.9	25.1	86.7	8.5	57.5	27.2	83.8	6.3	45.7	35.1
	Carbamazepine	124.2	1.4			97.6	7.7	7.6	1.9	106.2	3.1	-43.3	-46.7
	Carbendazim	124.1	3.1	27.2	15.0	98.9	8.6	57.9	65.4	101.5	3.1	-21.6	17.6
	Chlorpyrifos	98.8	7.4	94.7	93.0	109.3	6.5	99.2	99.1	62.4	11.7	88.0	88.5
	Clothianidin	120.7	1.5	44.0	-2.8	100.1	4.1	62.6	67.6	121.9	4.7	0.8	0.6
	Codeine	107.5	19.0	36.5	-34.0	102.6	14.4	80.5	79.3	49.5	4.7	7.5	-19.9
	Cortisone	126.5	4.3	46.9	15.5	92.1	12.5	64.3	72.8	99.3	2.7	41.4	10.9
	Cotinine	90.4	9.9	52.7	48.9	83.3	13.6			146.1	8.5	28.2	-4.2
	Cyprodinil	126.5	7.0	52.4	-7.6	184.2	10.3	96.2	93.2	182.9	9.6	67.1	56.2
	Diazinon	90.0	1.0	95.5	88.7	104.8	8.3	93.4	87.2	46.0	4.2	51.6	53.8
	Estrone	108.0	12.3	61.4	47.2	101.6	11.2	38.9	19.2	106.7	8.6	36.3	44.4
	Fluoxetine	123.5	0.6	59.8	15.3	100.4	10.1	95.9	94.9	8.8	12.6	48.7	33.7
	Hydrocortisone	115.9	10.4	44.1	14.7	81.1	10.0	71.5	74.3	71.0	30.4	24.9	-13.2
	Imidacloprid	122.1	2.5	44.6	-2.3	99.6	4.7	64.1	66.6	115.2	2.7	8.1	-9.1
	Ketoprofen	131.5	5.0	33.2	-17.3	96.6	10.5	13.1	-5.4	73.8	13.9	-13.1	-21.2
	Malathion	120.9	0.4	76.1	37.7	120.1	4.9	78.5	52.1	67.9	4.4	50.9	32.7
	Nicotine	60.7	11.8	83.8	66.3					26.3	15.6		
	Paracetamol	130.3	7.7	79.5	56.9	96.8	9.7	82.0	85.4	74.1	5.1	47.4	64.6
	Paroxetine	118.8	3.7	55.6	9.2	99.9	9.0	97.3	95.7	1.5	35.7	37.6	28.3
	Piperine	102.5	13.3	56.7	14.9	124.7	17.7	79.3	78.1	153.1	13.7	66.7	73.4
	Pravastatin	98.2	2.5	-24.2	-84.1	87.2	9.1	88.1	91.5	84.4	7.5	-20.8	-33.1
	Prochloraz	120.7	2.1	66.3	25.3	93.7	22.7	93.8	93.6	124.1	3.6	64.2	60.7
	Progesterone	117.7	2.3	72.4	48.8	165.6	26.2	78.2	68.2	113.0	9.0	61.6	55.9
	Propiconazole	121.2	0.8	65.8	21.8	107.9	7.9	85.1	72.3	113.3	4.2	59.0	48.8
	Sertraline	100.6	5.6	70.3	61.5	90.7	12.0	97.8	96.9	82.1	11.2	61.7	51.9
	Solanidine	104.9	4.4	50.9	8.0	106.3	14.2	93.6	85.2	99.3	23.0	44.3	33.6
	Tebuconazole	122.3	2.2	59.4	11.0	101.2	8.5	73.8	64.5	105.3	4.4	45.8	37.7
	Testosterone	124.2	6.1	58.3	21.4	100.5	9.7	42.7	24.5	114.5	8.1	27.3	19.4
	Thiacloprid	136.7	3.1	49.7	27.4	94.0	7.0	60.0	66.3	121.1	4.8	15.9	24.8
	Thiamethoxam	121.3	1.8	72.5	46.3	99.7	7.3	87.3	85.1	101.1	6.3	53.6	39.0
	Venlafaxine	122.5	3.4	42.2	3.4	96.0	6.3	58.0	54.8	108.4	21.6	8.7	13.9
NEGATIVE IONISATION	2-Phenylphenol											32.3	6.0
	5-Hydroxyindole-3-acetic acid	93.8	20.0	59.2	79.1	82.2	28.2	97.0	98.7			-8.3	62.2
	Acetylsalicylic acid											78.6	-5.3
	Arachidonic acid							67.9	24.7				
	Diclofenac	134.7	3.2	33.6	31.8	98.8	8.1	36.4	48.1	81.4	12.1	-16.0	24.0
	Dimethyldithiophosphate	60.7	16.4	34.3	-37.9	79.7	12.3	74.2	70.0			-16.6	-87.4
	Ibuprofen					62.2	46.9	16.8	49.2			-10.4	43.0
	Leukotriene B4	144.5	17.7	76.9	81.7					103.5	15.6	65.3	77.2
	Leukotriene D4	125.8	10.1	93.0	88.1			99.4	99.7	47.4	31.6	95.6	86.7
	Prostaglandin D2	104.1	13.7	65.2	64.2					65.0	26.9	35.5	46.7
	Prostaglandin E2	92.9	13.8	86.9	81.4	101.0	19.7	72.0	84.0	97.7	87.1	63.6	72.8
	Prostaglandin F2a	102.3	9.6	84.1	75.4	91.4	13.9	76.8	84.5	41.7	13.4	58.2	73.6
	Prostaglandin J2							15.1	56.0				
	Triclosan	112.4	7.4	-92.8	39.1	109.7	12.1	35.7	73.0	74.7	15.0	54.7	43.5

Table A3 – (continued) Preselection: Recovery, repeatability (recovery coefficient of variation CV) and matrix effect (ME) at 20 and 150 ng/mL of all sample preparation methods on individual compounds

		PHOSPHOLIPID AND PROTEIN REMOVAL PLATES								SUPPORTED LIQUID EXTRACTION CARTRIDGE			
		PL				PLUltra				Isolute			
	Component	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)
POSITIVE IONISATION	2-Aminobenzimidazole	57.6	10.9	-39.2	-33.1	45.3	10.7	-46.8	-27.0	69.2	2.6	47.9	26.1
	4-Androstene-3,17-dione	87.2	4.1	19.1	4.7	87.4	7.9	17.8	9.8	100.5	4.6	59.1	36.2
	Acetochlor	80.7	6.3	37.8	11.9	81.7	8.9	32.3	19.4	103.8	2.4	63.2	39.9
	Aflatoxin B1	60.5	8.9	4.5	5.1	46.4	3.3	-14.0	2.7	90.8	4.1	36.3	35.9
	Azoxystrobin	92.2	6.4	16.9	-16.0	88.5	7.4	7.1	-16.3	102.2	1.7	48.9	23.5
	Boscalid	106.5	5.5	35.6	16.6	82.9	8.9	37.9	27.5	93.2	3.6	55.9	29.8
	Carbamazepine	79.5	3.6	-72.8	-98.1	77.6	7.3	-86.4	-96.3	106.4	2.9	-4.4	-23.4
	Carbendazim	67.7	5.4	-45.1	-8.1	56.3	9.9	-54.1	-2.8	106.9	3.5	39.4	34.4
	Chlorpyrifos	82.7	9.1	88.8	91.2	79.7	7.7	90.5	91.7	79.9	2.4	91.8	90.7
	Clothianidin	51.2	8.2	-27.4	-33.7	43.7	9.7	-29.6	-19.7	101.4	4.1	34.0	2.2
	Codeine	35.9	3.2	-14.0	-57.6	29.6	4.3	-22.3	-47.9	91.4	4.9	47.7	3.6
	Cortisone	48.4	6.5	12.8	-18.7	41.0	6.9	12.7	-15.0	99.3	5.2	-67.1	2.2
	Cotinine	36.8	12.7	62.9	-25.2	26.7	25.0	37.9	-34.5	16.1	29.5	81.6	60.0
	Cyprodinil	137.3	8.5	66.0	62.6	96.8	10.9	65.5	62.3	105.7	11.9	87.1	81.4
	Diazinon	64.2	9.1	42.8	39.0	52.9	28.8	40.6	46.6	97.2	12.9	74.2	61.4
	Estrone	85.7	6.0	18.6	32.8	94.7	7.9	26.8	43.0	98.1	6.8	29.8	5.9
	Fluoxetine	86.3	2.9	43.7	22.2	78.8	7.8	45.2	43.0	98.3	2.7	86.4	78.3
	Hydrocortisone	48.7	10.9	46.6	-33.1	37.7	13.4	27.1	-26.5	106.3	18.9	85.8	88.9
	Imidacloprid	53.4	7.5	-30.9	-41.8	44.5	9.4	-29.2	-40.4	96.5	2.4	36.0	5.9
	Ketoprofen	83.8	2.3	-33.2	-55.2	84.4	6.8	-34.6	-50.3	1.5	5.5	45.2	-18.7
	Malathion	94.0	5.7		18.8	86.9	8.5	43.0	17.7	34.8	4.3	61.0	32.3
	Nicotine	26.9	9.8	44.6	65.9	21.2	9.1			96.3	2.5	48.1	30.0
	Paracetamol	35.2	10.4	29.7	47.2	28.6	10.2	33.1	52.3	84.1	3.8	39.7	18.2
	Paroxetine	74.1	7.4	29.3	17.3	64.9	13.6	31.3	11.4	106.8	2.0		79.5
	Piperine			69.7	68.5			64.6	74.4	87.7	11.1	47.3	23.5
	Pravastatin	48.6	6.6	-76.9	-91.3	46.6	6.5	-82.7	-88.7			61.4	56.1
	Prochloraz	90.2	8.4	60.9	63.1	98.0	9.5	65.3	76.9	110.6	6.0	73.8	69.8
	Progesterone	109.0	1.5	56.4	48.7	113.9	4.2	55.3	57.9	95.7	2.5	61.3	48.3
	Propiconazole	118.9	3.4	53.3	44.4	118.1	6.5	55.6	56.2	104.4	1.5	66.8	54.4
	Sertraline	84.3	15.6	61.3	55.3	104.3	2.4	65.2	71.3	83.1	9.1	92.6	89.3
	Solanidine	125.9	5.7	38.5	28.6	140.1	8.2	39.4	41.4	88.9	5.0	78.4	70.5
	Tebuconazole	107.5	2.9	37.9	23.0	99.3	5.9	40.7	38.6	102.9	1.1	65.1	43.3
	Testosterone	91.7	1.9	22.5	-1.5	93.8	7.0	16.1	1.6	102.3	2.0	30.3	21.9
	Thiacloprid	63.3	4.6	-5.7	-3.6	55.2	10.1	-13.4	0.6	98.3	5.5	60.0	31.4
	Thiamethoxam	48.8	9.9	37.8	16.1	39.1	8.8	34.5	23.0	30.7	8.1	51.0	28.2
	Venlafaxine	67.3	5.2	48.0	-18.6	56.3	9.9	-25.4	-23.1	108.5	1.4	49.8	26.7
NEGATIVE IONISATION	2-Phenylphenol	10.1	31.7	88.0	54.3	11.1	14.0	86.7	53.4			96.3	84.1
	5-Hydroxyindole-3-acetic acid	12.3	25.0	-0.4	49.4	9.3	48.2	19.6	54.1			75.4	82.7
	Acetylsalicylic acid											52.2	59.3
	Arachidonic acid									85.0	20.0	31.4	-17.3
	Diclofenac	112.4	2.2	51.2	27.4	106.7	3.7	53.3	30.7	18.6	10.3	0.1	5.2
	Dimethyldithiophosphate	52.8	10.2	15.4	-69.2	44.5	8.2	49.1	-54.9			85.4	70.9
	Ibuprofen	150.5	15.6	19.1	36.0	144.0	28.7	31.1	39.0			39.6	65.3
	Leukotriene B4	11.4	10.2	65.0	72.4	11.4	6.8	66.1	73.6				
	Leukotriene D4	15.0	0.6	73.5	83.9	16.0	6.7	75.7	85.1			68.1	77.9
	Prostaglandin D2	15.7	19.0	93.9	63.6	16.8	63.1	44.2	58.4			2.9	36.5
	Prostaglandin E2	65.2	8.9	43.3	59.3	64.1	5.1	97.0	64.2	2.7	15.7	42.6	51.2
	Prostaglandin F2a	70.8	5.9	69.2	62.1	70.3	4.1	70.7	63.3	2.9	4.9	50.8	56.0
	Prostaglandin J2			71.0	78.2			57.8	74.9				
	Triclosan	142.0	50.6	63.6	68.0	150.2	28.8	62.9	74.9	80.0	8.0	-86.9	41.6

Table A3 – (continued) Preselection: Recovery, repeatability (recovery coefficient of variation CV) and matrix effect (ME) at 20 and 150 ng/mL of all sample preparation methods on individual compounds

		SOLID PHASE EXTRACTION CARTRIDGES											
		HLB				Strata X				Strata XC			
	Component	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)	Mean recov. (%)	Recovery CV (%)	Low-level ME (%)	High-level ME (%)
POSITIVE IONISATION	2-Aminobenzimidazole	78.4	10.9	77.9	75.5	66.1	12.9	-4.8	22.9			11.4	16.0
	4-Androstene-3,17-dione	111.6	2.9	-68.8	-86.4	85.1	9.5	49.1	42.0	97.2	3.4	25.0	21.5
	Acetochlor	26.0	86.0	-11.8	-16.5	109.4	5.6	86.8	79.9	126.0	3.4	89.3	87.1
	Aflatoxin B1	112.8	3.9	27.3	35.7	73.4	19.3	49.7	44.3	134.5	20.0	55.0	60.9
	Azoxystrobin	119.8	2.1	64.3	49.0	87.7	13.1	57.7	41.2	97.6	2.6	45.6	25.9
	Boscalid	108.6	7.7	61.6	55.4	81.8	9.7	73.8	71.8	107.1	1.8	26.5	19.8
	Carbamazepine			-32.5	-36.0	97.9	5.6			90.1	5.8	-38.8	-37.8
	Carbendazim	110.6	3.1	11.1	32.0	114.5	2.2	2.8	23.2	105.2	6.1	-6.8	27.3
	Chlorpyrifos	94.3	6.4	99.6	99.3	71.7	13.3			109.4	22.3	99.4	98.5
	Clothianidin	110.2	4.2	46.5	36.8	107.9	4.5	28.8	8.8	102.0	1.9	41.7	38.5
	Codeine	98.9	22.3	29.4	11.7	110.5	3.2	21.5	-5.7	114.8	15.2	11.9	2.6
	Cortisone	124.9	2.6	25.2	17.9	107.8	4.3	31.5	7.4	86.9	17.0	33.5	23.7
	Cotinine			47.2	26.1	66.8	7.8	-25.6	16.6			65.7	47.1
	Cyprodinil	128.9	21.0	94.4	91.6	62.9	6.9			88.8	4.2	67.2	60.1
	Diazinon	90.7	7.8	87.8	84.8	86.8	18.5	98.9	97.2	110.1	6.6	96.3	90.5
	Estrone	130.2	1.5	67.5	67.0	70.2	14.4	57.4	58.6	89.4	4.1	-74.3	-49.6
	Fluoxetine	116.9	8.5	93.7	89.2	110.0	8.7	76.9	78.7	103.1	1.6	80.8	76.0
	Hydrocortisone	126.4	7.8	22.7	24.5	109.1	11.2	53.8	19.5	103.9	8.8	44.5	25.9
	Imidacloprid	113.6	1.6	52.6	40.8	111.0	2.0	34.1	6.6	98.4	2.7	46.4	35.0
	Ketoprofen	117.8	0.9	59.1	49.5	106.2	6.8	26.8	2.2	109.1	2.9	44.5	37.5
	Malathion	113.7	3.9	82.1	73.9	69.2	14.3	82.5	69.7	2.8	63.7	94.5	91.3
	Nicotine	0.5	21.0	28.1	-5.5			32.9	-12.8	5.7	24.9		
	Paracetamol	104.2	7.5	58.2	50.3	97.5	3.1	95.2	79.6	75.5	14.7	50.6	49.9
	Paroxetine	128.8	4.3	92.4	86.6	102.6	9.3	68.3	73.2	96.7	5.0	76.8	67.6
	Piperine	156.9	16.6	75.5	65.0	82.1	59.9	79.4	81.1	76.8	13.3	59.6	33.1
	Pravastatin	108.3	4.3	40.2	41.6	88.6	6.3	13.6	12.2	65.1	18.3	13.7	29.0
	Prochloraz	132.8	11.3	90.2	86.4	149.8	22.5	91.1	93.6	71.1	3.5	49.7	46.5
	Progesterone	99.9	9.8	75.8	70.2	101.1	7.2	83.2	88.4	170.1	9.5	-62.1	-94.8
	Propiconazole	119.2	3.3	78.0	71.7	110.9	2.3	89.2	91.5	100.0	1.7	34.1	32.4
	Sertraline	121.6	33.2	95.2	94.9	123.0	13.7	87.9	91.1	99.5	5.0	80.1	75.4
	Solanidine	132.0	8.6	86.1	84.2	121.5	5.3	89.1	78.3	101.9	2.7	69.3	63.9
	Tebuconazole	115.1	2.7	71.6	64.3	104.9	6.3	76.5	84.4	106.5	2.2	29.2	23.7
	Testosterone	132.1	5.2	56.2	46.0	153.9	6.3	58.1	51.6	110.0	6.5	37.3	25.5
	Thiacloprid	106.3	3.2	63.0	63.9	105.4	5.2	30.6	20.2	99.5	2.2	58.2	62.0
	Thiamethoxam	111.2	2.6	74.8	61.1	112.1	1.2	73.9	53.7	88.3	4.8	71.1	59.9
	Venlafaxine	101.7	5.6	74.4	68.5	112.9	4.9	33.0	12.9	105.1	3.0	74.5	70.7
NEGATIVE IONISATION	2-Phenylphenol			94.7	81.8								95.3
	5-Hydroxyindole-3-acetic acid	127.0	22.2	65.3	80.8	132.6	58.4	40.9	21.3	73.1	33.8	62.8	81.9
	Acetylsalicylic acid	130.9	14.0					45.4	19.4	16.4	48.0		
	Arachidonic acid									153.7	36.8		
	Diclofenac	108.8	2.0	51.8	62.9	99.7	14.4	89.5	93.0	107.9	3.7	21.4	41.7
	Dimethyldithiophosphate	11.1	9.6	12.7	-18.4	26.9	6.0	46.5	16.2	44.6	6.6	32.3	-0.8
	Ibuprofen	115.7	49.1	93.2	93.8	104.2	103.8			156.8	17.3	90.3	88.1
	Leukotriene B4	114.0	5.7	75.2	80.6	105.9	10.7	66.3	69.1	53.9	48.0	83.2	74.3
	Leukotriene D4	122.1	35.8	85.9	93.0	163.4	30.2	39.6	15.9	107.5	5.7	51.2	76.1
	Prostaglandin D2	112.5	2.8	-37.4	19.0	65.9	68.6	-0.6	3.1	101.2	6.9	-0.9	29.6
	Prostaglandin E2	120.9	9.9	57.3	73.2	74.3	18.2	-8.5	67.7	105.3	3.0	68.2	76.5
	Prostaglandin F2a	116.0	6.4	57.5	70.6	94.8	12.0	50.3	73.6	112.3	2.3	60.9	72.1
	Prostaglandin J2	40.8	134.1	88.8	95.0	135.2	24.3	-77.0	17.6	151.5	79.3	-61.2	-42.8
	Triclosan	160.8	22.8	81.9	95.6	172.8	15.2	94.1	96.2	108.8	3.6	-20.8	62.7

2.4. Table A4a – Comparison to PPT (Serum): Detection, repeatability, S/N and spiking significance of preselected preparation methods on individual compounds

Table A4a – Comparison of sample preparation methods to PPT in serum: median area, repeatability (area coefficient of variation CV), signal/noise ratio (S/N), and p-value of areas in four spiked vs. four non-spiked samples

		PPT				Phree				StrataX				Phree+StrataX			
	Component	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value $$
POSITIVE IONISATION	2-Aminobenzimidazole	168885	7	789	1.7E-07	152233	7	695	1.5E-07	171971	18	1116	3.1E-05	112054	23	707	1.3E-04
	4-Androstene-3,17-dione	191923	9	2868	5.9E-07	47650	6	1151	3.1E-07	150553	13	2832	5.1E-06	94218	11	1935	2.6E-06
	Acetochlor	10563	3	179	7.7E-10	3509	16	73	1.4E-05	8315	17	166	2.6E-05				
	Aflatoxin B1	7514	18	519	2.8E-05	41863	4	2315	5.8E-09	52222	15	3151	9.7E-06	32722	16	1635	1.4E-05
	Azoxystrobin	248938	8	10312	3.1E-07	54313	5	2699	2.2E-08	178762	14	8872	8.0E-06	121984	9	6779	5.7E-07
	Boscalid	107469	11	3234	1.6E-06	13141	7	884	1.1E-07	66410	18	7401	3.0E-05	39119	17	3339	2.1E-05
	Carbamazepine	157472	9	4397	4.9E-07	78172	1	2118	7.0E-13	136496	13	2820	5.9E-06	90113	15	1774	9.2E-06
	Carbendazim	130835	8	1717	2.0E-07	88661	8	1277	2.5E-07	111303	18	1739	2.9E-05	74097	13	1124	3.8E-06
	Chlorpyrifos	16680	12	3407	3.2E-06					7120	27	3323	2.8E-04	771	32	24756	1.2E-03
	Clothianidin	8563	7	532	1.2E-07	6952	14	382	6.9E-06	9478	14	557	7.4E-06	5934	21	320	7.7E-05
	Codeine	191972	6	3274	1.0E-07	163579	9	2849	6.9E-07	193676	17	3795	3.0E-05	118877	14	2450	8.8E-06
	Cortisone	151845	23	6019	1.1E-02	170331	7	6167	1.8E-05	227886	3	6822	3.6E-08	136684	12	3879	1.2E-03
	Cotinine	983674	4	18717	1.8E-05	843132	14	11176	2.5E-02	4079	4	28824	2.1E-07	283673	38	5803	5.4E-04
	Cyprodinil	509013	8	7287	3.1E-07	8515	14	155	7.0E-06	195296	33	2879	9.3E-04	57770	22	910	9.4E-05
	Diazinon	455855	3	13720	1.1E-06	9496	20	371	4.1E-03	169970	29	5261	1.0E-02	42403	78	1492	1.7E-01
	Estrone	38956	15	648	1.0E-05	6198	8	144	2.3E-07	22432	12	394	3.6E-06	14510	16	272	1.7E-05
	Fluoxetine	129336	5	2086	1.8E-08	4379	8	88	3.8E-07	71209	25	891	2.0E-04	27871	28	380	3.6E-04
	Hydrocortisone	840557	12	29101	8.6E-04	669610	8	22996	3.2E-03	913228	8	24658	5.3E-05	515564	10	14568	6.9E-01
	Imidacloprid	23078	8	2083	3.6E-07	20014	7	1464	8.5E-08	24441	15	2167	1.3E-05	16801	13	1192	4.5E-06
	Ketoprofen	52925	7	240	1.8E-07	26807	3	105	1.4E-09	57919	13	216	5.3E-06	38584	8	147	2.2E-07
	Malathion					1560	6	89	6.2E-08	8526	12	777	2.7E-06	4522	31	250	6.5E-04
	Nicotine	20945	11	176	4.7E-05	22208	10	170	2.1E-05					5178	25	48	4.9E-01
	Paracetamol	15072	1	190	7.5E-13	12304	2	152	9.0E-11								
	Paroxetine	317360	9	7914	1.3E-04	10556	8	260	1.7E-04	173573	28	3109	9.2E-03	60558	34	1202	1.8E-02
	Piperine	3437214	10	22785	1.0E-06	173761	8	1459	3.6E-01	877977	23	5335	4.6E-04	476506	21	2943	8.2E-04
	Pravastatin	6312	14	44	8.7E-06	2302	13	29	5.0E-06	3409	31	24	6.5E-04	2179	5	24	1.6E-08
	Prochloraz	62878	8	11665	3.4E-07	1529	15	151	9.5E-06	30254	31	12263	6.2E-04	12973	31	2449	6.7E-04
	Progesterone	261837	10	4083	2.2E-04	7670	11	299	4.7E-04	1141	18	2637	1.9E-03	49446	24	1429	5.7E-03
	Propiconazole	269169	9	9695	6.8E-07	5766	6	235	6.6E-08	136439	27	6427	3.4E-04	56249	22	2284	1.0E-04

Table A4a – (continued) Comparison of sample preparation methods to PPT in serum: median area, repeatability (area coefficient of variation CV), signal/noise ratio (S/N), and p-value of areas in four spiked vs. four non-spiked samples

		PPT				Phree				Strata X				Phree + Strata X			
	Component	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value
POSITIVE	Sertraline	39354	7	1215	1.1E-07					21629	24	570	1.7E-04	7395	31	203	7.0E-04
	Solanidine	286540	9	17609	5.3E-07	10756	34	1024	1.0E-03	178210	33	45111	9.4E-04	67834	31	38493	6.9E-04
	Tebuconazole	340486	11	6042	1.5E-06	10130	7	225	1.4E-07	180920	21	3197	8.5E-05	82258	20	1526	5.2E-05
	Testosterone	244917	10	5802	1.7E-06	47622	4	1718	6.6E-07	179614	11	4597	3.0E-06	111394	12	3170	4.6E-06
	Thiacloprid	59923	8	5852	1.9E-07	44756	6	3203	4.1E-08	60133	15	6213	9.4E-06	40721	15	3374	9.3E-06
	Thiamethoxam	13063	7	876	1.8E-07	14312	9	757	7.0E-07	16908	15	1422	1.2E-05	12060	15	798	9.2E-06
	Venlafaxine	178703	5	14722	1.3E-08	125662	7	5814	1.3E-07	167243	15	16474	1.3E-05	96171	22	4650	9.6E-05
NEGATIVE IONISATION	2-Phenylphenol	2918856	3	56373	1.3E-04	1321204	7	29157	5.0E-01	2938277	8	57962	2.4E-04	2014471	4	44556	5.7E-03
	5-Hydroxyindole-3-acetic acid	14004	13	344	4.6E-06					24015	8	309	2.2E-07	31460	13	588	4.0E-06
	Acetylsalicylic acid	5506375	2	54917	6.2E-06	3767287	7	33621	2.9E-01	6027751	6	47136	3.5E-05	3746474	7	38863	3.2E-01
	Arachidonic acid	8909809	14	22998841	9.4E-06	236642	143	721170	9.7E-01	5050161	29	3671468	7.4E-04	793832	54	218759	9.6E-02
	Diclofenac	16400	8	1071	2.5E-07	1783	6	115	5.6E-08	13792	19	966	3.8E-05	9428	11	600	1.7E-06
	Dimethyldithiophosphate	16609	32	2783	7.7E-04	458	80	70	4.7E-02								
	Ibuprofen					1402	13	78	2.1E-05								
	Leukotriene B4	165375	9	5830	2.7E-06	30321	20	4797	7.7E-01	181828	11	5112	6.6E-06	94865	42	4208	1.6E-02
	Leukotriene D4	17303	10	1150	9.4E-07					6511	11	411	1.5E-06				
	Prostaglandin D2	135409	6	3467	3.3E-04	90582	18	4818	8.3E-01	205159	6	6306	6.6E-06	118770	8	5289	3.8E-03
	Prostaglandin E2	187066	4	4590	2.3E-05	118566	10	5057	1.9E-03	243065	4	6027	3.3E-06	151570	7	5217	2.0E-04
	Prostaglandin F2a	242325	5	19900	4.6E-08	180703	9	137090	1.3E-06	263279	11	30284	2.9E-06	128622	67	26327	3.6E-02
	Prostaglandin J2	31280	7	2955	4.0E-06	27588	11	3450	2.7E-05	47768	10	4957	5.3E-06	31572	8	3622	5.5E-06
	Triclosan	1308	23	373	3.1E-04					388	12	208	1.2E-05				
	Detection frequency	96	90	92	88												
	Median S/N	3437	1024	3260	2109												
	Semi-quantification performance (% detected compounds with CV < 20%)	94	93	72	55												
	Median p-value	1.1E-06	5.0E-06	1.9E-05	1.2E-04												
	Speed of implementation	4	3	2	1												

2.5. Table A4b – Comparison to PPT (Plasma): Detection, repeatability, S/N and spiking significance of preselected preparation methods on individual compounds

Table A4b – Comparison of sample preparation methods to PPT in plasma: median area, repeatability (area coefficient of variation CV), signal/noise ratio (S/N), and p-value of areas in four spiked vs. four non-spiked samples

		PPT				Phree				Strata X				Phree + Strata X			
	Component	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value	Mean area	Area CV (%)	S/N	p-value
POSITIVE IONISATION	2-Aminobenzimidazole	119925	9	850	7.5E-07	110613	15	693	1.2E-05	198425	6	1302	4.8E-08	98928	18	751	3.1E-05
	4-Androstene-3,17-dione	107968	14	3136	8.0E-06	74565	8	1954	5.0E-07	146575	10	3123	1.2E-06	78070	14	1930	9.6E-06
	Acetochlor	8089	7	148	3.6E-04	6052	15	137	2.8E-03	10158	29	168	2.0E-03	4384	33	44	2.8E-02
	Aflatoxin B1	1794	13	120	3.8E-06	24293	12	1204	2.6E-06	39285	13	2184	5.3E-06	16947	39	809	2.0E-03
	Azoxystrobin	143400	18	12421	3.1E-05	94640	8	5809	2.1E-07	184050	12	16547	2.8E-06	94188	20	6337	5.7E-05
	Boscalid	66505	20	4268	6.0E-05	29388	4	1611	5.8E-09	77270	18	8719	3.2E-05	42040	15	3944	1.2E-05
	Carbamazepine	89048	11	2819	1.6E-06	75585	12	2200	3.3E-06	130975	6	3116	5.4E-08	79290	9	1722	4.2E-07
	Carbendazim	85070	10	1471	1.4E-06	77730	13	1279	5.1E-06	127850	1	1959	2.1E-12	76923	7	1400	1.8E-07
	Chlorpyrifos	7386	36	1569	1.4E-03	909	45	1154	4.5E-03	4077	25	466712	1.9E-04				
	Clothianidin	4952	13	313	4.9E-06	4735	12	291	3.6E-06	8127	6	574	3.7E-08	5093	7	309	1.2E-07
	Codeine	125800	10	4097	1.1E-06	105925	13	3143	4.2E-06	159475	23	4504	1.4E-04	76145	18	2155	3.3E-05
	Cortisone	69430	12	4563	6.4E-02	102163	14	5165	1.2E-03	198250	23	7673	8.5E-04	97038	13	2525	1.2E-03
	Cotinine	68790	8	1959	1.1E-06	62820	11	1579	9.9E-06	142175	12	2803	4.7E-06	98930	26	1526	4.9E-04
	Cyprodinil	330550	19	5794	3.8E-05	36065	16	616	1.6E-05	242600	38	3570	2.0E-03	70283	37	1057	1.6E-03
	Diazinon	269900	8	8686	3.9E-07	81685	19	3302	8.7E-05	101185	20	3722	1.0E-04	2550	146	90	2.6E-02
	Estrone	23205	19	739	4.4E-05	10574	7	249	2.8E-07	24033	19	493	4.6E-05	15145	15	213	1.4E-05
	Fluoxetine	58570	22	1189	9.3E-05	21205	10	372	1.0E-06	75745	26	957	2.5E-04	28685	50	322	7.2E-03
	Hydrocortisone	380800	12	24229	7.9E-01	394750	15	20203	5.5E-01	833225	32	25554	1.4E-02	380450	12	8139	8.0E-01
	Imidacloprid	13615	10	1150	1.2E-06	14173	15	1039	1.2E-05	21910	3	1865	1.1E-09	14698	10	1060	1.1E-06
	Ketoprofen	29941	13	177	5.4E-04	28607	8	124	7.7E-05	46685	10	248	2.0E-04	30280	6	145	2.0E-05
	Malathion					4420	3	189	1.9E-09	8483	6	604	6.9E-08	1598	37	120	1.6E-03
	Nicotine	1070	46	69	7.6E-02	10020	17	419	3.6E-03	2160	120	243	4.5E-01	338	200	99	3.5E-02
	Paracetamol	10091	14	240	9.2E-06	11177	20	249	6.7E-05	63458	22	1065	9.5E-05	63138	14	1272	6.4E-06
	Paroxetine	139875	23	4656	1.3E-04	44115	12	1125	3.3E-06	176000	26	3735	2.4E-04	55408	55	1109	1.1E-02
	Piperine	359650	17	6420	5.1E-04	134550	4	2489	1.3E-01	341500	51	4991	7.0E-02	149503	34	2586	9.4E-01
	Pravastatin	7605	10	787	1.3E-06	5805	14	426	8.0E-06	5859	14	214	6.8E-06	3423	23	83	1.1E-04
	Prochloraz	36083	22	2944	1.0E-04	5510	9	445	7.4E-07	40598	23	17210	1.3E-04	16627	37	5059	1.8E-03
	Progesterone	158625	17	4786	2.2E-05	35210	9	1760	5.8E-07	140875	17	4001	2.6E-05	44060	32	1859	8.0E-04
	Propiconazole	172000	19	7222	4.5E-05	39855	5	1561	3.2E-08	190800	20	8702	6.1E-05	88903	22	3515	1.0E-04
	Sertraline	22738	21	625	8.1E-05	3498	7	81	1.3E-07	24053	22	615	8.8E-05	8318	41	194	2.9E-03

Table A4b – (continued) Comparison of sample preparation methods to PPT: median area, repeatability (area coefficient of variation CV), signal/noise ratio (S/N), and p-value of areas in four spiked vs. four non-spiked samples

		PPT				Phree				Strata X				Phree + Strata X			
	Component	**Mean area**	**Area CV (%)**	**S/N**	**p-value**	**Mean area**	**Area CV (%)**	**S/N**	**p-value**	**Mean area**	**Area CV (%)**	**S/N**	**p-value**	**Mean area**	**Area CV (%)**	**S/N**	**p-value**
POSITIVE	Solanidine	250825	20	11228	6.2E-05	52193	10	3289	1.8E-06	191195	43	43909	3.4E-03	58968	39	3035	2.3E-03
	Tebuconazole	215950	18	5617	3.7E-05	57930	5	1379	1.5E-08	245375	22	4146	1.0E-04	119800	24	2434	1.7E-04
	Testosterone	122550	16	4621	1.6E-05	76673	8	2827	2.3E-07	156500	11	4553	2.3E-06	88753	8	2856	3.9E-07
	Thiacloprid	35550	12	2586	3.1E-06	33135	13	2243	4.6E-06	55410	5	5378	9.6E-09	34043	9	2766	4.7E-07
	Thiamethoxam	7122	12	545	3.3E-06	8693	18	616	3.0E-05	13413	8	1153	1.9E-07	9197	11	783	1.9E-06
	Venlafaxine	105735	10	13456	1.3E-06	95028	13	4150	5.5E-06	160775	11	23569	1.4E-06	81265	15	3462	1.1E-05
NEGATIVE IONISATION	2-Phenylphenol	3867	20	101	7.1E-01	3626	8	97	6.5E-01	5014	24	127	7.7E-02	4914	7	135	6.5E-04
	5-Hydroxyindole-3-acetic acid	8032	12	183	1.2E-01	9522	18	195	8.7E-01	16040	29	298	3.1E-02	10653	7	211	9.4E-02
	Acetylsalicylic acid					4580	8	96	1.1E-06	13828	11	135	2.6E-06	5287	39	98	3.9E-03
	Arachidonic acid	1587750	5	2821075	2.2E-08	61620	12	1888391	4.6E-05	492975	63	365127	5.1E-02	58495	57	19408	1.3E-02
	Diclofenac	18945	10	1129	1.3E-06	4799	10	282	7.7E-07	15553	21	1210	7.1E-05	9951	11	669	1.4E-06
	Dimethyldithiophosphate	28970	4	1381	4.0E-06	10171	22	477	5.1E-03								
	Ibuprofen	8432	6	761	1.5E-05	4377	15	318	2.1E-02	13618	46	918	1.4E-02	1968	25	105	1.8E-01
	Leukotriene B4	88910	8	45976	2.4E-07	23270	8	14083	2.3E-07	49465	42	9744	3.0E-03	6752	83	1463	5.3E-02
	Leukotriene D4	29878	11	3633	1.4E-06	871	53	225	9.1E-03	4598	84	360	5.8E-02				
	Prostaglandin D2					8626	17	632	2.1E-05	16003	15	906	1.2E-05	7597	46	448	4.7E-03
	Prostaglandin E2	28600	16	2535	1.7E-05	38355	13	3087	4.8E-06	61820	14	4114	6.2E-06	29988	41	1956	2.8E-03
	Prostaglandin F2a	105975	7	181698	1.3E-07	97193	14	90725	7.9E-06	144150	19	57909	4.7E-05	54948	59	15756	1.5E-02
	Prostaglandin J2	11630	9	2290	1.0E-06	9721	9	842	1.6E-06	17138	27	914	4.4E-04	8656	30	416	9.5E-04
	Triclosan	3583	4	485	7.9E-09	571	6	63	3.1E-07	1727	20	390	5.9E-05				

Detection frequency	94	100	98	92
Median S/N	2535	1082	2803	1190
Semi-quantification performance (% detected compounds with CV < 20%)	81	94	47	46
Median p-value	1.5E-05	5.0E-06	7.1E-05	8.8E-04
Speed of implementation	4	3	2	1

2.6. Table A5a – Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification

Table A5a – Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	SMILES	CI m/z		CI Rt								CI isotopic fit		Global CI	
				Experimental		RTI-predicted		Retip-predicted		logP-predicted		CI overall			
		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
1,8-Epoxy-p-menthan-3-ol glucoside *	CC1(C2CCC(O1)(CC2OC3C(C(C(C(O3)CO)O)O)O)C)C	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
25-Hydroxyvitamin D3 26,23-lactol *	CC(CC1CC(C(O1)O)(C)O)C2CCC3C2(CCCC3=CC=C4CC(CCC4=C)O)C	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2-naphthylamine	Nc1ccc2ccccc2c1	0.94	n.a.	n.a.	n.a.	**0.85**	n.a.	0.84	n.a.	0.84	n.a.	0.80	n.a.	G3_0.86	n.a.
3-[2-(5-Methylthiophen-2-yl)-2-oxoethoxy]benzonitrile *	CC1=CC=C(S1)C(=O)COC2=CC=CC(=C2)C#N	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	OC(=O)c1cccc(O)c1	n.a.	0.85	n.a.	n.a.	n.a.	**0.87**	n.a.	0.87	n.a.	0.86	n.a.	n.a.	n.a.	G2_0.86
4-chlorophenol	Oc1ccc(Cl)cc1	n.a.	0.78	n.a.	n.a.	n.a.	**0.83**	n.a.	0.78	n.a.	0.83	n.a.	0.91	n.a.	G3_0.84
4-hydroxy-2,5,6-trichloroisophthalonitrile	Oc1c(Cl)c(Cl)c(C#N)c(Cl)c1C#N	n.a.	0.82	n.a.	n.a.	n.a.	n.a.	n.a.	**0.80**	n.a.	0.76	n.a.	0.78	n.a.	G3_0.8
4-hydroxybenzoic acid	OC(=O)c1ccc(O)cc1	n.a.	0.85	n.a.	n.a.	n.a.	**0.85**	n.a.	0.83	n.a.	0.81	n.a.	n.a.	n.a.	G2_0.85
4-Hydroxyquinoline *	C1=CC=C2C(=C1)C(=O)C=CN2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-Nitrophenol *	C1=CC(=CC=C1[N+](=O)[O-])O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-Sulfamoylbenzoic acid *	CC1=CC=C(S1)C(=O)COC2=CC=CC(=C2)C#N	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Acetaminophen sulfate	CC(=O)NC1=CC=C(C=C1)OS(=O)(=O)O	n.a.	0.84	n.a.	n.a.	n.a.	**0.82**	n.a.	0.99	n.a.	n.a.	n.a.	0.46	n.a.	G3_0.43
Azelaic acid *	C(CCCC(=O)O)CCCC(=O)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Benzophenone-4	COc1cc(O)c(cc1[S](O)(=O)=O)C(=O)c2ccccc2	n.a.	0.72	n.a.	n.a.	n.a.	**0.85**	n.a.	0.66	n.a.	0.95	n.a.	n.a.	n.a.	G2_0.79
Caffeic acid	OC(=O)\C=C/c1ccc(O)c(O)c1	n.a.	0.91	n.a.	n.a.	n.a.	**0.97**	n.a.	0.89	n.a.	0.85	n.a.	n.a.	n.a.	G2_0.94
Caffeine	Cn1cnc2N(C)C(=O)N(C)C(=O)c12	0.97	n.a.	**1.00**	n.a.	0.94	n.a.	0.77	n.a.	0.61	n.a.	n.a.	n.a.	G2_0.98	n.a.
Carveol *	CC(=C)C1CC=C(C)C(O)C1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	C=CCC1=CC=C(C=C1)OS(=O)(=O)O	n.a.	0.85	n.a.	**0.40**	n.a.	0.24	n.a.	0.86	n.a.	0.83	n.a.	0.24	n.a.	G3_0.5
Coumaric acid	OC(=O)\C=C\c1cccc(O)c1	1.00	0.98	n.a.	n.a.	**0.83**	**0.83**	0.83	0.83	0.83	0.83	0.93	n.a.	G3_0.92	G2_0.9
Cresol sulfate	CC1=CC=CC=C1OS(=O)(=O)O	n.a.	0.99	n.a.	**0.90**	n.a.	0.95	n.a.	0.77	n.a.	0.84	n.a.	0.85	n.a.	G3_0.91
Diethylphthalate *	CCOC(=O)C1=CC=CC=C1C(=O)OCC	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diphenylsulfone *	C1=CC=C(C=C1)S(=O)(=O)C2=CC=CC=C2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carbinol *	C1=CC=C2C(=C1)C(=CN2)CO	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	C1=CC=C2C(=C1)C(=CN2)C=O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indolelactic acid *	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	SMILES	CI m/z		CI Rt								CI isotopic fit		Global CI	
				Experimental		RTI-predicted		Retip-predicted		logP-predicted		CI overall			
		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
Indoxyl sulfate	C1=CC=C2C(=C1)C(=CN2)OS(=O)(=O)O	n.a.	0.85	n.a.	**0.86**	n.a.	n.a.	n.a.	n.a.	n.a.	0.75	n.a.	n.a.	n.a.	G2_0.85
Isobutylparaben	CC(C)COC(=O)c1ccc(O)cc1	0.97	n.a.	n.a.	n.a.	**0.81**	n.a.	0.67	n.a.	0.21	n.a.	0.70	n.a.	G3_0.83	n.a.
Isopropylparaben	CC(C)OC(=O)c1ccc(O)cc1	n.a.	0.97	n.a.	n.a.	n.a.	n.a.	n.a.	**0.86**	n.a.	0.91	n.a.	n.a.	n.a.	G2_0.49
Jasmonic acid *	CCC=CCC1C(CCC1=O)CC(=O)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Loliolid *	CC1(CC(CC2(C1=CC(=O)O2)C)O)C	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
L-Phenylalanine	N[C@@H](Cc1ccccc1)C(O)=O	0.92	n.a.	n.a.	n.a.	**0.82**	n.a.	n.a.	n.a.	n.a.	n.a.	0.97	n.a.	G3_0.9	n.a.
Naphthalene-2-sulfonic acid *	C1=CC=C2C=C(C=CC2=C1)S(=O)(=O)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Octaethylene glycol	OCCOCCOCCOCCOCCOCCOCCOCCO	0.94	n.a.	n.a.	n.a.	**0.63**	n.a.	0.88	n.a.	n.a.	n.a.	0.94	n.a.	G3_0.84	n.a.
Paraxanthine	Cn1cnc2NC(=O)N(C)C(=O)c12	n.a.	0.85	n.a.	n.a.	n.a.	**0.84**	n.a.	0.89	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.85
PEG18	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO	0.89	n.a.	n.a.	n.a.	**0.69**	n.a.	0.98	n.a.	n.a.	n.a.	0.85	n.a.	G3_0.81	n.a.
Piperine	O=C(/C=C/C=C/c1ccc2OCOc2c1)N3CCCCC3	0.83	n.a.	**0.96**	n.a.	0.29	n.a.	0.98	n.a.	0.38	n.a.	0.92	n.a.	G3_0.9	n.a.
Propylparaben sulfate	CCCOC(=O)C1=CC=C(C=C1)OS(=O)(=O)O	n.a.	0.86	n.a.	**0.87**	n.a.	0.33	n.a.	0.81	n.a.	0.69	n.a.	0.82	n.a.	G3_0.85
Stachydrine (Proline betaine) *	C[N+]1(CCCC1C(=O)[O-])C	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Thymol *	CC1=CC(=C(C=C1)C(C)C)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan sulfate	C1=CC(=C(C=C1Cl)OS(=O)(=O)O)OC2=C(C=C(C=C2)Cl)Cl	n.a.	0.86	n.a.	**0.94**	n.a.	n.a.	n.a.	n.a.	n.a.	0.93	n.a.	1.00	n.a.	G3_0.93
Tridecalactone *	CCCCCCCCC1CCCC(=O)O1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triethylphosphate	CCO[P](=O)(OCC)OCC	0.93	n.a.	n.a.	n.a.	**0.78**	n.a.	0.93	n.a.	0.91	n.a.	n.a.	n.a.	G2_0.86	n.a.
Tris(2-butoxyethyl)phosphate	CCCCOCCO[P](=O)(OCCOCCCC)OCCOCCCC	0.78	n.a.	n.a.	n.a.	**1.00**	n.a.	0.89	n.a.	n.a.	n.a.	0.95	n.a.	G3_0.91	n.a.
Tryptophan *	N[C@@H](Cc1c[nH]c2ccccc12)C(O)=O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	MS/MS				Confidence level
	Theoretical fragments		Experimental fragments		
	(+)	(-)	(+)	(-)	
1,8-Epoxy-p-menthan-3-ol glucoside *	n.a.	57.0346, 75.0088, 85.0295, 113.0244, 153.1286	n.a.	57.0345, 75.0089, 85.0295, 113.0242, 153.1284	2b
25-Hydroxyvitamin D3 26,23-lactol *	n.a.	411.2906	n.a.	411.2913	2b
2-naphthylamine	91.0556, 115.0542, 117.0699, 127.0542	n.a.	91.0556, 115.0545, 117.0703, 127.0554	n.a.	2a
3-[2-(5-Methylthiophen-2-yl)-2-oxoethoxy]benzonitrile *	109.9821, 111.9978, 123.9978, 140.0291	n.a.	109.9824, 111.9974, 123.9977, 140.0291	n.a.	2b
3-hydroxybenzoic acid	n.a.	93.0343	n.a.	93.0347	2a
4-chlorophenol	n.a.	91.019	n.a.	0	3
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	146.9765, 174.9704, 181.9447, 209.9401	n.a.	146.9756, 174.9704, 181.9444, 209.9394	2a
4-hydroxybenzoic acid	n.a.	93.0343	n.a.	93.0341	2a
4-Hydroxyquinoline *	77.0415, 91.0555, 104.0494, 128.0476	n.a.	77.0395, 91.0549, 104.0493, 128.0491	n.a.	2a
4-Nitrophenol *	n.a.	92.0260, 108.0229	n.a.	92.0260, 108.0235	2a
4-Sulfamoylbenzoic acid *	77.0386, 105.0336	n.a.	77.0386, 105.0338	n.a.	2b
Acetaminophen sulfate	n.a.	79.9570, 107.0374, 150.0556	n.a.	79.9572, 107.0372, 150.0560	2a
Azelaic acid *	n.a.	57.0342, 97.0655, 123.0811, 125.0970	n.a.	57.0345, 97.0652, 123.0810, 125.0962	2a
Benzophenone-4	n.a.	93.0346, 121.0295, 211.0400, 227.0714	n.a.	93.0346, 121.0295, 211.0398, 227.0713	2a
Caffeic acid	n.a.	135.0452	n.a.	135.0449	2a
Caffeine	83.0601, 110.0719, 123.0435, 138.0668	n.a.	83.0611, 110.0721, 123.0434, 138.0670	n.a.	1
Carveol *	79.0544, 91.0543, 107.0856, 119.0856	n.a.	79.0547, 91.0545, 107.0858, 119.0858	n.a.	2a
Chavicol sulfate	n.a.	105.0710, 133.0659	n.a.	105.0703, 133.0656	2b
Coumaric acid	77.0382, 91.0530, 95.0488, 103.0533, 123.0423, 147.0425	93.0348, 119.0503	77.0391, 91.0547, 95.0498, 103.0542, 123.0447, 147.0449	93.0349, 119.0505	2a
Cresol sulfate	n.a.	92.0279, 107.0493	n.a.	92.0268, 107.0499	1
Diethylphthalate *	121.0284, 149.0233, 163.0390, 177.0546	n.a.	121.0288, 149.0234, 163.0389, 177.0549	n.a.	2a
Diphenylsulfone *	77.0386, 95.0491, 125.0066, 141.0004	n.a.	77.0388, 95.0491, 125.0063, 141.0009	n.a.	2a
Indole-3-carbinol *	77.0380, 103.0555, 130.0634	n.a.	77.0383, 103.0545, 130.0643	n.a.	2a
Indole-3-carboxaldehyde *	n.a.	65.9998, 115.0422, 126.0354	n.a.	65.9999, 115.0432, 126.0354	2a
Indolelactic acid *	n.a.	72.9947, 116.0486, 130.0661, 142.0633, 158.0625, 186.0553	n.a.	72.9937, 116.0491, 130.0677, 142.0642, 158.0615, 186.0558	2a
Indoxyl sulfate	n.a.	79.9578, 132.0460	n.a.	79.9573, 132.0457	2a
Isobutylparaben	95.049, 121.0282, 139.0388	n.a.	95.0498, 121.0293, 139.0397	n.a.	2a
Isopropylparaben	n.a.	121.0297, 137.0239	n.a.	121.0297, 137.0243	2a

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	MS/MS				Confidence level
	Theoretical fragments		Experimental fragments		
	(+)	(-)	(+)	(-)	
Jasmonic acid *	105.0697, 133.1013, 151.1121, 165.1263, 193.1225	n.a.	105.0706, 133.1019, 151.1107, 165.1275, 193.1230	n.a.	2a
Loliolid *	79.0529, 91.0544, 105.0690, 117.0708, 133.1020, 161.0972, 179.1078	n.a.	79.0540, 91.0548, 105.0703, 117.0698, 133.1020, 161.0967, 179.1088	n.a.	2a
L-Phenylalanine	77.0381, 79.0538, 91.0539, 103.0540, 120.0806	n.a.	77.0387, 79.0548, 91.0546, 103.0542, 120.0808	n.a.	2a
Naphthalene-2-sulfonic acid *	n.a.	79.9576, 115.0549, 143.0503	n.a.	79.9574, 115.0553, 143.0503	2a
Octaethylene glycol	89.0603, 133.0864, 177.1127	n.a.	89.0601, 133.0867, 177.1126	n.a.	2b
Paraxanthine	n.a.	122.0365, 164.0341	n.a.	122.0357, 164.0340	2a
PEG18	89.0597, 133.0860, 177.1122	n.a.	89.0603, 133.0865, 177.1131	n.a.	2b
Piperine	115.0553, 135.0446, 143.0495, 171.0453, 201.0548	n.a.	115.0554, 135.0448, 143.0502, 171.0437, 201.0557	n.a.	1
Propylparaben sulfate	n.a.	121.0297, 137.0239, 179.0716	n.a.	121.0295, 137.0246, 179.0714	2b
Stachydrine (Proline betaine) *	58.0650, 72.0805, 84.0810, 98.0962	n.a.	58.0656, 72.0809, 84.0809, 98.0962	n.a.	2a
Thymol *	81.0705, 93.0704, 107.0859, 123.0789, 133.1013	n.a.	81.0706, 93.0705, 107.0867, 123.0801, 133.1020	n.a.	2a
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	1
Tridecalactone *	83.0850, 95.0871, 121.1006, 135.1164, 177.1632	n.a.	83.0858, 95.0859, 121.1016, 135.1173, 177.1642	n.a.	2a
Triethylphosphate	127.0158, 155.0470	n.a.	127.0154, 155.0467	n.a.	2a
Tris(2-butoxyethyl)phosphate	101.0962, 199.0731, 299.1621, 399.2511	n.a.	101.0973, 199.0733, 299.1633, 399.2499	n.a.	2a
Tryptophan *	n.a.	74.0234, 116.0494, 142.0652	n.a.	74.0248, 116.0504, 142.0666	2a

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Internal standard-corrected areas in sample prepared with Phree																					
	(+)											(-)										
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10
1,8-Epoxy-p-menthan-3-ol glucoside *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.01	0.07	0.05	0.11	0.01	0.02	0.10	0.01	0.02	0.02
25-Hydroxyvitamin D3 26,23-lactol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3.09	0.79	0.03	1.77	1.37	0.85	0.89	2.36	0.60	1.72
2-naphthylamine	n.a.	39.35	42.97	192.35	36.83	75.28	118.40	162.66	52.57	107.24	38.49	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-[2-(5-Methylthiophen-2-yl)-2-oxoethoxy]benzonitrile *	0.03	0.38	0.93	1.83	0.07	0.26	0.49	0.68	0.36	0.73	0.34	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-chlorophenol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.02	n.a.	n.a.	0.03	n.a.	n.a.	n.a.	n.a.	0.03	n.a.
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-Hydroxyquinoline *	n.a.	0.31	1.09	0.59	0.16	0.98	0.28	0.91	0.01	0.24	0.10	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-Nitrophenol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.16	7.36	10.36	1.32	26.68	1.54	5.78	10.74	1.85	9.37	23.72
4-Sulfamoylbenzoic acid *	0.54	1.29	3.30	14.09	1.37	3.69	4.48	4.55	4.03	2.34	2.11	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Acetaminophen sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.23	0.25	0.25	0.11	0.03	0.19	0.11	0.21	0.03
Azelaic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	28.60	5.38	0.80	n.a.	4.49	n.a.	n.a.	11.93	n.a.	10.98	7.42
Benzophenone-4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.03	n.a.
Caffeic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Caffeine	0.09	6.66	6.94	4.28	6.91	4.73	5.21	5.09	7.15	5.80	0.10	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Carveol *	n.a.	n.a.	0.02	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.03	0.14	0.17	1.94	0.10	0.05	0.16	0.22	0.18	0.27
Coumaric acid	10.06	0.45	2.35	0.91	8.02	3.58	9.81	8.93	10.45	3.81	4.55	n.a.	0.87	1.12	0.05	2.66	0.13	0.38	0.64	0.91	0.38	1.40
Cresol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.09	205.86	349.66	15.18	1204.4	79.21	110.20	234.48	311.49	164.21	461.50
Diethylphthalate *	n.a.	0.15	0.96	2.50	0.14	0.24	0.97	0.62	0.53	1.03	0.86	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diphenylsulfone *	4.14	n.a.	n.a.	16.20	4.89	2.27	16.69	0.00	0.27	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Internal standard-corrected areas in sample prepared with Phree																					
	(+)											(-)										
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10
Indole-3-carbinol *	0.07	8.20	8.39	8.61	11.57	9.42	10.57	11.28	17.46	19.52	12.94	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	n.a.	2.49	3.26	3.38	3.67	1.29	4.35	3.68	2.60	4.52	3.34	0.11	5.32	4.66	4.96	45.24	9.00	4.32	6.15	9.28	2.96	17.84
Indolelactic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3.72	5.69	0.22	21.65	0.51	2.65	3.97	4.63	2.23	7.49
Indoxyl sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	16.46	44.92	2.10	147.34	2.21	6.99	18.07	18.74	12.11	48.62
Isobutylparaben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Isopropylparaben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.04	21.93	28.23	13.93	96.32	3.53	22.81	24.36	13.93	12.49	15.11
Jasmonic acid *	n.a.	0.27	0.86	1.46	0.11	0.28	0.38	0.72	0.23	0.60	0.25	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Loliolid *	n.a.	0.52	1.65	2.83	0.31	0.55	0.68	1.41	0.39	1.09	0.46	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
L-Phenylalanine	0.17	112.35	69.57	26.31	155.89	123.38	77.64	57.34	49.87	44.91	89.60	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Naphthalene-2-sulfonic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.36	1.42	0.98	0.47	0.72	1.23	0.14	1.00	0.19	0.31
Octaethylene glycol	1.21	12.52	12.82	15.29	10.89	12.08	22.38	26.47	10.86	23.02	7.23	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Paraxanthine	0.26	62.39	29.70	25.64	63.17	45.42	55.06	49.87	31.55	44.95	2.92	n.a.	0.94	0.20	0.07	1.00	0.47	0.19	0.22	0.71	0.28	1.01
PEG18	n.a.	0.67	1.26	0.65	0.71	0.93	0.68	0.91	0.92	1.39	1.04	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Piperine	n.a.	n.a.	0.12	0.18	0.22	0.02	1.77	0.14	2.02	0.08	0.09	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.51	1.50	0.33	n.a.	0.15	0.98	0.59	0.30	0.38	0.36
Stachydrine (Proline betaine) *	1.10	1.43	0.12	0.24	0.31	n.a.	19.25	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Thymol *	n.a.	0.24	0.75	1.34	0.13	0.26	0.38	0.66	0.21	0.48	0.23	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.28	n.a.	n.a.	n.a.	0.93	0.22	3.23
Tridecalactone *	n.a.	0.41	2.86	6.01	0.51	0.79	1.34	2.53	0.74	2.02	0.81	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triethylphosphate	0.03	n.a.	0.08	0.28	0.02	0.03	0.19	0.12	0.09	0.09	0.04	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tris(2-butoxyethyl) phosphate	1.36	n.a.	0.88	5.52	n.a.	15.69	265.09	0.30	n.a.	0.62	0.26	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tryptophan *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.12	70.10	59.35	9.81	248.85	22.85	47.38	69.01	105.63	46.94	131.19

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Internal standard-corrected areas in sample prepared with PPT																					
	(+)											(-)										
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10
1,8-Epoxy-p-menthan-3-ol glucoside *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.06	2.81	0.90	2.54	5.70	2.10	0.47	0.08	0.15	2.46
25-Hydroxyvitamin D3 26,23-lactol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	5.86	5.27	2.96	5.34	7.53	5.51	5.51	5.57	6.16	6.39
2-naphthylamine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-[2-(5-Methylthiophen-2-yl)-2-oxoethoxy]benzonitrile *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-chlorophenol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-Hydroxyquinoline *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-Nitrophenol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.34	0.48	n.a.	n.a.	0.10	n.a.	0.09	n.a.	0.25	n.a.	0.01
4-Sulfamoylbenzoic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Acetaminophen sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.93	2.75	0.36	3.95	0.12	0.89	0.15	0.28	0.10
Azelaic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	57.47	25.64	13.45	21.42	17.28	26.65	20.48	44.99	13.32	15.89
Benzophenone-4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.56	0.01	n.a.	4.28	0.04	0.02	0.01	0.10	0.02
Caffeic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.67	1.20	0.28	0.31	2.03	0.60	0.55	2.09	1.15	1.72
Caffeine	n.a.	6.61	7.85	5.85	8.27	7.28	7.57	7.09	8.70	7.84	0.12	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Carveol *	n.a.	0.09	0.25	0.50	0.08	0.08	0.24	0.23	0.25	0.39	0.17	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.15	2.73	0.51	7.02	410.7	4.54	0.55	0.70	0.34	1.25
Coumaric acid	n.a.	10.88	8.98	5.03	20.11	14.51	9.61	15.98	23.79	17.04	18.80	n.a.	3.59	1.25	0.46	2.76	1.60	1.97	1.69	1.81	1.71	2.23
Cresol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.17	790.9	1142	626.1	1030	1769	742.9	1092	1118	1647	1916
Diethylphthalate *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diphenylsulfone *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Internal standard-corrected areas in sample prepared with PPT																					
	(+)											(-)										
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10
Indole-3-carbinol *	n.a.	11.08	15.47	12.99	21.96	53.80	11.32	14.52	22.29	20.23	17.86	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	n.a.	0.44	0.55	0.40	0.13	0.19	0.15	0.85	0.14	0.69	0.14	n.a.	3.77	2.09	1.34	2.27	2.10	1.97	3.73	4.01	2.56	2.78
Indolelactic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	27.90	16.67	8.51	25.86	22.12	23.99	29.87	24.45	29.72	32.02
Indoxyl sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	86.98	120.3	88.92	192.6	154.4	136.8	175.5	137.1	170.0	485.9
Isobutylparaben	n.a.	0.37	0.17	0.10	0.21	0.25	0.15	0.36	0.24	0.19	0.29	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Isopropylparaben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	66.74	127.2	35.56	299.3	11092	219.6	76.19	42.08	71.63	98.71
Jasmonic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Loliolid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
L-Phenylalanine	0.03	111.9	55.02	26.00	119.8	101.1	55.75	80.34	61.79	58.81	73.69	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Naphthalene-2-sulfonic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	51.15	1.92	1.36	2.41	2.05	2.01	1.90	1.60	1.71	1.31
Octaethylene glycol	0.91	13.81	5.61	5.56	4.26	3.98	7.28	23.53	0.34	6.55	3.63	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Paraxanthine	0.01	51.29	33.38	65.39	55.58	46.73	48.69	67.72	12.28	66.33	11.07	n.a.	1.92	3.72	1.27	11.05	2.89	4.44	1.31	3.48	3.93	0.82
PEG18	n.a.	0.39	0.75	0.12	0.53	0.59	0.54	0.74	0.38	0.44	0.45	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Piperine	n.a.	n.a.	0.12	0.13	0.20	0.02	1.91	0.16	1.10	0.11	0.11	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2.81	3.55	1.08	n.a.	438.8	5.71	1.65	0.74	1.07	1.77
Stachydrine (Proline betaine) *	0.21	35.27	3.37	3.72	10.37	0.04	85.90	0.45	12.21	37.08	1.91	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Thymol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tridecalactone *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triethylphosphate	0.04	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tris(2-butoxyethyl) phosphate	0.59	n.a.	n.a.	n.a.	n.a.	0.02	n.a.	n.a.	0.04	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tryptophan *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	467.0	341.9	205.5	411.2	559.2	343.3	479.5	504.8	481.0	519.9

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Fold changes (Area Phree / Area PPT)																					
	(+)											(-)										
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10
1,8-Epoxy-p-menthan-3-ol glucoside *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	7.9	0.2	Inf	0.1	Inf	0.1	1.4	2.7	6.0	0.0
25-Hydroxyvitamin D3 26,23-lactol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2.4	1.0	0.1	1.0	1.3	1.0	1.0	5.7	0.8	1.2
2-naphthylamine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-[2-(5-Methylthiophen-2-yl)-2-oxoethoxy]benzonitrile *	n.a.	5.4	5.6	44.2	8.6	2.9	3.6	4.9	11.4	9.1	8.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-chlorophenol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.7	2.5	n.a.	n.a.	156.1	3.2	5.8	65.0	n.a.	19.3	73.2
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Inf	n.a.	n.a.	n.a.	Inf	n.a.	n.a.	n.a.	n.a.
4-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2.6	2.7	20.5	0.8	0.0	0.8	2.4	6.6	2.6	0.6
4-Hydroxyquinoline *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	15.6	2.0	23.1	14.4	6.5	31.1	36.0	11.3	27.9	19.7
4-Nitrophenol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4-Sulfamoylbenzoic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Acetaminophen sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Azelaic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.8	Inf	6.9	0.2	Inf	1.4	Inf	5.9	Inf
Benzophenone-4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Caffeic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Caffeine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Inf	Inf	n.a.	Inf	Inf	Inf	Inf	2.7	Inf
Carveol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Chavicol sulfate	n.a.	3.1	2.8	6.0	7.4	3.3	1.8	2.8	3.8	2.0	Inf	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Coumaric acid	n.a.	n.a.	40.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Cresol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	10.0	0.4	3.8	0.8	0.0	0.1	1.9	4.3	4.1	0.9
Diethylphthalate *	n.a.	3.1	4.6	31.7	9.9	4.8	5.8	4.8	4.9	2.3	2.8	n.a.	1.1	6.2	1.3	2.8	0.6	1.3	2.4	6.8	1.7	2.8
Diphenylsulfone *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	11.9	1.1	2.1	0.2	3.4	0.3	1.0	1.4	3.8	0.8	1.0

Table A5a – (continued) Application to cohort samples (Serum-Pelagie): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Fold changes (Area Phree / Area PPT)																					
	(+)											(-)										
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Mean 9	Mean 10
Indole-3-carbinol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indolelactic acid *	n.a.	2.3	1.7	5.5	4.7	0.9	2.6	3.1	3.7	2.7	2.6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indoxyl sulfate	n.a.	18.1	19.6	70.5	261.1	35.7	81.8	17.3	90.0	18.9	88.5	n.a.	6.0	14.3	3.1	56.7	2.6	13.7	10.1	30.7	8.1	27.2
Isobutylparaben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.6	2.4	0.3	2.4	0.2	0.7	0.9	2.6	0.6	1.0
Isopropylparaben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.8	2.6	0.3	2.2	0.1	0.3	0.7	1.8	0.6	0.4
Jasmonic acid *	n.a.	Inf	Inf	36.8	Inf	Inf	Inf	Inf	Inf	Inf	Inf	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Loliolid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.5	1.5	4.3	0.9	0.0	0.7	2.0	4.5	1.3	0.7
L-Phenylalanine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Naphthalene-2-sulfonic acid *	2.6	0.7	4.2	22.1	5.7	1.4	5.5	2.9	3.9	2.2	4.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Octaethylene glycol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	18.5	Inf	Inf	Inf	0.2	Inf	42.4	Inf	Inf	Inf	Inf
Paraxanthine	0.6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	37.1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
PEG18	8.9	3.8	2.9	3.3	10.3	5.0	3.2	2.9	12.4	1.9	0.6	n.a.	2.2	0.4	0.6	0.3	Inf	0.3	1.1	2.8	0.6	5.4
Piperine	n.a.	n.a.	3.1	11.9	23.0	5.7	2.6	5.7	8.7	2.2	2.1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.8	3.0	3.4	4.2	0.0	1.1	2.3	5.4	2.8	0.9
Stachydrine (Proline betaine) *	2.5	18.6	n.a.	n.a.	155.8	n.a.	0.6	n.a.	Inf	Inf	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Thymol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.4	n.a.	n.a.
Tridecalactone *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triethylphosphate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tris(2-butoxyethyl) phosphate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Tryptophan *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.7	1.2	0.5	1.8	0.3	0.9	0.9	2.8	0.8	1.1

2.7. Table A5b – Application to cohort samples (Plasma-Danish cohort): Annotations and semi-quantification

Table A5b – Application to cohort samples (Plasma – Danish cohort): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	SMILES	CI m/z		CI Rt								CI isotopic fit		Global CI	
				Experimental		RTI-predicted		Retip-predicted		logP-predicted		CI overall			
		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
2-Hydroxybenzoic Acid	OC(=O)c1ccccc1O	n.a.	1.00	n.a.	n.a.	n.a.	**0.89**	n.a.	n.a.	n.a.	0.23	n.a.	n.a.	n.a.	G2_0.94
2-Methoxyacetophenone	COCC(=O)c1ccccc1	0.95	n.a.	n.a.	n.a.	**1.00**	n.a.	0.90	n.a.	0.99	n.a.	0.85	n.a.	G3_0.93	n.a.
2-naphthylamine	Nc1ccc2ccccc2c1	0.94	n.a.	n.a.	n.a.	**0.00**	n.a.	n.a.	n.a.	n.a.	n.a.	0.75	n.a.	G3_0.56	n.a.
3,4,5-trimethoxycinnamic acid	COc1cc(C=CC(O)=O)cc(OC)c1OC	0.77	n.a.	n.a.	n.a.	**0.93**	n.a.	0.47	n.a.	0.99	n.a.	0.85	n.a.	G3_0.85	n.a.
3-hydroxybenzoic acid	OC(=O)c1cccc(O)c1	n.a.	1.00	n.a.	n.a.	n.a.	**0.42**	n.a.	n.a.	n.a.	0.39	n.a.	n.a.	n.a.	G2_0.7
4-hydroxy-2,5,6-trichloroisophthalonitrile	Oc1c(Cl)c(Cl)c(C#N)c(Cl)c1C#N	n.a.	0.98	n.a.	n.a.	n.a.	n.a.	n.a.	**0.44**	n.a.	0.46	n.a.	0.99	n.a.	G3_0.8
4-hydroxybenzoic acid	OC(=O)c1ccc(O)cc1	n.a.	0.95	n.a.	n.a.	n.a.	**0.64**	n.a.	0.69	n.a.	0.53	n.a.	n.a.	n.a.	G2_0.8
Acetaminophen glucuronide	CC(=O)NC1=CC=C(C=C1)OC2C(C(C(O2)C(=O)O)O)O)O	0.92	0.85	**0.92**	**0.81**	0.92	0.94	0.92	0.94	n.a.	n.a.	0.73	n.a.	G3_0.86	G2_0.83
Azelaic acid *	C(CCCC(=O)O)CCCC(=O)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Bourbonal	CCOc1cc(C=O)ccc1O	0.98	n.a.	n.a.	n.a.	**0.31**	n.a.	n.a.	n.a.	0.08	n.a.	0.69	n.a.	G3_0.66	n.a.
Bupivacaine	CCCCN1CCCCC1C(=O)Nc2c(C)cccc2C	0.79	n.a.	**0.96**	n.a.	0.43	n.a.	0.29	n.a.	n.a.	n.a.	0.79	n.a.	G3_0.85	n.a.
Caffeine	Cn1cnc2N(C)C(=O)N(C)C(=O)c12	0.73	n.a.	**0.96**	n.a.	0.91	n.a.	0.63	n.a.	0.80	n.a.	0.81	n.a.	G3_0.83	n.a.
Carveol	CC(=C)C1CC=C(C)C(O)C1	0.88	n.a.	n.a.	n.a.	**0.52**	n.a.	n.a.	n.a.	0.07	n.a.	n.a.	n.a.	G2_0.7	n.a.
Chavicol sulfate	C=CCC1=CC=C(C=C1)OS(=O)(=O)O	n.a.	0.78	n.a.	**0.90**	n.a.	0.36	n.a.	0.57	n.a.	0.31	n.a.	n.a.	n.a.	G2_0.84
Cotinine	CN1C(CCC1=O)c2cccnc2	0.95	0.95	**0.89**	**0.66**	0.97	0.93	0.81	0.89	0.99	0.97	0.64	n.a.	G3_0.83	G2_0.81
Cresol sulfate	CC1=CC=CC=C1OS(=O)(=O)O	n.a.	0.74	n.a.	**0.93**	n.a.	0.95	n.a.	0.59	n.a.	0.74	n.a.	0.82	n.a.	G3_0.83
Curcumenol *	CC1CCC2C13CC(=C(C)C)C(O3)(C=C2C)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diazepam	CN1C(=O)CN=C(c2ccccc2)c3cc(Cl)ccc13	0.70		n.a.	n.a.	**0.99**	n.a.	0.54	n.a.	0.91	n.a.	0.95	n.a.	G3_0.88	n.a.
Diethyl phthalate	CCOC(=O)c1ccccc1C(=O)OCC	0.88	0.92	n.a.	n.a.	**0.94**	**0.96**	0.62	0.63	0.80	0.81	0.81	n.a.	G3_0.88	G2_0.94
Docosahexaenoic acid	CCCCCCCCCC=CC=CC=CC=CC=CC(O)=O	0.96	0.76	**1.00**	**1.00**	0.79	0.80	0.93	0.94	0.35	0.35	0.92	n.a.	G3_0.96	G2_0.88
Eicosapentaenoic acid	CCCCCCCCCC=CC=CC=CC=CC(O)=O	n.a.	0.81	n.a.	n.a.	n.a.	**0.77**	n.a.	0.84	n.a.	0.46	n.a.	n.a.	n.a.	G2_0.79
Ethyl paraben	CCOC(=O)c1ccc(O)cc1	n.a.	0.93	n.a.	**1.00**	n.a.	0.13	n.a.	0.01	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.96
Ibuprofen	CC(C)Cc1ccc(cc1)C(C)C(O)=O	n.a.	0.91	n.a.	**0.96**	n.a.	n.a.	n.a.	0.59	n.a.	0.81	n.a.	n.a.	n.a.	G2_0.94
Indole-3-carbinol *	C1=CC=C2C(=C1)C(=CN2)CO	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	C1=CC=C2C(=C1)C(=CN2)C=O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indolelactic acid *	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)O	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indoxyl sulfate	C1=CC=C2C(=C1)C(=CN2)OS(=O)(=O)O	n.a.	0.73	n.a.	**0.82**	n.a.	n.a.	n.a.	n.a.	n.a.	0.50	n.a.	0.90	n.a.	G3_0.82

Table A5b – (continued) Application to cohort samples (Plasma – Danish cohort): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	SMILES	CI m/z		CI Rt								CI isotopic fit CI overall		Global CI	
				Experimental		RTI-predicted		Retip-predicted		logP-predicted					
		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
Nordazepam	Clc1ccc2NC(=O)CN=C(c3ccccc3)c2c1	0.89	0.83	n.a.	n.a.	**0.90**	**0.87**	0.47	0.45	0.89	0.87	0.86	0.72	G3_0.88	G3_0.81
Octaethylene glycol	OCCOCCOCCOCCOCCOCCOCCOCCO	0.80	n.a.	n.a.	n.a.	n.a.	n.a.	**0.78**	n.a.	n.a.	n.a.	0.82	n.a.	G3_0.80	n.a.
Oxazepam	OC1N=C(c2ccccc2)c3cc(Cl)ccc3NC1=O	0.70	n.a.	**1.00**	n.a.	0.45	n.a.	0.29	n.a.	0.88	n.a.	n.a.	n.a.	G2_0.85	n.a.
Paracetamol	CC(=O)Nc1ccc(O)cc1	0.77	1.00	**0.78**	**0.62**	0.90	0.89	0.80	0.80	0.93	0.93	n.a.	n.a.	G2_0.77	G2_0.81
Paraxanthine	Cn1cnc2NC(=O)N(C)C(=O)c12	n.a.	0.93	n.a.	n.a.	n.a.	**0.74**	n.a.	0.76	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.83
Pentachlorophenol	Oc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	n.a.	0.73	n.a.	n.a.	n.a.	**0.86**	n.a.	0.22	n.a.	0.94	n.a.	0.92	n.a.	G3_0.83
Phenol sulfate	C1=CC=C(C=C1)OS(=O)(=O)O	n.a.	0.89	n.a.	**0.85**	n.a.	n.a.	n.a.	0.79	n.a.	0.78	n.a.	0.78	n.a.	G3_0.84
Piperine	O=C(/C=C/C=C/c1ccc2OCOc2c1)N3CCCCC3	0.93	n.a.	**0.96**	n.a.	0.26	n.a.	0.71	n.a.	0.64	n.a.	0.91	n.a.	G3_0.93	n.a.
Propylparaben sulfate	CCCOC(=O)C1=CC=C(C=C1)OS(=O)(=O)O	n.a.	0.98	n.a.	**0.94**	n.a.	0.54	n.a.	0.59	n.a.	0.20	n.a.	n.a.	n.a.	G2_0.96
Theobromine	Cn1cnc2N(C)C(=O)NC(=O)c12	0.79	n.a.	n.a.	n.a.	**0.65**	n.a.	0.49	n.a.	n.a.	n.a.	0.89	n.a.	G3_0.78	n.a.
Trans-3-hydroxycotinine	CN1C(CC(O)C1=O)c2cccnc2	0.90	n.a.	n.a.	n.a.	n.a.	n.a.	**0.84**	n.a.	0.98	n.a.	n.a.	n.a.	G2_0.87	n.a.
Triclosan glucuronide	C1=CC(=C(C=C1Cl)OC2C(C(C(C(O2)C(=O)O)O)O)O)OC3=C(C=C(C=C3)Cl)Cl	n.a.	0.86	n.a.	**0.99**	n.a.	0.51	n.a.	n.a.	n.a.	0.89	n.a.	0.88	n.a.	G3_0.91
Triclosan sulfate	C1=CC(=C(C=C1Cl)OS(=O)(=O)O)OC2=C(C=C(C=C2)Cl)Cl	n.a.	0.84	n.a.	**0.96**	n.a.	n.a.	n.a.	n.a.	n.a.	0.88	n.a.	0.96	n.a.	G3_0.92
Tryptophan	N[C@@H](Cc1c[nH]c2ccccc12)C(O)=O	0.90	0.78	**0.65**	**0.66**	n.a.	n.a.	0.23	0.23	n.a.	n.a.	0.92	n.a.	G3_0.82	G2_0.72

Table A5b – (continued) Application to cohort samples (Plasma – Danish cohort): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	MS/MS				Confidence level
	Theoretical fragments		Experimental fragments		
	(+)	(-)	(+)	(-)	
2-Hydroxybenzoic Acid	n.a.	93.0343	n.a.	93.0345	2a
2-Methoxyacetophenone	63.0229, 79.0542, 105.0335, 119.0491, 133.0648	n.a.	63.0226, 79.0537, 105.0333, 119.0491, 133.0647	n.a.	2b
2-naphthylamine	91.0556, 115.0542, 117.0699, 127.0542	n.a.	91.0547, 115.0541, 117.0690, 127.0542	n.a.	2a
3,4,5-trimethoxycinnamic acid	107.0491, 137.0597, 149.0597, 161.0597, 177.0546, 193.0859, 221.0808	n.a.	107.0490, 137.0604, 149.0601, 161.0602, 177.0549, 193.0862, 221.0810	n.a.	2a
3-hydroxybenzoic acid	n.a.	93.0343	n.a.	93.0345	2a
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	146.9765, 174.9704, 181.9447, 209.9401	n.a.	146.9769, 174.9708, 181.9449, 209.9407	2a
4-hydroxybenzoic acid	n.a.	93.0343	n.a.	93.0347	2a
Acetaminophen glucuronide	110.0607, 134.0606, 152.0712	175.0252, 150.0561, 113.0252	110.0607, 134.0606, 152.0712	175.0248, 150.0561, 113.0245	1
Azelaic acid *	n.a.	97.0655, 123.0811, 125.0970	n.a.	97.0660, 123.0816, 125.0972	2a
Bourbonal	121.0290, 139.0395, 149.0603	n.a.	121.0287, 139.0395, 149.0601	n.a.	2a
Bupivacaine	140.1445	n.a.	140.1434	n.a.	1
Caffeine	83.0609, 110.0708, 123.0417, 138.0659, 195.0881	n.a.	83.0602, 110.0707, 123.0421, 138.0653, 195.0871	n.a.	1
Carveol	107.0855, 119.0855, 135.1168	n.a.	107.0857, 119.0857, 135.1172	n.a.	2a
Chavicol sulfate	n.a.	105.0710, 133.0659	n.a.	105.0710, 133.0656	2b
Cotinine	106.0633, 118.0646, 120.0794	n.a.	106.0642, 118.0653, 120.0801	n.a.	1
Cresol sulfate	n.a.	92.0279, 107.0493	n.a.	92.0270, 107.0501	2a
Curcumenol *	93.0698, 105.0698, 119.0855, 133.1010, 175.1116	n.a.	93.0700, 105.0700, 119.0856, 133.1016, 175.1124	n.a.	2a
Diazepam	154.0408, 193.0879, 222.1146, 228.0569, 257.0837	n.a.	154.0418, 193.0885, 222.1154, 228.0579, 257.0847	n.a.	2a
Diethyl phthalate	93.0326, 111.0437, 121.0284, 149.0233, 177.0546	71.0502, 121.0296, 134.0374, 149.0972, 177.0921	93.0334, 111.0444, 121.0282, 149.0245, 177.0553	71.0501, 121.0292, 134.0365, 149.0970, 177.0917	2a
Docosahexaenoic acid	119.0848, 131.0847, 145.0989, 161.1313, 175.1434, 269.2256, 293.2272, 311.2344	229.1958, 283.2446	119.0854, 131.0850, 145.0999, 161.1323, 175.1444, 269.2267, 293.2272, 311.2354	229.1962, 283.2437	1
Eicosapentaenoic acid	n.a.	203.1802, 229.1957, 257.2274	n.a.	203.1811, 229.1967, 257.2276	2a
Ethyl paraben	n.a.	92.0272, 137.0244	n.a.	92.0269, 137.0242	2a
Ibuprofen	n.a.	161.1332	n.a.	161.1334	1
Indole-3-carbinol *	77.0380, 103.0555	n.a.	77.0383, 103.0547	n.a.	2a
Indole-3-carboxaldehyde *	n.a.	115.0422, 126.0354	n.a.	115.0421, 126.0345	2a
Indolelactic acid *	n.a.	72.9947, 116.0486, 130.0661, 142.0633, 158.0625, 186.0553	n.a.	72.9932, 116.0491, 130.0661, 142.0642, 158.0619, 186.0560	2a
Indoxyl sulfate	n.a.	79.9578, 132.0460	n.a.	79.9572, 132.0452	2a
Nordazepam	140.0252, 165.0201, 208.0986, 226.0406, 243.0677	241.0299	140.0261, 165.0207, 208.0997, 226.0416, 243.0686	241.0302	2a
Octaethylene glycol	89.0603, 133.0864, 177.1127	n.a.	89.0601, 133.0861, 177.1128	n.a.	2b
Oxazepam	231.0668, 241.0516, 269.0464	n.a.	231.0674, 241.0524, 269.0470	n.a.	2a
Paracetamol	110.0608	107.0366	110.0602	107.0372	1
Paraxanthine	n.a.	122.0365, 164.0341	n.a.	122.0362, 164.0341	2a
Pentachlorophenol	n.a.		n.a.	n.a.	3
Phenol sulfate	n.a.	79.9551, 93.0325	n.a.	79.9554, 93.0326	2a
Piperine	115.0544, 135.0441, 143.0491, 171.0446, 201.0548	n.a.	115.0540, 135.0445, 143.0493, 171.0442, 201.0543	n.a.	1
Propylparaben sulfate	n.a.	179.0715	n.a.	179.0714	2b
Theobromine	108.0554, 110.0713, 122.0589, 138.0668, 163.0611	n.a.	108.0554, 110.0710, 122.0583, 138.0660, 163.0614	n.a.	2a
Trans-3-hydroxycotinine	80.0493, 86.0606, 106.0676, 118.0674, 134.0602, 149.0714	n.a.	80.0495, 86.0600, 106.0666, 118.0664, 134.0601, 149.0709	n.a.	2a
Triclosan glucuronide	n.a.	286.9448	n.a.	286.9452	1
Triclosan sulfate	n.a.	286.9448	n.a.	286.9445	1
Tryptophan	118.0650, 146.0596, 159.0912, 170.0596, 188.0700	116.0500, 142.0655, 159.0915	118.0646, 146.0592, 159.0915, 170.0599, 188.0702	116.0506, 142.0659, 159.0922	1

Table A5b – (continued) Application to cohort samples (Plasma – Danish cohort): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Internal standard-corrected areas in sample prepared with Phree																	
	(+)									(-)								
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8
2-Hydroxybenzoic Acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.03	0.04	0.05	n.a.	n.a.	0.04	3.65	0.10
2-Methoxyacetophenone	n.a.	2.14	2.69	2.63	0.70	0.03	2.25	0.03	2.88	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2-naphthylamine	n.a.	4.78	5.54	6.47	6.36	6.58	5.19	4.15	5.95	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3,4,5-trimethoxycinnamic acid	n.a.	14.25	2.09	2.08	2.00	1.72	2.16	5.32	2.70	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.98	0.68	1.12	1.20	0.97	0.78	0.41	110.96	1.08
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.11	0.07	0.23	0.13	0.09	0.15	0.06	0.27
4-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3.35	2.28	1.84	8.02	5.72	2.78	4.09	2.30
Acetaminophen glucuronide	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	8.93	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	180.77	n.a.
Azelaic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	158.88	n.a.	n.a.	n.a.	9.87	6.09	n.a.	n.a.	n.a.
Bourbonal	0.02	0.13	0.10	0.00	0.04	0.00	n.a.	n.a.	0.01	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Bupivacaine	0.08	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.82	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Caffeine	n.a.	5.63	5.99	0.47	5.05	5.51	5.96	1.45	6.18	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Carveol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	23.23	93.86	118.79	n.a.	n.a.	n.a.	98.85	2.56	142.35
Cotinine	n.a.	0.00	n.a.	0.00	0.00	0.00	n.a.	0.00	n.a.	n.a.	0.00	n.a.	0.00	0.00	0.00	n.a.	0.00	n.a.
Cresol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	184.24	217.61	744.70	1230.8	64.53	189.51	408.03	254.03
Curcumenol *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diazepam	n.a.	46.41	25.50	0.01	n.a.	n.a.	30.29	33.26	26.80	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diethyl phthalate	n.a.	17.93	23.91	7.14	5.51	7.69	18.57	9.06	25.28	0.02	0.62	0.76	0.28	0.87	0.37	0.71	0.30	0.86
Docosahexaenoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	34.06	112.43	50.07	78.96	41.56	45.68	15.56	36.69
Eicosapentaenoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	11.47	37.61	23.78	15.25	21.03	28.36	3.18	16.92
Ethyl paraben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.06	6.09	7.53	3.37	7.88	3.67	6.34	3.96	8.45
Ibuprofen	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.31	0.71	1.21	n.a.	n.a.	n.a.
Indole-3-carbinol *	n.a.	0.90	0.45	1.60	2.10	1.46	1.24	0.36	1.18	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	n.a.	0.76	0.71	2.43	1.75	1.95	0.63	0.85	0.75	n.a.	4.63	4.96	12.30	10.86	11.27	4.12	5.66	5.23
Indolelactic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	5.96	6.69	6.67	7.34	4.53	5.86	2.69	3.98
Indoxyl sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	20.99	5.31	77.15	114.70	112.24	42.36	17.26	28.47
Nordazepam	n.a.	1.33	2.61	0.00	0.00	0.00	1.77	0.00	1.90	n.a.	0.06	0.11	n.a.	n.a.	n.a.	0.10	0.00	0.07
Octaethylene glycol	0.00	0.10	0.09	0.21	0.09	0.25	0.10	0.92	0.10	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Oxazepam	n.a.	0.00	0.00	n.a.	0.00	n.a.	0.00	0.00	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Paracetamol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.77	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	71.53	n.a.
Paraxanthine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.01	26.86	35.32	0.99	15.27	14.11	28.57	0.84	36.76
Pentachlorophenol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phenol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	64.51	85.70	41.83	41.01	23.59	70.35	34.97	91.34
Piperine	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Theobromine	n.a.	15.16	24.90	0.95	1.86	15.10	15.74	8.52	19.47	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Trans-3-hydroxycotinine	n.a.	n.a.	0.00	0.00	0.00	0.00	n.a.	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan glucuronide	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.54	0.56	0.01	0.01	0.04	0.53	n.a.	0.70
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	n.a.	0.00	0.00	n.a.	0.00
Tryptophan	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A5b – (continued) Application to cohort samples (Plasma – Danish cohort): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Internal standard-corrected areas in sample prepared with PPT																	
	(+)									(-)								
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8
2-Hydroxybenzoic Acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.02	0.02	0.02	n.a.	n.a.	0.02	1.94	0.04
2-Methoxyacetophenone	n.a.	2.11	3.11	2.62	0.75	0.04	2.39	0.03	3.46	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2-naphthylamine	4.22	1.18	2.73	2.03	2.27	2.44	1.39	2.08	3.48	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3,4,5-trimethoxycinnamic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.41	0.46	0.40	0.32	0.54	0.35	73.01	0.45
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.05	0.05	0.10	0.07	0.05	0.06	0.04	0.08
4-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.32	1.53	0.63	6.81	3.08	1.17	1.06	1.70
Acetaminophen glucuronide	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	34.19	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	489.91	n.a.
Azelaic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	4.85	32.14	26.29	61.64	166.14	186.75	23.30	23.67	34.51
Bourbonal	0.14	0.18	0.24	0.31	0.05	0.37	n.a.	n.a.	0.20	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Bupivacaine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.16	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Caffeine	0.00	5.80	6.42	0.38	5.19	5.39	5.82	1.36	6.59	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Carveol	n.a.	0.02	0.03	0.01	0.01	0.01	0.03	0.02	0.03	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	103.67	138.91	24.29	16.84	19.68	109.84	23.13	152.04
Cotinine	n.a.	0.00	n.a.	0.00	0.00	0.00	n.a.	0.00	n.a.	n.a.	0.00	n.a.	0.00	0.00	0.00	n.a.	0.00	n.a.
Cresol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	162.10	204.16	828.17	1455.9	78.15	162.02	324.58	230.76
Curcumenol *	n.a.	0.40	0.50	0.36	0.28	0.29	0.50	0.74	0.79	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diazepam	n.a.	41.36	32.66	0.02	n.a.	n.a.	33.30	31.73	34.52	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diethyl phthalate	n.a.	14.14	12.42	0.23	1.50	0.89	13.29	2.51	16.69	0.08	0.63	0.48	0.16	0.31	0.31	0.20	0.19	0.81
Docosahexaenoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	106.06	150.86	71.19	88.20	53.23	109.58	100.40	104.32
Eicosapentaenoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.01	39.83	52.36	65.33	23.89	22.71	24.47	12.71	60.25
Ethyl paraben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.06	4.55	6.28	3.52	10.23	4.68	4.48	2.77	7.04
Ibuprofen	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.20	0.07	3.70	1.40	2.13	n.a.	0.05	0.05
Indole-3-carbinol *	n.a.	1.85	2.66	1.58	2.43	1.64	2.07	1.98	2.91	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	n.a.	0.52	0.69	2.02	1.02	1.19	0.57	0.70	0.70	n.a.	2.76	3.81	12.97	5.75	8.54	3.00	3.52	4.20
Indolelactic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	13.06	16.41	10.64	11.05	13.21	12.26	6.26	16.86
Indoxyl sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.07	68.58	81.74	116.61	158.38	162.49	66.32	98.17	87.99
Nordazepam	n.a.	2.35	3.65	0.00	0.01	0.00	2.81	0.00	4.13	n.a.	0.12	0.18	n.a.	n.a.	n.a.	0.16	0.00	0.20
Octaethylene glycol	n.a.	0.08	0.07	0.04	0.06	0.24	0.06	0.81	0.08	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Oxazepam	n.a.	0.00	0.00	n.a.	0.00	n.a.	0.00	0.00	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Paracetamol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3.03	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	66.36	n.a.
Paraxanthine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.01	28.66	36.92	1.20	21.34	19.21	29.21	0.89	39.56
Pentachlorophenol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phenol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	56.51	82.32	45.92	48.15	27.34	63.58	31.69	85.30
Piperine	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Theobromine	n.a.	11.72	15.05	0.62	1.24	11.71	12.16	7.75	17.26	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Trans-3-hydroxycotinine	n.a.	n.a.	0.00	0.00	0.00	0.00	n.a.	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan glucuronide	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.42	0.71	0.01	0.01	0.03	0.53	n.a.	0.71
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.00	0.00	0.00	n.a.	0.00	0.00	n.a.	0.00
Tryptophan	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	n.a.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A5b – (continued) Application to cohort samples (Plasma – Danish cohort): Annotations and semi-quantification. Red asterisks indicate compounds annotated through manual annotation (i.e. without confidence indices-CI values)

Annotation	Fold changes (Area Phree / Area PPT)																	
	(+)									(-)								
	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8	Blank	Mean 1	Mean 2	Mean 3	Mean 4	Mean 5	Mean 6	Mean 7	Mean 8
2-Hydroxybenzoic Acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2.2	1.7	2.8	n.a.	n.a.	2.1	1.9	2.7
2-Methoxyacetophenone	n.a.	1.0	0.9	1.0	0.9	0.7	0.9	1.0	0.8	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2-naphthylamine	n.a.	4.1	2.0	3.2	2.8	2.7	3.7	2.0	1.7	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3,4,5-trimethoxycinnamic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.7	2.4	3.0	3.0	1.4	1.2	1.5	2.4
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2.0	1.4	2.2	1.9	1.9	2.4	1.6	3.3
4-hydroxybenzoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2.5	1.5	2.9	1.2	1.9	2.4	3.9	1.3
Acetaminophen glucuronide	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.4	n.a.
Azelaic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	32.8	Inf	Inf	Inf	0.1	0.0	Inf	Inf	Inf
Bourbonal	0.1	0.7	0.4	0.0	0.7	0.0	n.a.	n.a.	0.0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Bupivacaine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.7	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Caffeine	n.a.	1.0	0.9	1.2	1.0	1.0	1.0	1.1	0.9	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Carveol	n.a.	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.9	0.9	Inf	Inf	Inf	0.9	0.1	0.9
Cotinine	n.a.	0.7	n.a.	0.3	0.5	0.4	n.a.	1.0	n.a.	n.a.	0.4	n.a.	0.8	0.9	0.8	n.a.	0.4	n.a.
Cresol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.1	1.1	0.9	0.8	0.8	1.2	1.3	1.1
Curcumenol *	n.a.	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diazepam	n.a.	1.1	0.8	0.9	n.a.	n.a.	0.9	1.0	0.8	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Diethyl phthalate	n.a.	1.3	1.9	31.0	3.7	8.6	1.4	3.6	1.5	0.2	1.0	1.6	1.8	2.8	1.2	3.5	1.6	1.1
Docosahexaenoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.3	0.7	0.7	0.9	0.8	0.4	0.2	0.4
Eicosapentaenoic acid	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.3	0.7	0.4	0.6	0.9	1.2	0.3	0.3
Ethyl paraben	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.0	1.3	1.2	1.0	0.8	0.8	1.4	1.4	1.2
Ibuprofen	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Inf	Inf	0.4	0.5	0.6	n.a.	Inf	Inf
Indole-3-carbinol *	n.a.	0.5	0.2	1.0	0.9	0.9	0.6	0.2	0.4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indole-3-carboxaldehyde *	n.a.	1.5	1.0	1.2	1.7	1.6	1.1	1.2	1.1	n.a.	1.7	1.3	0.9	1.9	1.3	1.4	1.6	1.2
Indolelactic acid *	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.5	0.4	0.6	0.7	0.3	0.5	0.4	0.2
Indoxyl sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.3	0.1	0.7	0.7	0.7	0.6	0.2	0.3
Nordazepam	n.a.	0.6	0.7	0.9	0.2	0.2	0.6	0.7	0.5	n.a.	0.5	0.6	n.a.	n.a.	n.a.	0.6	1.3	0.4
Octaethylene glycol	n.a.	1.2	1.3	5.4	1.7	1.1	1.7	1.1	1.2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Oxazepam	n.a.	1.1	1.8	n.a.	15.5	n.a.	2.1	1.6	3.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Paracetamol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.1	n.a.
Paraxanthine	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.9	0.9	1.0	0.8	0.7	0.7	1.0	0.9	0.9
Pentachlorophenol	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.3	0.3	0.3	0.3	0.6	0.2	0.3	0.2
Phenol sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.1	1.0	0.9	0.9	0.9	1.1	1.1	1.1
Piperine	n.a.	1.4	1.8	2.8	3.5	3.1	1.6	2.4	1.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.5	0.6	1.0	0.9	0.9	0.6	0.9	0.6
Theobromine	n.a.	1.3	1.7	1.5	1.5	1.3	1.3	1.1	1.1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Trans-3-hydroxycotinine	n.a.	n.a.	1346.6	8.4	7.7	7.4	n.a.	1.8	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Triclosan glucuronide	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.3	0.8	0.9	1.1	1.2	1.0	n.a.	1.0
Triclosan sulfate	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	7.8	4.4	4.4	n.a.	4.5	3.5	n.a.	5.8
Tryptophan	n.a.	0.3	0.7	0.7	0.8	0.6	0.5	0.8	0.9	n.a.	0.8	0.7	0.8	0.7	0.6	0.7	0.1	0.7

2.8. Table A6 – Phree and PPT methods detection limits on 30 xenobiotics

Table A6 – Methods detection limits on 30 xenobiotics for Phree and protein precipitation (PPT)

	Serum		Plasma	
Compounds	**Phree**	**PPT**	**Phree**	**PPT**
2-Aminobenzimidazole	0.1	0.1	0.1	0.5
Acetochlor	0.1	0.1	0.1	0.1
Aflatoxin B1	0.1	0.5	0.1	0.5
Azoxystrobine	0.1	0.1	0.1	0.1
Boscalid	0.1	0.1	0.1	1
Carbamazepine	0.5	0.5	0.1	0.1
Carbendazim	0.1	0.1	0.1	0.1
Chlorpyrifos	0.5	0.5	0.5	0.1
Clothianidin	0.5	0.5	1	0.1
Cotinine	0.1	0.1	0.1	0.1
Cyprodinil	0.1	0.5	0.1	1
Diazinon	0.1	0.1	0.1	0.1
Diclofenac	0.1	0.5	0.5	0.5
Fluoxetine	0.1	0.1	0.1	0.1
Ibuprofen	10	40	20	20
Imidacloprid	0.1	0.1	0.1	0.5
Ketoprofen	0.1	0.1	0.5	0.5
Malathion	0.5	1	0.1	5
Nicotine	0.1	0.1	0.1	0.1
Paracetamol	0.1	0.1	0.1	0.1
Piperine	0.1	0.1	0.1	0.1
Pravastatin	N/A	0.5	N/A	0.5
Prochloraz	0.1	0.5	0.1	1
Propiconazole	0.1	0.5	0.1	1
Sertraline	0.1	1	0.5	1
Tebuconazole	0.5	0.5	0.5	0.1
Thiacloprid	0.1	0.1	0.1	0.1
Thiamethoxam	0.1	0.1	0.5	1
Triclosan	10	20	10	20
Venlafaxine	0.1	0.5	0.1	0.1

Mean	0.9	2.3	1.2	1.9
Median	0.1	0.3	0.1	0.3

2.9. Appendix S.1 – Solvents and chemicals

Native and isotopically labeled standard compounds were purchased from suppliers Bertin, LGC, Sigma Aldrich and VWR and were stored at -20°C. Details can be found in Supporting Information (SI, Table A1). Ultrapure water was generated using a Millipore Milli-Q Gradient system. UPLC-MS-grade acetonitrile and formic acid were purchased from Biosolve (Dieuze, France). UPLC-MS-grade methanol was purchased from Carlo Erba (Val-de-Reuil, France). HPLC-MS-grade methyl tert-butyl ether (MTBE) and ethyl acetate were purchased from Fisher Scientific (Illkirch-Graffenstaden, France). Aqueous ammonia was purchased from VWR (Strasbourg, France).

2.10. Appendix S.2 – Data acquisition

Samples were analyzed on QTOF-MS (AB Sciex X500R) interfaced with an AB SCIEX ExionLC AD UPLC. Compound chromatographic separation was achieved using an Acquity UPLC HSS T3 C18 column (1.8µm, 1.0 × 150mm) maintained at 40°C. Injection volume was set at 2 µL. Flow rate was set at 100 µL/min with mobile phases of ultrapure water (A) and acetonitrile (B) both modified with 0.01% formic acid. The gradient was set as: 0-2.5 min, 10-20% B; 2.5-20 min, 20-30% B; 20-38 min, 30-45% B; 38-45 min, 45-100% B; 45-55 min, 100% B; 55-60 min, 10% B. Full-scan mass spectra was acquired in both – and + electrospray ionization (ESI) modes between 50-1100 m/z using ESI source settings: temperature 550°C, ionspray voltage 4,5kV (-4,5kV in negative mode), declustering potential 80V (-80V in negative mode), accumulation time 300 ms, spray N2 gas 35 arbitrary units, heat conduction gas 35 arbitrary units; curtain gas 7 arbitrary units, collisionally activated dissociation gas 7 arbitrary units, run time 60min. MS/MS fragmentation was performed on selected samples using sequential window acquisition of the theoretical mass spectrum (SWATH) or data dependent acquisition (DDA). SWATH experiments were performed in both – and + ESI modes, using the following source settings: MS1 accumulation time 80ms, MS2 accumulation time 30 ms, collision energy 35eV, collision energy spread 15eV, cycle time 469ms, mass range 50-1100 m/z. Acquisition windows were established for each matrix and mode using an vendor-provided automated SWATH window calculator based on results from full scan injections. DDA experiments were performed in both – and + ESI modes, using the following source settings: MS1 accumulation time 250ms, MS2 accumulation time 100ms, collision energy 35eV, cycle time 2.35s, mass range 50-1100 m/z. Precursor ion selection parameters were as follows: a maximum of 20 candidate ions per cycle, intensity threshold 1cps, and dynamic background subtraction was enabled (candidate ions only includes ions increasing in intensity).

2.11. Appendix S.3 – Quality control procedures

A solvent blank (i.e. acetonitrile/ultrapure water 90:10 (v/v)) and an extracted ultrapure water blank (i.e. extraction performed with ultrapure water in place of sample) were systematically injected with each batch to respectively ensure lack of carryover in the UPLC system and monitor contamination linked to the sample preparation process. Composite QC samples were injected after the blanks to equilibrate the analytical system, and repeatedly throughout the batch (every 5 samples). Samples were injected randomly. IS peak areas were monitored to assess analytical drift.

2.12. Appendix S.4 – Sample preparation procedures

The twelve sample preparation methods used for this work are described below. As the spiking level, sample volume and recovery volume vary between experiments; they are not specified in each procedure and are recapitulated in Table B1.

Experiment	Spiking level (ng/mL)	Sample volume (µL)	Recovery volume (µL)
Preselection	40	200	100
Comparison to protein precipitation	10	100	20
Method detection limit	0.1, 0.5, 1, 5, 10, 20, 40	100	20

Table B1 – Spiking levels, sample volumes and recovery volumes used for all sample preparation procedures for three spiking experiments.

- **Protein precipitation**

Protein precipitation was carried out using a 4:1 (v/v) cold methanol to matrix ratio. Samples were left at -20°C for 1h to improve protein removal. Centrifugation was performed at 4°C and 17,000g for 20 min, after which supernatants were collected and evaporated to dryness under vacuum. Samples were recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- **Phospholipid and protein removal**
    - Ostro (Waters), Phree (Phenomenex)- Acetonitrile, PL (Supelco), PL Ultra (Supelco)

A 99:1 (v/v) acetonitrile to formic acid mixture was added to the matrix using a 3:1 (v/v) ratio. Samples were vortexed then placed on the plate and drawn through it drop by drop under vacuum. An additional volume of 100 µL of the 99:1 (v/v) acetonitrile to formic acid mixture was drawn through the plate for rinsing. The resulting solutions were evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- Phree (Phenomenex)- Methanol

A 99:1 (v/v) methanol to formic acid mixture was added to the matrix using a 4:1 (v/v) ratio. Samples were vortexed then placed on the plate and drawn through it drop by drop under vacuum. An additional volume of 100 µL of the 99:1 (v/v) methanol to formic acid mixture was drawn through the plate for rinsing. The resulting solutions were evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- PLD (Biotage)

A 99:1 (v/v) acetonitrile to formic acid mixture was added to the matrix using a 4:1 (v/v) ratio. Samples were vortexed then placed on the plate and drawn through it drop by drop under vacuum. An additional volume of 100 µL of the 99:1 (v/v) acetonitrile to formic acid mixture was drawn through the plate for rinsing. The resulting solutions were evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- Prime HLB (Waters)

Samples were placed on the plate and drawn though it drop by drop under vacuum. An additional volume of 2 mL of a 95:5 (v/v) ultrapure water to methanol mixture was drawn through the plate for rinsing. Elution was performed with 2 mL of a 90:10 (v/v) acetonitrile to methanol mixture. The resulting solutions were evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- **Solid phase extraction**
  - HLB Oasis, Strata X (Phenomenex)

A 98:2 (v/v) ultrapure water to formic acid mixture was added to the matrix using a 1:1 (v/v) ratio. Solid phase was conditioned with 1 mL of methanol followed by 1 mL of ultrapure water. Samples were placed on the plate and drawn through it drop by drop under vacuum. An additional volume of 2 mL of a 95:5 (v/v) ultrapure water to methanol mixture was drawn

through the plate for rinsing. After drying, elution was performed using 1 mL of methanol (first extract), then 1 mL of ethyl acetate (second extract). Extracts were separately evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- Strata XC (Phenomenex)

A 98:2 (v/v) ultrapure water to formic acid mixture was added to the matrix using a 1:1 (v/v) ratio. Solid phase was conditioned with 1 mL of methanol followed by 1 mL of ultrapure water. Samples were placed on the plate and drawn through it drop by drop under vacuum. An additional volume of 2 mL of a 95:5 (v/v) ultrapure water to methanol mixture was drawn through the plate for rinsing. After drying, elution was performed using 1 mL of a 95:5 (v/v) methanol to aqueous ammonia ratio (first extract), then 1 mL of methanol (second extract). Extracts were separately evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

- **Supported liquid extraction**

Samples were placed on the plate and drawn though it drop by drop under vacuum. Elution was performed with twice 900 µL of methyl tert-butyl ether (MTBE). The resulting solutions were evaporated to dryness under vacuum, and recovered in 90:10 (v/v) ultrapure water to acetonitrile mixture as to obtain the desired sample concentration factor.

2.13. Appendix S.5 – Application of PPT and Phree to cohort samples

Sample preparation methods PPT and Phree (acetonitrile) were applied to serum samples from the Pelagie cohort and plasma samples from a Danish birth cohort. Quality control was performed on the injected batches, both at the targeted and non-targeted scales. Results are presented in Figure S1.

Suspect screening was performed on the associated datasets, resulting in 44 xenobiotic annotations in serum and 41 xenobiotic annotations in plasma. For each annotated compound, fold changes (i.e. area ratio of features in samples prepared with Phree and protein precipitation) were computed for annotated compounds. Fold change values were also computed at the non-targeted level on quality control samples. Results are presented in Figure S2.

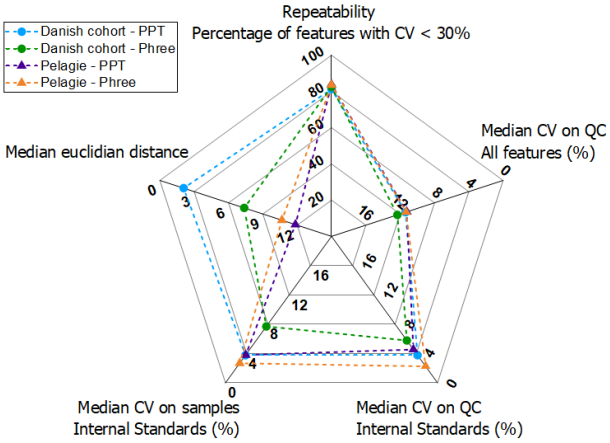

Figure S1 - Quality control parameters for the application of two sample preparation methods to two sets of cohort samples (n=8 plasma samples from the Danish cohort, and n=10 serum samples for Pelagie). Outer edges identify best performances.

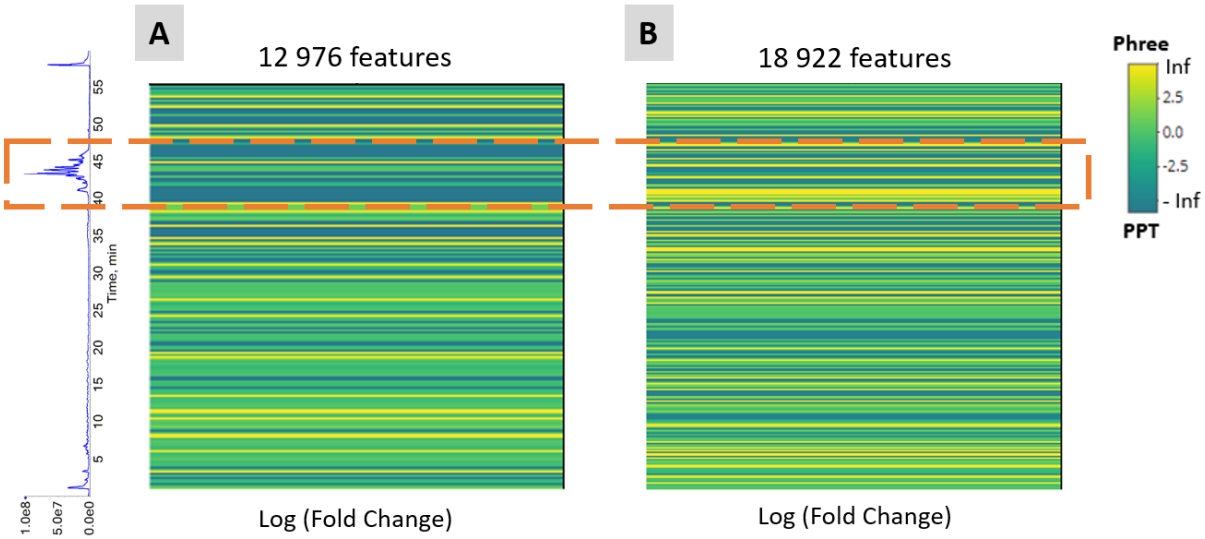

Figure S2 – Comparison of fold change values (i.e. area ratio of features in samples prepared with Phree and protein precipitation) for quality control samples in Pelagie serum samples (A) and Danish plasma samples (B). Yellow indicates features only visible in Phree-prepared samples and blue indicates features only visible in protein-precipitated samples. Features are organized by retention time value (from bottom to top). The orange dashed rectangle indicates the range where lysophospholipids and peptides are mostly observed.

3. Appendix 3. Supporting information – Chapter IV

3.1. Table A1 – Standard compounds form and suppliers

Table A1 – Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Arachidonic Acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	Bertin	Powder
Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	Bertin	Powder
Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N	Bertin	Powder
Prostaglandin D2	CCCCCC(C=CC1C(C(CC1=O)O)CC=CCCCC(=O)O)O	Bertin	Powder
Prostaglandin E2	CCCCCC(C=CC1C(CC(=O)C1CC=CCCCC(=O)O)O)O	Bertin	Powder
Prostaglandin F2a	CCCCCC(C=CC1C(CC(C1CC=CCCCC(=O)O)O)O)O	Bertin	Powder
Acetochlor	CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C	LGC	Powder
Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	LGC	Powder
Carbendazim	COC(=O)NC1=NC2=CC=CC=C2N1	LGC	Powder
Clothianidin	CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl	LGC	Powder
Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)CO)O)C	LGC	Powder
Dimethyldithiophosphate	COP(=S)(OC)S	LGC	Powder
Estrone	CC12CCC3C(C1CCC2=O)CCC4=C3C=CC(=C4)O	LGC	Powder
Fluoxetine	CNCCC(C1=CC=CC=C1)OC2=CC=C(C=C2)C(F)(F)F	LGC	1.0 mg/mL in MeOH
Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	LGC	Powder
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	LGC	Powder
Paracetamol	CC(=O)NC1=CC=C(C=C1)O	LGC	Powder
Paroxetine	C1CNCC(C1C2=CC=C(C=C2)F)COC3=CC4=C(C=C3)OCO4	LGC	1.0 mg/mL in MeOH
Progesterone	CC(=O)C1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	LGC	Powder
Sertraline	CNC1CCC(C2=CC=CC=C12)C3=CC(=C(C=C3)Cl)Cl	LGC	1.0 mg/mL in MeOH
Tebuconazole	CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O	LGC	Powder
Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	LGC	Powder
Thiacloprid	C1CSC(=NC#N)N1CC2=CN=C(C=C2)Cl	LGC	Powder
Venlafaxine	CN(C)CC(C1=CC=C(C=C1)OC)C2(CCCCC2)O	LGC	Powder
2-chloro-4-methylbenzoic acid	CC1=CC(=C(C=C1)C(=O)O)Cl	LGC	Powder
Acetamiprid	CC(=NC#N)N(C)CC1=CN=C(C=C1)Cl	LGC	Powder
Amidosulfuron	CN(S(=O)(=O)C)S(=O)(=O)NC(=O)NC1=NC(=CC(=N1)OC)OC	LGC	Powder
Atrazine	CCNC1=NC(=NC(=N1)Cl)NC(C)C	LGC	Powder
Atrazine-2-hydroxy	CCNC1=NC(=O)NC(=N1)NC(C)C	LGC	Powder
Beflubutamid	CCC(C(=O)NCC1=CC=CC=C1)OC2=CC(=C(C=C2)F)C(F)(F)F	LGC	Powder
Bixafen	CN1C=C(C(=N1)C(F)F)C(=O)NC2=C(C=C(C=C2)F)C3=CC(=C(C=C3)Cl)Cl	LGC	Powder
Bromacil	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	LGC	Powder

Table A1 – (continued) Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Carbaryl	CNC(=O)OC1=CC=CC2=CC=CC=C21	LGC	Powder
Carbetamide	CCNC(=O)C(C)OC(=O)NC1=CC=CC=C1	LGC	Powder
Chlorantraniliprole	CC1=CC(=CC(=C1NC(=O)C2=CC(=NN2C3=C(C=CC=N3)Cl)Br)C(=O)NC)Cl	LGC	Powder
Dimethenamid	CC1=CSC(=C1N(C(C)COC)C(=O)CCl)C	LGC	Powder
Estradiol-2-hydroxy	CC12CCC3C(C1CCC2O)CCC4=CC(=C(C=C34)O)O	LGC	Powder
Estrone-2-hydroxy	CC12CCC3C(C1CCC2=O)CCC4=CC(=C(C=C34)O)O	LGC	Powder
Ethidimuron	CCS(=O)(=O)C1=NN=C(S1)N(C)C(=O)NC	LGC	Powder
Fenamidone	CC1(C(=O)N(C(=N1)SC)NC2=CC=CC=C2)C3=CC=CC=C3	LGC	Powder
Fenpropimorph	CC1CN(CC(O1)C)CC(C)CC2=CC=C(C=C2)C(C)(C)C	LGC	Powder
Flonicamid	C1=CN=CC(=C1C(F)(F)F)C(=O)NCC#N	LGC	Powder
Fluroxypyr	C(C(=O)O)OC1=NC(=C(C(=C1Cl)N)Cl)F	LGC	Powder
Flurtamone	CNC1=C(C(=O)C(O1)C2=CC=CC=C2)C3=CC(=CC=C3)C(F)(F)F	LGC	Powder
Fosthiazate	CCO[P](=O)(SC(C)CC)N1CCSC1=O	LGC	Powder
Imazamethabenz-methyl	CC1=CC(=C(C=C1)C(=O)OC)C2=NC(C(=O)N2)(C)C(C)C	LGC	Powder
Imazamox	CC(C)C1(C(=O)NC(=N1)C2=C(C=C(C=N2)COC)C(=O)O)C	LGC	Powder
Imazaquin	CC(C)C1(C(=O)NC(=N1)C2=NC3=CC=CC=C3C=C2C(=O)O)C	LGC	Powder
Iodosulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=C(C=CC(=C2)I)C(=O)OC	LGC	Powder
Irgarol	CC(C)(C)NC1=NC(=NC(=N1)NC2CC2)SC	LGC	Powder
Isoxaben	CCC(C)(CC)C1=NOC(=C1)NC(=O)C2=C(C=CC=C2OC)OC	LGC	Powder
Isoxaflutole	CS(=O)(=O)C1=C(C=CC(=C1)C(F)(F)F)C(=O)C2=C(ON=C2)C3CC3	LGC	Powder
Metamitron	CC1=NN=C(C(=O)N1N)C2=CC=CC=C2	LGC	Powder
Metobromuron	CN(C(=O)NC1=CC=C(C=C1)Br)OC	LGC	Powder
Metolachlor	CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C	LGC	Powder
Metosulam	CC1=C(C(=C(C=C1)Cl)NS(=O)(=O)C2=NN3C(=CC(=NC3=N2)OC)OC)Cl	LGC	Powder
Metribuzine	CSC1=NN=C(C(=O)N1N)C(C)(C)C	LGC	Powder
Metsulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(=O)OC	LGC	Powder
Nicosulfuron	CN(C)C(=O)C1=C(N=CC=C1)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC	LGC	Powder
Oryzalin	CCCN(CCC)C1=C(C=C(C=C1[N+](=O)[O-])S(=O)(=O)N)[N+](=O)[O-]	LGC	Powder
Pencycuron	C1CCC(C1)N(CC2=CC=C(C=C2)Cl)C(=O)NC3=CC=CC=C3	LGC	Powder
Propachlor	CC(C)N(C1=CC=CC=C1)C(=O)CCl	LGC	Powder
Propamocarb	CCCOC(=O)NCCCN(C)C	LGC	Powder
Propoxycarbazone	CCCOC1=NN(C(=O)N1C)C(=O)NS(=O)(=O)C2=CC=CC=C2C(=O)OC	LGC	Powder
Pymetrozine	CC1=NNC(=O)N(C1)N=CC2=CN=CC=C2	LGC	Powder
Pyraclostrobin	COC(=O)N(C1=CC=CC=C1COC2=NN(C=C2)C3=CC=C(C=C3)Cl)OC	LGC	Powder
Pyroxsulam	COC1=CC(=NC2=NC(=NN12)NS(=O)(=O)C3=C(C=CN=C3OC)C(F)(F)F)OC	LGC	Powder
Quinmerac	CC1=CC2=C(C(=C(C=C2)Cl)C(=O)O)N=C1	LGC	Powder
Spiroxamine	CCCN(CC)CC1COC2(CCC(CC2)C(C)(C)C)O1	LGC	Powder
Sulcotrione	CS(=O)(=O)C1=CC(=C(C=C1)C(=O)C2C(=O)CCCC2=O)Cl	LGC	Powder

Table A1 – (continued) Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Terbuthylazine	CCNC1=NC(=NC(=N1)Cl)NC(C)(C)C	LGC	Powder
Tertbutylazine-2-hydroxy	CCNC1=NC(=O)NC(=N1)NC(C)(C)C	LGC	Powder
Thifensulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=C(SC=C2)C(=O)OC	LGC	Powder
Triazoxide	C1=CC2=C(C=C1Cl)[N+](=NC(=N2)N3C=CN=C3)[O-]	LGC	Powder
Triclopyr	C1=C(C(=NC(=C1Cl)Cl)OCC(=O)O)Cl	LGC	Powder
Triflusulfuron-methyl	CC1=C(C(=CC=C1)C(=O)OC)S(=O)(=O)NC(=O)NC2=NC(=NC(=N2)OCC(F)(F)F)N(C)C	LGC	Powder
Trinexapac-ethyl	CCOC(=O)C1CC(=O)C(=C(C2CC2)O)C(=O)C1	LGC	Powder
Triticonazole	CC1(CCC(=CC2=CC=C(C=C2)Cl)C1(CN3C=NC=N3)O)C	LGC	Powder
Tritosulfuron	COC1=NC(=NC(=N1)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(F)(F)F)C(F)(F)F	LGC	Powder
17b-Estradiol	CC12CCC3C(C1CCC2O)CCC4=C3C=CC(=C4)O	LGC	Powder
Acetylsalicylic acid	CC(=O)OC1=CC=CC=C1C(=O)O	LGC	Powder
Aniline	C1=CC=C(C=C1)N	LGC	Powder
Dehydroepiandrosterone	CC12CCC3C(C1CCC2=O)CC=C4C3(CCC(C4)O)C	LGC	Powder
Estriol	CC12CCC3C(C1CC(C2O)O)CCC4=C3C=CC(=C4)O	LGC	Powder
L-thyroxine	C1=C(C=C(C(=C1I)OC2=CC(=C(C(=C2)I)O)I)I)CC(C(=O)O)N	LGC	Powder
Pregnenolone	CC(=O)C1CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C	LGC	Powder
Progesterone-17-hydroxy	CC(=O)C1(CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C)O	LGC	Powder
Tryptophan	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N	LGC	Powder
Mesotrione	CS(=O)(=O)C1=CC(=C(C=C1)C(=O)C2C(=O)CCCC2=O)[N+](=O)[O-]	Servilab	Powder
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	Servilab	Powder
Aflatoxin B1	COC1=C2C3=C(C(=O)CC3)C(=O)OC2=C4C5C=COC5OC4=C1	Sigma Aldrich	Powder
Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	Sigma Aldrich	Powder
Hydroxyindoleacetic acid	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	Sigma Aldrich	Powder
Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	Sigma Aldrich	Powder
Pravastatin	CCC(C)C(=O)OC1CC(C=C2C1C(C(C=C2)C)CCC(CC(CC(=O)O)O)O)O	Sigma Aldrich	Powder
Solanidine	CC1CCC2C(C3C(N2C1)CC4C3(CCC5C4CC=C6C5(CCC(C6)O)C)C)C	Sigma Aldrich	Powder
Foramsulfuron	CN(C)C(=O)C1=C(C=C(C=C1)NC=O)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC	Sigma Aldrich	Powder
4-Aminophenol	C1=CC(=CC=C1N)O	Sigma Aldrich	Powder
Acetylcholine	CC(=O)OCC[N+](C)(C)C	Sigma Aldrich	Powder
Aldosterone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4C(=O)CO)C=O)O	Sigma Aldrich	Powder
Allopregnanolone	CC(=O)C1CCC2C1(CCC3C2CCC4C3(CCC(C4)O)C)C	Sigma Aldrich	Powder
Amoxicillin	CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=C(C=C3)O)N)C(=O)O)C	Sigma Aldrich	Powder
Dopamine	C1=CC(=C(C=C1CCN)O)O	Sigma Aldrich	Powder
Epinephrine	CNCC(C1=CC(=C(C=C1)O)O)O	Sigma Aldrich	Powder
Ethinylestradiol	CC12CCC3C(C1CCC2(C#C)O)CCC4=C3C=CC(=C4)O	Sigma Aldrich	Powder
Ketoprofen	CC(C1=CC(=CC=C1)C(=O)C2=CC=CC=C2)C(=O)O	Sigma Aldrich	Powder
Methylparaben	COC(=O)C1=CC=C(C=C1)O	Sigma Aldrich	Powder
Morphine	CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O	Sigma Aldrich	Powder

Table A1 – (continued) Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Oxazepam	C1=CC=C(C=C1)C2=NC(C(=O)NC3=C2C=C(C=C3)Cl)O	Sigma Aldrich	Powder
Oxybenzone	COC1=CC(=C(C=C1)C(=O)C2=CC=CC=C2)O	Sigma Aldrich	Powder
Pivmecillinam	CC1(C(N2C(S1)C(C2=O)N=CN3CCCCCC3)C(=O)OCOC(=O)C(C)(C)C)C	Sigma Aldrich	Powder
Propylparaben	CCCOC(=O)C1=CC=C(C=C1)O	Sigma Aldrich	Powder
Salicylic acid	C1=CC=C(C(=C1)C(=O)O)O	Sigma Aldrich	Powder
Tryptamine-5-hydroxy	C1=CC2=C(C=C1O)C(=CN2)CCN	Sigma Aldrich	Powder
2-Phenylphenol	C1=CC=C(C=C1)C2=CC=CC=C2O	VWR	Powder
Aminobenzimidazole	C1=CC=C2C(=C1)NC(=N2)N	VWR	Powder
Azoxystrobin	COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC	VWR	Powder
Boscalid	C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl	VWR	Powder
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	VWR	Powder
Chlorpyrifos	CCOP(=S)(OCC)OC1=NC(=C(C=C1Cl)Cl)Cl	VWR	Powder
Cotinine	CN1C(CCC1=O)C2=CN=CC=C2	VWR	Powder
Cyprodinil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C3CC3	VWR	Powder
Diazinon	CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C	VWR	Powder
Diclofenac	C1=CC=C(C(=C1)CC(=O)O)NC2=C(C=CC=C2Cl)Cl	VWR	Powder
Imidacloprid	C1CN(C(=N[N+](=O)[O-])N1)CC2=CN=C(C=C2)Cl	VWR	Powder
Nicotine	CN1CCCC1C2=CN=CC=C2	VWR	Powder
Prochloraz	CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=CN=C2	VWR	Powder
Propiconazole	CCCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl	VWR	Powder
Thiamethoxam	CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl	VWR	Powder
1-(3,4-Dichlorophenyl)-3-methylurea	CNC(=O)NC1=CC(=C(C=C1)Cl)Cl	VWR	Powder
1-(3,4-Dichlorophenyl)urea	C1=CC(=C(C=C1NC(=O)N)Cl)Cl	VWR	Powder
1-(4-Isopropylphenyl)urea	CC(C)C1=CC=C(C=C1)NC(=O)N	VWR	Powder
2,4-mcpa	CC1=C(C=CC(=C1)Cl)OCC(=O)O	VWR	Powder
Alachlor	CCC1=C(C(=CC=C1)CC)N(COC)C(=O)CCl	VWR	Powder
Ametryn	CCNC1=NC(=NC(=N1)SC)NC(C)C	VWR	Powder
Atrazine-deisopropyl	CCNC1=NC(=NC(=N1)N)Cl	VWR	Powder
Carbofuran	CC1(CC2=C(O1)C(=CC=C2)OC(=O)NC)C	VWR	Powder
Chloridazon	C1=CC=C(C=C1)N2C(=O)C(=C(C=N2)N)Cl	VWR	Powder
Chlortoluron	CC1=C(C=C(C=C1)NC(=O)N(C)C)Cl	VWR	Powder
Dichlorprop	CC(C(=O)O)OC1=C(C=C(C=C1)Cl)Cl	VWR	Powder
Dimethomorph	COC1=C(C=C(C=C1)C(=CC(=O)N2CCOCC2)C3=CC=C(C=C3)Cl)OC	VWR	Powder
Diuron	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl	VWR	Powder
Fenpropidine	CC(CC1=CC=C(C=C1)C(C)(C)C)CN2CCCCC2	VWR	Powder
Flufenacet	CC(C)N(C1=CC=C(C=C1)F)C(=O)COC2=NN=C(S2)C(F)(F)F	VWR	Powder
Iprodione	CC(C)NC(=O)N1CC(=O)N(C1=O)C2=CC(=CC(=C2)Cl)Cl	VWR	Powder

Table A1 – (continued) Standard compounds form and suppliers

Compound name	SMILES	Supplier	Form
Isoproturon	CC(C)C1=CC=C(C=C1)NC(=O)N(C)C	VWR	Powder
Isoproturon-didemethyl	CC(C)C1=CC=C(C=C1)NC(=O)N	VWR	Powder
Linuron	CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC	VWR	Powder
Mesosulfuron-methyl	COC1=CC(=NC(=N1)NC(=O)NS(=O)(=O)C2=C(C=CC(=C2)CNS(=O)(=O)C)C(=O)OC)OC	VWR	Powder
Metalaxyl	CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)COC	VWR	Powder
Metazachlor	CC1=C(C(=CC=C1)C)N(CN2C=CC=N2)C(=O)CCl	VWR	Powder
Methabenzthiazuron	CNC(=O)N(C)C1=NC2=CC=CC=C2S1	VWR	Powder
Paclobutrazol	CC(C)(C)C(C(CC1=CC=C(C=C1)Cl)N2C=NC=N2)O	VWR	Powder
Pirimicarb	CC1=C(N=C(N=C1OC(=O)N(C)C)N(C)C)C	VWR	Powder
Propyzamide	CC(C)(C#C)NC(=O)C1=CC(=CC(=C1)Cl)Cl	VWR	Powder
Prosulfuron	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2CCC(F)(F)F	VWR	Powder
Pyrimethanil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C	VWR	Powder
Simazine	CCNC1=NC(=NC(=N1)Cl)NCC	VWR	Powder
Tebutame	CC(C)N(CC1=CC=CC=C1)C(=O)C(C)(C)C	VWR	Powder
Terbutryne	CCNC1=NC(=NC(=N1)SC)NC(C)(C)C	VWR	Powder
Triadimenol	CC(C)(C)C(C(N1C=NC=N1)OC2=CC=C(C=C2)Cl)O	VWR	Powder
Chlorpyrifos-methyl	COP(=S)(OC)OC1=NC(=C(C=C1Cl)Cl)Cl	VWR	Powder
Malathion	CCOC(=O)CC(C(=O)OCC)SP(=S)(OC)OC	VWR	Powder
Triclosan	C1=CC(=C(C=C1Cl)O)OC2=C(C=C(C=C2)Cl)Cl	VWR	Powder

3.2. Table A2 – Computer specifications

Model	Dell OptiPlex XE2
Processor	Intel® Core™ i5-4570S CPU @ 2.90 GHz 2.89 GHz
RAM	32.0 GB
System type	64-bit operating system, x64-based processor
Operating system	Windows 10 Enterprise 2016 LTSB

3.3. Table A3 – Sets of compounds used for spiking samples (n=45), training models (n=134) and evaluating them (n=30)

Table A3 – Sets of compounds used for spiking samples (n=45), training models (n=134) and evaluating them (n=30)

	Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Spiking set	2-Phenylphenol	C1=CC=C(C=C1)C2=CC=CC=C2O	C12H10O	170.0732	90-43-7
	Acetochlor	CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C	C14H20ClNO2	269.1183	123113-74-6
	Aflatoxin B1	COC1=C2C3=C(C(=O)CC3)C(=O)OC2=C4C5C=COC5OC4=C1	C17H12O6	312.0634	27261-02-5
	Aminobenzimidazole	C1=CC=C2C(=C1)NC(=N2)N	C7H7N3	133.0640	934-32-7
	Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	C19H26O2	286.1933	63-05-8
	Arachidonic Acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	C20H32O2	304.2402	93444-49-6
	Azoxystrobin	COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC	C22H17N3O5	403.1168	215934-32-0
	Boscalid	C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl	C18H12Cl2N2O	342.0327	188425-85-6
	Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	C15H12N2O	236.0950	298-46-4
	Carbendazim	COC(=O)NC1=NC2=CC=CC=C2N1	C9H9N3O2	191.0695	63278-70-6
	Chlorpyrifos	CCOP(=S)(OCC)OC1=NC(=C(C=C1Cl)Cl)Cl	C9H11Cl3NO3PS	348.9263	39475-55-3
	Clothianidin	CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl	C6H8ClN5O2S	249.0087	205510-53-8
	Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	C18H21NO3	299.1521	76-57-3
	Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)CO)O)C	C21H28O5	360.1937	53-06-5
	Cotinine	CN1C(CCC1=O)C2=CN=CC=C2	C10H12N2O	176.0950	486-56-6
	Cyprodinil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C3CC3	C14H15N3	225.1266	121552-61-2
	Diazinon	CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C	C12H21N2O3PS	304.1011	30583-38-1
	Diclofenac	C1=CC=C(C(=C1)CC(=O)O)NC2=C(C=CC=C2Cl)Cl	C14H11Cl2NO2	295.0167	15307-86-5
	Dimethyldithiophosphate	COP(=S)(OC)S	C2H7O2PS2	157.9625	756-80-9
	Estrone	CC12CCC3C(C1CCC2=O)CCC4=C3C=CC(=C4)O	C18H22O2	270.1620	53-16-7
	Fluoxetine	CNCCC(C1=CC=CC=C1)OC2=CC=C(C=C2)C(F)(F)F	C17H18F3NO	309.1340	57226-07-0
	Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	C21H30O5	362.2093	50-23-7
	Hydroxyindoleacetic acid	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	C10H9NO3	191.0582	113303-91-6
	Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	C13H18O2	206.1307	58560-75-1
	Imidacloprid	C1CN(C(=N[N+](=O)[O-])N1)CC2=CN=C(C=C2)Cl	C9H10ClN5O2	255.0523	138261-41-3
	Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	C20H32O4	336.2301	71160-24-2
	Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N	C25H40N2O6S	496.2607	73836-78-9
	Nicotine	CN1CCCC1C2=CN=CC=C2	C10H14N2	162.1157	551-13-3
	Paracetamol	CC(=O)NC1=CC=C(C=C1)O	C8H9NO2	151.0633	8055-08-1
	Paroxetine	C1CNCC(C1C2=CC=C(C=C2)F)COC3=CC4=C(C=C3)OCO4	C19H20FNO3	329.1427	63952-24-9
	Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	C17H19NO3	285.1365	147030-08-8
	Pravastatin	CCC(C)C(=O)OC1CC(C=C2C1C(C(C=C2)C)CCC(CC(CC(=O)O)O)O)O	C23H36O7	424.2461	81093-37-0
	Prochloraz	CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=CN=C2	C15H16Cl3N3O2	375.0308	67747-09-5
	Progesterone	CC(=O)C1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	C21H30O2	314.2246	257630-50-5
	Propiconazole	CCCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl	C15H17Cl2N3O2	341.0698	75881-82-2
	Prostaglandin D2	CCCCCC(C=CC1C(C(CC1=O)O)CC=CCCCC(=O)O)O	C20H32O5	352.2250	41598-07-6
	Prostaglandin E2	CCCCCC(C=CC1C(CC(=O)C1CC=CCCCC(=O)O)O)O	C20H32O5	352.2250	363-24-6

Table A3 – Sets of compounds used for spiking samples (n=45), training models (n=134) and evaluating them (n=30)

	Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
	Prostaglandin F2a	CCCCCC(C=CC1C(CC(C1CC=CCCCC(=O)O)O)O)O	C20H34O5	354.2406	13535-33-6
	Sertraline	CNC1CCC(C2=CC=CC=C12)C3=CC(=C(C=C3)Cl)Cl	C17H17Cl2N	305.0738	79559-97-0
	Solanidine	CC1CCC2C(C3C(N2C1)CC4C3(CCC5C4CC=C6C5(CCC(C6)O)C)C)C	C27H43NO	397.3345	80-78-4
Spiking	Tebuconazole	CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O	C16H22ClN3O	307.1451	80443-41-0
set	Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	C19H28O2	288.2089	58-22-0
	Thiacloprid	C1CSC(=NC#N)N1CC2=CN=C(C=C2)Cl	C10H9ClN4S	252.0236	111988-49-9
	Thiamethoxam	CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl	C8H10ClN5O3S	291.0193	153719-23-4
	Venlafaxine	CN(C)CC(C1=CC=C(C=C1)OC)C2(CCCCC2)O	C17H27NO2	277.2042	93413-69-5
	1-(3,4-Dichlorophenyl)-3-methylurea	CNC(=O)NC1=CC(=C(C=C1)Cl)Cl	C8H8Cl2N2O	218.0014	3567-62-2
	1-(3,4-Dichlorophenyl)urea	C1=CC(=C(C=C1NC(=O)N)Cl)Cl	C7H6Cl2N2O	203.9857	2327-02-8
	1-(4-Isopropylphenyl)urea	CC(C)C1=CC=C(C=C1)NC(=O)N	C10H14N2O	178.1106	56046-17-4
	2,4-mcpa	CC1=C(C=CC(=C1)Cl)OCC(=O)O	C9H9ClO3	200.0240	94-74-6
	2-chloro-4-methylbenzoic acid	CC1=CC(=C(C=C1)C(=O)O)Cl	C8H7ClO2	170.0135	7697-25-8
	Acetamiprid	CC(=NC#N)N(C)CC1=CN=C(C=C1)Cl	C10H11ClN4	222.0672	135410-20-7
	Alachlor	CCC1=C(C(=CC=C1)CC)N(COC)C(=O)CCl	C14H20ClNO2	269.1183	15972-60-8
	Ametryn	CCNC1=NC(=NC(=N1)SC)NC(C)C	C9H17N5S	227.1205	834-12-8
	Amidosulfuron	CN(S(=O)(=O)C)S(=O)(=O)NC(=O)NC1=NC(=CC(=N1)OC)OC	C9H15N5O7S2	369.0412	120923-37-7
	Atrazine	CCNC1=NC(=NC(=N1)Cl)NC(C)C	C8H14ClN5	215.0938	1912-24-9
	Atrazine-2-hydroxy	CCNC1=NC(=O)NC(=N1)NC(C)C	C8H15N5O	197.1277	2163-68-0
	Atrazine-deisopropyl	CCNC1=NC(=NC(=N1)N)Cl	C5H8ClN5	173.0468	1007-28-9
	Beflubutamid	CCC(C(=O)NCC1=CC=CC=C1)OC2=CC(=C(C=C2)F)C(F)(F)F	C18H17F4NO2	355.1195	113614-08-7
	Bixafen	CN1C=C(C(=N1)C(F)F)C(=O)NC2=C(C=C(C=C2)F)C3=CC(=C(C=C3)Cl)Cl	C18H12Cl2F3N3O	413.0310	581809-46-3
	Bromacil	CCC(C)N1C(=O)NC(=C(Br)C1=O)C	C9H13BrN2O2	260.0160	314-40-9
	Carbaryl	CNC(=O)OC1=CC=CC2=CC=CC=C21	C12H11NO2	201.0790	51274-03-4
Training	Carbetamide	CCNC(=O)C(C)OC(=O)NC1=CC=CC=C1	C12H16N2O3	236.1161	16118-49-3
Set	Carbofuran	CC1(CC2=C(O1)C(=CC=C2)OC(=O)NC)C	C12H15NO3	221.1052	1563-66-2
	Chlorantraniliprole	CC1=CC(=CC(=C1NC(=O)C2=CC(=NN2C3=C(C=CC=N3)Cl)Br)C(=O)NC)Cl	C18H14BrCl2N5O2	480.9708	500008-45-7
	Chloridazon	C1=CC=C(C=C1)N2C(=O)C(=C(C=N2)N)Cl	C10H8ClN3O	221.0356	1698-60-8
	Chlortoluron	CC1=C(C=C(C=C1)NC(=O)N(C)C)Cl	C10H13ClN2O	212.0716	15545-48-9
	Dichlorprop	CC(C(=O)O)OC1=C(C=C(C=C1)Cl)Cl	C9H8Cl2O3	233.9851	120-36-5
	Dimethenamid	CC1=CSC(=C1N(C(C)COC)C(=O)CCl)C	C12H18ClNO2S	275.0747	87674-68-8
	Dimethomorph	COC1=C(C=C(C=C1)C(=CC(=O)N2CCOCC2)C3=CC=C(C=C3)Cl)OC	C21H22ClNO4	387.1237	110488-70-5
	Diuron	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl	C9H10Cl2N2O	232.0170	102962-29-8
	Estradiol-2-hydroxy	CC12CCC3C(C1CCC2O)CCC4=CC(=C(C=C34)O)O	C18H24O3	288.1725	362-05-0
	Estrone-2-hydroxy	CC12CCC3C(C1CCC2=O)CCC4=CC(=C(C=C34)O)O	C18H22O3	286.1569	362-06-1
	Ethidimuron	CCS(=O)(=O)C1=NN=C(S1)N(C)C(=O)NC	C7H12N4O3S2	264.0351	30043-49-3
	Fenamidone	CC1(C(=O)N(C(=N1)SC)NC2=CC=CC=C2)C3=CC=CC=C3	C17H17N3OS	311.1092	161326-34-7
	Fenpropidine	CC(CC1=CC=C(C=C1)C(C)(C)C)CN2CCCCC2	C19H31N	273.2456	67306-00-7
	Fenpropimorph	CC1CN(CC(O1)C)CC(C)CC2=CC=C(C=C2)C(C)(C)C	C20H33NO	273.2456	67564-91-4
	Flonicamid	C1=CN=CC(=C1C(F)(F)F)C(=O)NCC#N	C9H6F3N3O	229.0463	158062-67-0
	Flufenacet	CC(C)N(C1=CC=C(C=C1)F)C(=O)COC2=NN=C(S2)C(F)(F)F	C14H13F4N3O2S	363.0665	142459-58-3
	Fluroxypyr	C(C(=O)O)OC1=NC(=C(C(=C1Cl)N)Cl)F	C7H5Cl2FN2O3	253.9661	69377-81-7

Table A3 – Sets of compounds used for spiking samples (n=45), training models (n=134) and evaluating them (n=30)

	Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Training Set	Flurtamone	CNC1=C(C(=O)C(O1)C2=CC=CC=C2)C3=CC(=CC=C3)C(F)(F)F	C18H14F3NO2	333.0977	96525-23-4
	Foramsulfuron	CN(C)C(=O)C1=C(C=C(C=C1)NC=O)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC	C17H20N6O7S	452.1114	173159-57-4
	Fosthiazate	CCO[P](=O)(SC(C)CC)N1CCSC1=O	C9H18NO3PS2	283.0466	98886-44-3
	Imazamethabenz-methyl	CC1=CC(=C(C=C1)C(=O)OC)C2=NC(C(=O)N2)(C)C(C)C	C16H20N2O3	288.1474	81405-85-8
	Imazamox	CC(C)C1(C(=O)NC(=N1)C2=C(C=C(C=N2)COC)C(=O)O)C	C15H19N3O4	305.1376	114311-32-9
	Imazaquin	CC(C)C1(C(=O)NC(=N1)C2=NC3=CC=CC=C3C=C2C(=O)O)C	C17H17N3O3	311.1270	81335-37-7
	Iodosulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=C(C=CC(=C2)I)C(=O)OC	C14H13IN5NaO6S	528.9529	144550-36-7
	Iprodione	CC(C)NC(=O)N1CC(=O)N(C1=O)C2=CC(=CC(=C2)Cl)Cl	C13H13Cl2N3O3	329.0334	36734-19-7
	Irgarol	CC(C)(C)NC1=NC(=NC(=N1)NC2CC2)SC	C11H19N5S	253.1361	28159-98-0
	Isoproturon	CC(C)C1=CC=C(C=C1)NC(=O)N(C)C	C12H18N2O	206.1419	34123-59-6
	Isoproturon-didemethyl	CC(C)C1=CC=C(C=C1)NC(=O)N	C10H14N2O	178.1106	56046-17-4
	Isoxaben	CCC(C)(CC)C1=NOC(=C1)NC(=O)C2=C(C=CC=C2OC)OC	C18H24N2O4	332.1736	82558-50-7
	Isoxaflutole	CS(=O)(=O)C1=C(C=CC(=C1)C(F)(F)F)C(=O)C2=C(ON=C2)C3CC3	C15H12F3NO4S	359.0439	141112-29-0
	Linuron	CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC	C9H10Cl2N2O2	248.0119	56645-87-5
	Mesosulfuron-methyl	COC1=CC(=NC(=N1)NC(=O)NS(=O)(=O)C2=C(C=CC(=C2)CNS(=O)(=O)C)C(=O)OC)OC	C17H21N5O9S2	503.0781	208465-21-8
	Mesotrione	CS(=O)(=O)C1=CC(=C(C=C1)C(=O)C2C(=O)CCCC2=O)[N+](=O)[O-]	C14H13NO7S	339.0413	104206-82-8
	Metalaxyl	CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)COC	C15H21NO4	279.1471	57837-19-1
	Metamitron	CC1=NN=C(C(=O)N1N)C2=CC=CC=C2	C10H10N4O	202.0855	41394-05-2
	Metazachlor	CC1=C(C(=CC=C1)C)N(CN2C=CC=N2)C(=O)CCl	C14H16ClN3O	277.0982	67129-08-2
	Methabenzthiazuron	CNC(=O)N(C)C1=NC2=CC=CC=C2S1	C10H11N3OS	221.0623	18691-97-9
	Metobromuron	CN(C(=O)NC1=CC=C(C=C1)Br)OC	C9H11BrN2O2	258.0004	3060-89-7
	Metolachlor	CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C	C15H22ClNO2	283.1339	55762-76-0
	Metosulam	CC1=C(C(=C(C=C1)Cl)NS(=O)(=O)C2=NN3C(=CC(=NC3=N2)OC)OC)Cl	C14H13Cl2N5O4S	417.0065	139528-85-1
	Metribuzine	CSC1=NN=C(C(=O)N1N)C(C)(C)C	C8H14N4OS	214.0888	21087-64-9
	Metsulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(=O)OC	C14H15N5O6S	381.0743	74223-64-6
	Nicosulfuron	CN(C)C(=O)C1=C(N=CC=C1)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC	C15H18N6O6S	410.1009	111991-09-4
	Oryzalin	CCCN(CCC)C1=C(C=C(C=C1[N+](=O)[O-])S(=O)(=O)N)[N+](=O)[O-]	C12H18N4O6S	346.0947	19044-88-3
	Paclobutrazol	CC(C)(C)C(C(CC1=CC=C(C=C1)Cl)N2C=NC=N2)O	C30H40Cl2N6O2	586.2590	76738-62-0
	Pencycuron	C1CCC(C1)N(CC2=CC=C(C=C2)Cl)C(=O)NC3=CC=CC=C3	C19H21ClN2O	328.1342	66063-05-6
	Pirimicarb	CC1=C(N=C(N=C1OC(=O)N(C)C)N(C)C)C	C11H18N4O2	238.1430	23103-98-2
	Propachlor	CC(C)N(C1=CC=CC=C1)C(=O)CCl	C11H14ClNO	211.0764	1918-16-7
	Propamocarb	CCCOC(=O)NCCCN(C)C	C9H20N2O2	188.1525	24579-73-5
	Propoxycarbazone	CCCOC1=NN(C(=O)N1C)C(=O)NS(=O)(=O)C2=CC=CC=C2C(=O)OC	C15H17N4NaO7S	420.0716	181274-15-7
	Propyzamide	CC(C)(C#C)NC(=O)C1=CC(=CC(=C1)Cl)Cl	C12H11Cl2NO	255.0218	11097-11-3
	Prosulfuron	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2CCC(F)(F)F	C15H16F3N5O4S	419.0875	94125-34-5
	Pymetrozine	CC1=NNC(=O)N(C1)N=CC2=CN=CC=C2	C10H11N5O	217.0964	123312-89-0
	Pyraclostrobin	COC(=O)N(C1=CC=CC=C1COC2=NN(C=C2)C3=CC=C(C=C3)Cl)OC	C19H18ClN3O4	387.0986	175013-18-0
	Pyrimethanil	CC1=CC(=NC(=N1)NC2=CC=CC=C2)C	C12H13N3	199.1109	53112-28-0
	Pyroxsulam	COC1=CC(=NC2=NC(=NN12)NS(=O)(=O)C3=C(C=CN=C3OC)C(F)(F)F)OC	C14H13F3N6O5S	434.0620	422556-08-9
	Quinmerac	CC1=CC2=C(C(=C(C=C2)Cl)C(=O)O)N=C1	C11H8ClNO2	221.0244	90717-03-6
	Simazine	CCNC1=NC(=NC(=N1)Cl)NCC	C7H12ClN5	201.0781	119603-94-0
	Spiroxamine	CCCN(CC)CC1COC2(CCC(CC2)C(C)(C)C)O1	C18H35NO2	297.2668	118134-30-8
	Sulcotrione	CS(=O)(=O)C1=CC(=C(C=C1)C(=O)C2C(=O)CCCC2=O)Cl	C14H13ClO5S	328.0172	99105-77-8

Table A3 – Sets of compounds used for spiking samples (n=45), training models (n=134) and evaluating them (n=30)

	Compound name	SMILES	Chemical formula	Monoisotopic mass	CAS
Training Set	Tebutame	CC(C)N(CC1=CC=CC=C1)C(=O)C(C)(C)C	C15H23NO	233.1780	35256-85-0
	Terbuthylazine	CCNC1=NC(=NC(=N1)Cl)NC(C)(C)C	C9H16ClN5	229.1094	5915-41-3
	Terbutryne	CCNC1=NC(=NC(=N1)SC)NC(C)(C)C	C10H19N5S	241.1361	886-50-0
	Tertbutylazine-2-hydroxy	CCNC1=NC(=O)NC(=N1)NC(C)(C)C	C9H17N5O	211.1433	66753-07-9
	Thifensulfuron-methyl	CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=C(SC=C2)C(=O)OC	C12H13N5O6S2	387.0307	79277-27-3
	Triadimenol	CC(C)(C)C(C(N1C=NC=N1)OC2=CC=C(C=C2)Cl)O	C14H18ClN3O2	295.1088	55219-65-3
	Triazoxide	C1=CC2=C(C=C1Cl)[N+](=NC(=N2)N3C=CN=C3)[O-]	C10H6ClN5O	247.0261	72459-58-6
	Triclopyr	C1=C(C(=NC(=C1Cl)Cl)OCC(=O)O)Cl	C7H4Cl3NO3	254.9257	55335-06-3
	Triflusulfuron-methyl	CC1=C(C(=CC=C1)C(=O)OC)S(=O)(=O)NC(=O)NC2=NC(=NC(=N2)OCC(F)(F)F)N(C)C	C17H19F3N6O6S	492.1039	126535-15-7
	Trinexapac-ethyl	CCOC(=O)C1CC(=O)C(=C(C2CC2)O)C(=O)C1	C13H16O5	252.0998	95266-40-3
	Triticonazole	CC1(CCC(=CC2=CC=C(C=C2)Cl)C1(CN3C=NC=N3)O)C	C17H20ClN3O	317.1295	131983-72-7
	Tritosulfuron	COC1=NC(=NC(=N1)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(F)(F)F)C(F)(F)F	C13H9F6N5O4S	445.0279	142469-14-5
Validation set	17b-Estradiol	CC12CCC3C(C1CCC2O)CCC4=C3C=CC(=C4)O	C18H24O2	272.1776	50-28-2
	4-Aminophenol	C1=CC(=CC=C1N)O	C6H7NO	109.0528	123-30-8
	Acetylcholine	CC(=O)OCC[N+](C)(C)C	C7H16NO2	146.1181	51-84-3
	Acetylsalicylic acid	CC(=O)OC1=CC=CC=C1C(=O)O	C9H8O4	180.0423	50-78-2
	Aldosterone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4C(=O)CO)C=O)O	C21H28O5	360.1937	152-04-5
	Allopregnanolone	CC(=O)C1CCC2C1(CCC3C2CCC4C3(CCC(C4)O)C)C	C21H34O2	318.2559	516-54-1
	Amoxicillin	CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=C(C=C3)O)N)C(=O)O)C	C16H19N3O5S	365.1045	26787-78-0
	Aniline	C1=CC=C(C=C1)N	C6H7N	93.0578	62-53-3
	Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	C8H10N4O2	194.0804	58-08-2
	Chlorpyrifos-methyl	COP(=S)(OC)OC1=NC(=C(C=C1Cl)Cl)Cl	C7H7Cl3NO3PS	320.8950	5598-13-0
	Dehydroepiandrosterone	CC12CCC3C(C1CCC2=O)CC=C4C3(CCC(C4)O)C	C19H28O2	288.2089	53-43-0
	Dopamine	C1=CC(=C(C=C1CCN)O)O	C8H11NO2	153.0790	51-61-6
	Epinephrine	CNCC(C1=CC(=C(C=C1)O)O)O	C9H13NO3	183.0895	51-43-4
	Estriol	CC12CCC3C(C1CC(C2O)O)CCC4=C3C=CC(=C4)O	C18H24O3	288.1725	50-27-1
	Ethinylestradiol	CC12CCC3C(C1CCC2(C#C)O)CCC4=C3C=CC(=C4)O	C20H24O2	296.1776	77538-56-8
	Ketoprofen	CC(C1=CC(=CC=C1)C(=O)C2=CC=CC=C2)C(=O)O	C16H14O3	254.0943	172964-50-0
	L-thyroxine	C1=C(C=C(C(=C1I)OC2=CC(=C(C(=C2)I)O)I)I)CC(C(=O)O)N	C15H11I4NO4	776.6867	7488-70-2
	Malathion	CCOC(=O)CC(C(=O)OCC)SP(=S)(OC)OC	C10H19O6PS2	330.0361	121-75-5
	Methylparaben	COC(=O)C1=CC=C(C=C1)O	C8H8O3	152.0473	99-76-3
	Morphine	CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O	C17H19NO3	285.1365	47106-99-0
	Oxazepam	C1=CC=C(C=C1)C2=NC(C(=O)NC3=C2C=C(C=C3)Cl)O	C15H11ClN2O2	286.0509	35295-88-6
	Oxybenzone	COC1=CC(=C(C=C1)C(=O)C2=CC=CC=C2)O	C14H12O3	228.0786	58392-22-6
	Pivmecillinam	CC1(C(N2C(S1)C(C2=O)N=CN3CCCCCC3)C(=O)OCOC(=O)C(C)(C)C)C	C21H33N3O5S	439.2141	32886-97-8
	Pregnenolone	CC(=O)C1CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C	C21H32O2	316.2402	145-13-1
	Progesterone-17-hydroxy	CC(=O)C1(CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C)O	C21H30O3	330.2195	68-96-2
	Propylparaben	CCCOC(=O)C1=CC=C(C=C1)O	C10H12O3	180.0786	94-13-3
	Salicylic acid	C1=CC=C(C(=C1)C(=O)O)O	C7H6O3	138.0317	7681-06-3
	Triclosan	C1=CC(=C(C=C1Cl)O)OC2=C(C=C(C=C2)Cl)Cl	C12H7Cl3O2	287.9512	3380-34-5
	Tryptamine-5-hydroxy	C1=CC2=C(C=C1O)C(=CN2)CCN	C10H12N2O	176.0950	50-67-9
	Tryptophan	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N	C11H12N2O2	204.0899	73-22-3

3.4. Table A4 – Calibrant sets used in positive and negative mode for the RTI platform

Table A4 – Calibrant sets used in positive and negative ionization modes for the RTI platform

ESI (+)

Compound Name	Molecular formula	$[M+H]^+$
Guanylurea	C2H6N4O	103.0614
Amitrole	C2H4N4	85.0509
Histamine	C5H9N3	112.0869
Chlormequate	C5H13ClN	123.0809
Methamidophos	C2H8NO2PS	142.0086
Vancomycin	C66H75Cl2N9O24	1448.4375
Cefoperazone	C25H27N9O8S2	646.1497
Trichlorfon (Dylox)	C4H8Cl3O4P	256.9299
Butocarboxim	C7H14N2O2S	191.0849
Dichlorvos	C4H7Cl2O4P	220.9532
Tylosin	C46H77NO17	916.5264
TCMTB	C9H6N2S3	238.9766
Rifaximin	C43H51N3O11	786.3596
Spinosad A	C41H65NO10	732.4681
Emamectin B1a	C49H75NO13	886.5311
Avermectin B1a	C48H72O14	873.4995
Nigericin	C40H68O11	725.4834
Ivermectin B1a	C48H74O14	875.5151

ESI (-)

Compound Name	Molecular formula	$[M-H]^-$
Amitrole	C2H4N4	83.0363
Benzoic acid	C7H6O2	121.0295
Acephate	C4H10NO3PS	182.0046
Salicylic acid	C7H6O3	137.0244
Simazine 2-Hydroxy	C7H13N5O	182.1047
Tepraloxydim	C17H24ClNO4	340.1321
Bromoxynil	C7H3Br2NO	273.8509
MCPA	C9H9ClO3	199.0167
Valproic acid	C8H16O2	143.1078
Phenytoin	C15H12N2O2	251.0826
Flamprop	C16H13ClFNO3	320.0495
Benodanil	C13H10INO	321.9734
Dinoterb	C10H12N2O5	239.0673
Inabenfide	C19H15ClN2O2	337.0749
Coumaphos	C14H16ClO5PS	361.0072
Triclosan	C12H7Cl3O2	286.9438
AvermectinB1a	C48H72O14	871.4849
Salinomycin	C42H70O11	749.4845

3.5. Table A5.1 – Results of data processing workflows on individual compounds in serum

Table A5.1 – Results of data processing workflows on individual compounds in serum

			XCMS - Default settings - Noise 10					XCMS - Optimized settings - Noise 10				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74	7971	7489	1.9E-01	1.1	10.2					
Paracetamol	152.0706	4.98	15318	13185	3.4E-03	1.2	11.9	4283	3482	4.4E-01	1.2	22.3
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31						1742	1496	8.2E-02	1.2	86.8
Carbendazim	192.0768	5.69										
Cyprodinil	226.1339	33.22	132274	234	7.7E-04	565.2	17.9	19141	115	1.7E-03	166.0	18.4
Carbamazepine	237.1022	18.01										
Clothianidin	250.016	7.99						298	148	5.1E-02	2.0	17.5
Thiacloprid	253.0309	12.24						5335	448	4.7E-03	11.9	24.4
Imidacloprid	256.0596	8.57						287	52	2.8E-04	5.5	14.2
Acetochlor	270.1255	40.57	3492	141	1.2E-02	24.7	45.4	300	120	1.4E-02	2.5	21.6
estrone	271.1693	31.60						297	86	4.3E-03	3.4	23.6
venlafaxine	278.2115	9.84	46103	149	4.5E-03	310.2	32.8	49197	39	1.3E-03	1250.9	16.7
Piperine	286.1444	36.42	537209	350581	6.1E-03	1.5	14.2	33959	28902	2.1E-01	1.2	4.4
Androstenedione	287.2006	31.50	55204	2253	4.9E-04	24.5	1.1	15705	619	2.4E-03	25.4	20.1
Testosterone	289.2168	28.90	61331	5213	4.2E-04	11.8	15.8	18275	7392	2.0E-03	2.5	13.0
Thiamethoxam	292.0266	6.97										
Codeine	300.1594	5.12						1463	549	8.8E-03	2.7	12.3
Diazinon	305.1083	43.38										
sertraline	306.0811	24.34	13392	372	5.6E-04	36.0	15.7	1733	101	1.5E-04	17.2	11.7
Tebuconazole	308.1524	39.36	118602	840	7.3E-04	141.1	17.5	107685	246	4.9E-03	437.3	26.5
fluoxetine	310.1413	23.71										
Aflatoxin B1	313.0707	17.52	1514	202	2.0E-05	7.5	6.1					
Progesterone	315.2339	42.10	70692	5845	3.8E-02	12.1	69.0	55997	2517	4.1E-05	22.2	9.8
paroxetine	330.15	18.34	81401	179	1.3E-03	455.5	21.1	6089	24	1.3E-03	258.1	17.0
Propiconazole	342.0771	41.73	106341	391	7.2E-04	272.0	17.5	54815	732	4.1E-04	74.9	11.5
Boscalid	343.0399	38.00						15360	852	5.5E-03	18.0	28.7
Chlorpyrifos	349.9336	45.53	4085	98	5.8E-03	41.5	35.6	4636	96	7.8E-05	48.3	8.2
Cortisone	361.2006	16.12	46168	22183	5.7E-04	2.1	11.7	20860	10110	3.0E-02	2.1	22.2
hydrocortisone	363.2166	15.86	232832	179367	9.0E-03	1.3	11.2	155061	65909	3.7E-02	2.4	34.6
Prochloraz	376.0381	38.74										
Solanidine	398.342	24.54	177510	2537	1.1E-03	70.0	19.7	73463	326	1.1E-04	225.5	7.3
Azoxystrobine	404.1241	38.03	108006	170	6.0E-04	636.1	16.4	74841	53	4.7E-03	1399.3	26.3
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95	7590	785	2.9E-05	9.7	1.6	2528	16	2.5E-02	155.8	47.4
2-phenylphenol	169.0659	30.19						531	242	1.7E-04	2.2	8.7
Hydroxyindoleacetic acid	190.051	5.71										
Ibuprofen	205.1223	39.94	3184	2425	2.2E-03	1.3	7.9	3119	574	6.4E-04	5.4	13.2
Diclofenac	294.0094	39.59	11268	704	9.8E-06	16.0	7.7	5139	283	8.5E-03	18.1	31.5
Arachidonic Acid	303.233	47.00	782488	624374	1.1E-01	1.3	6.8	29062	23443	7.0E-02	1.2	13.9
Leukotriene B4	335.2228	39.52	664	494	3.8E-02	1.3	19.2	32484	330	6.0E-02	98.4	67.3
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50	2928	419	1.4E-02	7.0	44.9	3823	131	3.4E-04	29.3	13.4
Prostaglandin F2a	353.2333	25.60	69517	152	5.4E-05	458.2	7.3	45657	260	2.6E-02	175.7	48.1
Leukotriene D4	495.2534	33.04	22927	342	8.4E-03	67.0	40.7					

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			XCMS - Optimized settings - Noise 20					XCMS - Optimized settings - Noise 50				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74										
Paracetamol	152.0706	4.98	4133	3185	3.2E-01	1.3	19.7	4161	3215	3.2E-01	1.3	20.9
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31	1603	1645	6.9E-02	1.0	77.1					
Carbendazim	192.0768	5.69										
Cyprodinil	226.1339	33.22	20959	121	2.8E-04	173.4	10.1	21173	121	3.1E-04	174.4	10.4
Carbamazepine	237.1022	18.01										
Clothianidin	250.016	7.99	300	145	4.8E-02	2.1	11.2	299	144	4.8E-02	2.1	11.0
Thiacloprid	253.0309	12.24	4907	440	2.1E-03	11.2	19.1	4879	447	2.5E-03	10.9	20.2
Imidacloprid	256.0596	8.57	312	58	6.1E-04	5.4	16.1					
Acetochlor	270.1255	40.57	334	120	1.8E-02	2.8	30.8	339	120	2.0E-02	2.8	32.3
estrone	271.1693	31.60	322	96	2.1E-04	3.4	12.3	324	97	1.9E-04	3.4	11.5
venlafaxine	278.2115	9.84	51685	40	3.1E-04	1307.1	10.5	51527	39	2.6E-04	1314.9	9.8
Piperine	286.1444	36.42	38129	27607	2.6E-02	1.4	8.4	38231	28098	3.3E-02	1.4	8.8
Androstenedione	287.2006	31.50	16605	639	1.8E-03	26.0	18.3	16713	631	2.0E-03	26.5	19.2
Testosterone	289.2168	28.90	16475	7934	8.5E-03	2.1	15.0	16367	7925	8.7E-03	2.1	15.8
Thiamethoxam	292.0266	6.97										
Codeine	300.1594	5.12	1602	581	7.0E-03	2.8	9.1	1599	576	6.1E-03	2.8	9.8
Diazinon	305.1083	43.38										
sertraline	306.0811	24.34	1603	105	1.6E-06	15.2	6.9	1599	104	1.4E-06	15.3	6.7
Tebuconazole	308.1524	39.36	106629	263	1.1E-05	405.6	3.4	107178	267	2.2E-05	401.5	4.3
fluoxetine	310.1413	23.71										
Aflatoxin B1	313.0707	17.52										
Progesterone	315.2339	42.10	59068	2575	2.3E-05	22.9	8.9	58660	2610	4.4E-05	22.5	9.9
paroxetine	330.15	18.34	6212	24	1.4E-03	263.1	17.3	6322	24	1.4E-03	265.9	17.5
Propiconazole	342.0771	41.73	54964	649	3.9E-05	84.7	5.4	54965	644	4.3E-05	85.3	5.6
Boscalid	343.0399	38.00	13699	974	8.0E-04	14.1	19.4	13577	979	6.2E-04	13.9	18.6
Chlorpyrifos	349.9336	45.53	4627	108	8.6E-04	42.9	15.5	4627	110	8.4E-04	42.0	15.4
Cortisone	361.2006	16.12	21315	9677	1.6E-02	2.2	8.8	21423	17043	1.6E-02	1.3	8.8
hydrocortisone	363.2166	15.86	150246	69702	2.4E-03	2.2	13.2	154997	74342	3.5E-04	2.1	10.1
Prochloraz	376.0381	38.74										
Solanidine	398.342	24.54	70258	319	1.2E-04	220.0	7.5	70272	322	1.6E-04	218.1	8.3
Azoxystrobine	404.1241	38.03	74106	60	4.0E-03	1233.1	24.7	74117	61	3.5E-03	1222.2	23.8
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95	2515	28	2.6E-02	90.9	47.9	2515	28	2.6E-02	90.9	47.9
2-phenylphenol	169.0659	30.19	479	242	4.6E-03	2.0	17.4	479	242	4.6E-03	2.0	17.4
Hydroxyindoleacetic acid	190.051	5.71										
Ibuprofen	205.1223	39.94	3119	596	8.8E-04	5.2	13.2	3119	596	8.8E-04	5.2	13.2
Diclofenac	294.0094	39.59										
Arachidonic Acid	303.233	47.00	29020	23443	7.5E-02	1.2	14.1	29020	23443	7.5E-02	1.2	14.1
Leukotriene B4	335.2228	39.52	36247	284	3.5E-02	127.5	53.9	36247	284	3.5E-02	127.5	53.9
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50	3823	121	3.2E-04	31.5	13.4	3823	121	3.2E-04	31.5	13.4
Prostaglandin F2a	353.2333	25.60	37130	258	3.3E-02	144.1	52.9	37130	258	3.3E-02	144.1	52.9
Leukotriene D4	495.2534	33.04										

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			XCMS - Optimized settings - Noise 100					Markerview - Noise 10				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74						787	4	1.1E-04	217.3	9.3
Paracetamol	152.0706	4.98	3570	3021	5.6E-01	1.2	23.1	109	62	7.1E-02	1.7	17.1
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31						373	0	2.7E-04	Infinity	12.6
Carbendazim	192.0768	5.69						754	0	1.9E-04	Infinity	11.1
Cyprodinil	226.1339	33.22	21875	110	4.0E-03	198.9	24.7	3802	0	9.0E-04	Infinity	18.9
Carbamazepine	237.1022	18.01						692	0	5.1E-05	Infinity	7.2
Clothianidin	250.016	7.99	312	123	1.2E-02	2.5	17.5	20	0	3.7E-02	Infinity	73.8
Thiacloprid	253.0309	12.24	4903	437	9.0E-03	11.2	30.6	416	0	2.0E-04	Infinity	11.4
Imidacloprid	256.0596	8.57						140	0	3.7E-04	Infinity	14.0
Acetochlor	270.1255	40.57	340	117	2.3E-02	2.9	34.5					
estrone	271.1693	31.60	324	95	5.9E-04	3.4	16.1	248	0	2.8E-03	Infinity	27.8
venlafaxine	278.2115	9.84	52120	35	8.1E-04	1476.3	14.4	1505	0	1.5E-04	Infinity	10.4
Piperine	286.1444	36.42	35990	29145	2.3E-01	1.2	12.8	17599	11290	1.1E-02	1.5	17.7
Androstenedione	287.2006	31.50	15786	552	1.6E-03	28.6	17.7	1515	0	4.3E-04	Infinity	14.7
Testosterone	289.2168	28.90	15303	7209	4.2E-03	2.1	13.2	1683	0	6.5E-04	Infinity	16.9
Thiamethoxam	292.0266	6.97						63	0	1.4E-04	Infinity	10.2
Codeine	300.1594	5.12	1536	503	2.8E-03	3.1	19.4	1808	0	1.6E-04	Infinity	10.6
Diazinon	305.1083	43.38						3956	0	9.8E-05	Infinity	8.9
sertraline	306.0811	24.34	1525	96	1.8E-04	15.9	12.0	193	0	7.4E-03	Infinity	39.5
Tebuconazole	308.1524	39.36	96732	253	5.6E-04	382.5	12.7	3298	0	9.7E-04	Infinity	19.4
fluoxetine	310.1413	23.71						701	0	1.2E-02	Infinity	46.6
Aflatoxin B1	313.0707	17.52						2462	0	6.8E-04	Infinity	17.1
Progesterone	315.2339	42.10	59890	2536	6.4E-05	23.6	9.9					
paroxetine	330.15	18.34	5471	23	9.7E-04	240.7	15.3	2289	0	1.6E-03	Infinity	23.0
Propiconazole	342.0771	41.73	53083	629	9.0E-04	84.5	14.8	2874	0	8.6E-04	Infinity	18.6
Boscalid	343.0399	38.00	13766	991	1.7E-03	13.9	22.5	1082	0	1.2E-03	Infinity	21.1
Chlorpyrifos	349.9336	45.53	4583	101	1.5E-03	45.4	18.3					
Cortisone	361.2006	16.12	20361	18119	1.0E-01	1.1	2.7	1251	539	2.4E-03	2.3	15.5
hydrocortisone	363.2166	15.86	151256	75039	1.3E-03	2.0	12.2	6954	5331	1.1E-02	1.3	11.9
Prochloraz	376.0381	38.74						467	0	6.5E-03	Infinity	37.2
Solanidine	398.342	24.54	70308	275	2.0E-04	256.1	9.0	5204	0	1.3E-03	Infinity	21.2
Azoxystrobine	404.1241	38.03	72797	55	5.4E-03	1315.5	27.6	3039	0	8.1E-04	Infinity	18.2
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95	2012	26	8.4E-02	76.6	77.4	63	0	4.6E-03	Infinity	33.4
2-phenylphenol	169.0659	30.19	355	210	3.3E-02	1.7	23.9					
Hydroxyindoleacetic acid	190.051	5.71										
Ibuprofen	205.1223	39.94	1426	662	1.7E-01	2.2	59.6	54	33	2.5E-03	1.7	12.5
Diclofenac	294.0094	39.59						230	0	1.5E-04	Infinity	10.4
Arachidonic Acid	303.233	47.00	28959	23443	7.6E-02	1.2	14.1	23349	18582	1.1E-01	1.3	6.4
Leukotriene B4	335.2228	39.52	26531	644	6.5E-02	41.2	68.5	1494	0	1.3E-04	Infinity	9.9
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50	8190	36	1.6E-01	225.8	105.7	379	0	1.9E-03	Infinity	24.2
Prostaglandin F2a	353.2333	25.60	31087	53	9.5E-02	587.6	83.0	1990	0	6.2E-05	Infinity	7.7
Leukotriene D4	495.2534	33.04						733	0	2.8E-04	Infinity	12.7

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			Markerview - Noise 20					Markerview - Noise 50				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74	184	163	6.1E-02	1.1	9.7	642	0	9.5E-05	Infinity	8.9
Paracetamol	152.0706	4.98	1171	624	1.5E-05	1.9	4.5	523	349	5.8E-02	1.5	30.4
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31	206	0	4.4E-03	Infinity	32.7					
Carbendazim	192.0768	5.69	700	0	1.7E-04	Infinity	10.8	586	0	4.2E-04	Infinity	14.6
Cyprodinil	226.1339	33.22	3644	0	1.0E-03	Infinity	19.7	3371	0	1.2E-03	Infinity	21.1
Carbamazepine	237.1022	18.01	659	0	7.3E-05	Infinity	8.1	579	0	1.1E-05	Infinity	4.3
Clothianidin	250.016	7.99	3	0	2.0E-01	Infinity	200.0					
Thiacloprid	253.0309	12.24	371	0	2.5E-04	Infinity	12.2	193	0	3.1E-02	Infinity	69.1
Imidacloprid	256.0596	8.57	100	0	3.9E-03	Infinity	31.4	19	0	2.0E-01	Infinity	200.0
Acetochlor	270.1255	40.57										
estrone	271.1693	31.60	169	0	1.4E-02	Infinity	49.4	10	0	2.0E-01	Infinity	200.0
venlafaxine	278.2115	9.84	1408	0	2.3E-04	Infinity	12.0	1230	0	2.2E-04	Infinity	11.7
Piperine	286.1444	36.42	17262	10940	1.0E-02	1.6	17.9	16342	10025	9.9E-03	1.6	19.0
Androstenedione	287.2006	31.50	1414	0	5.3E-04	Infinity	15.8	1185	0	6.8E-04	Infinity	17.2
Testosterone	289.2168	28.90	1594	0	8.5E-04	Infinity	18.5	1389	0	1.1E-03	Infinity	19.9
Thiamethoxam	292.0266	6.97	8	0	2.0E-01	Infinity	200.0					
Codeine	300.1594	5.12	1690	0	1.1E-04	Infinity	9.3	1523	0	1.3E-04	Infinity	9.8
Diazinon	305.1083	43.38	3746	0	1.2E-04	Infinity	9.6	3078	0	1.5E-04	Infinity	10.3
sertraline	306.0811	24.34	57	0	1.1E-02	Infinity	45.3					
Tebuconazole	308.1524	39.36	3172	0	1.0E-03	Infinity	19.9	2842	0	1.6E-03	Infinity	22.8
fluoxetine	310.1413	23.71	682	0	4.3E-03	Infinity	32.3	308	0	7.2E-02	Infinity	101.8
Aflatoxin B1	313.0707	17.52	2339	0	8.0E-04	Infinity	18.2	2028	0	7.4E-04	Infinity	17.6
Progesterone	315.2339	42.10										
paroxetine	330.15	18.34	2141	0	1.9E-03	Infinity	24.3	1811	0	1.7E-03	Infinity	23.7
Propiconazole	342.0771	41.73	2731	0	8.4E-04	Infinity	18.5	2428	0	1.2E-03	Infinity	21.0
Boscalid	343.0399	38.00	972	0	2.1E-03	Infinity	25.1	587	0	2.3E-02	Infinity	61.1
Chlorpyrifos	349.9336	45.53										
Cortisone	361.2006	16.12	1153	478	2.2E-03	2.4	16.1	1004	315	1.1E-03	3.2	16.1
hydrocortisone	363.2166	15.86	6819	5196	9.8E-03	1.3	11.9	6532	4978	1.2E-02	1.3	12.6
Prochloraz	376.0381	38.74	418	0	7.8E-03	Infinity	40.2	86	0	2.0E-01	Infinity	200.0
Solanidine	398.342	24.54	4981	0	1.4E-03	Infinity	22.2	4491	0	1.5E-03	Infinity	22.6
Azoxystrobine	404.1241	38.03	2859	0	1.2E-03	Infinity	20.6	2441	0	2.6E-03	Infinity	27.3
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95										
2-phenylphenol	169.0659	30.19										
Hydroxyindoleacetic acid	190.051	5.71										
Ibuprofen	205.1223	39.94	10	0	1.1E-01	Infinity	125.9					
Diclofenac	294.0094	39.59	187	0	2.1E-04	Infinity	11.5	27	0	2.1E-01	Infinity	173.2
Arachidonic Acid	303.233	47.00	23091	18397	1.1E-01	1.3	6.3	23123	18542	1.2E-01	1.2	5.9
Leukotriene B4	335.2228	39.52	1432	0	1.6E-04	Infinity	10.6	1295	302	2.0E-02	4.3	15.0
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50	329	0	2.0E-03	Infinity	24.9	85	0	2.1E-01	Infinity	173.2
Prostaglandin F2a	353.2333	25.60	1909	0	7.1E-05	Infinity	8.0	55	0	9.3E-02	Infinity	87.5
Leukotriene D4	495.2534	33.04	617	0	4.3E-04	Infinity	14.7	311	0	5.1E-02	Infinity	60.2

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			Markerview - Noise 100					MzMine - CWT pipeline - Default settings - Noise 10				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74	219	0	9.1E-02	Infinity	115.5					
Paracetamol	152.0706	4.98						16252	12257	6.0E-02	1.3	4.2
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31						20628	1308	7.7E-06	15.8	6.4
Carbendazim	192.0768	5.69	286	0	5.6E-02	Infinity	90.1	31313	1953	1.5E-04	16.0	9.7
Cyprodinil	226.1339	33.22	2929	0	1.9E-03	Infinity	24.4	140937	5962	6.5E-04	23.6	16.8
Carbamazepine	237.1022	18.01	277	0	4.8E-03	Infinity	33.8	37137	2381	6.8E-04	15.6	17.7
Clothianidin	250.016	7.99						2797	632	2.6E-04	4.4	12.5
Thiacloprid	253.0309	12.24						17484	3017	2.9E-04	5.8	10.9
Imidacloprid	256.0596	8.57						7009	703	1.8E-04	10.0	10.0
Acetochlor	270.1255	40.57						2324	1712	1.7E-02	1.4	14.8
estrone	271.1693	31.60						14105	2602	1.1E-03	5.4	16.7
venlafaxine	278.2115	9.84	916	0	9.8E-04	Infinity	19.5	52957	49	1.1E-04	1085.8	9.3
Piperine	286.1444	36.42	3206	0	1.1E-03	Infinity	20.2	170988	115806	3.2E-02	1.5	22.2
Androstenedione	287.2006	31.50	690	0	6.9E-03	Infinity	38.5	55350	3405	3.0E-04	16.3	12.5
Testosterone	289.2168	28.90	1016	0	1.8E-03	Infinity	23.9	61448	1840	7.4E-04	33.4	17.3
Thiamethoxam	292.0266	6.97						4074	532	3.6E-04	7.7	12.4
Codeine	300.1594	5.12	1280	0	1.9E-04	Infinity	11.3	66600	9622	1.6E-05	6.9	11.4
Diazinon	305.1083	43.38	2017	0	7.4E-04	Infinity	17.7	148106	513	7.9E-05	288.9	8.3
sertraline	306.0811	24.34						12814	1635	5.1E-05	7.8	6.4
Tebuconazole	308.1524	39.36	2352	0	3.2E-03	Infinity	29.3	121191	3791	7.6E-04	32.0	17.3
fluoxetine	310.1413	23.71						34625	3606	2.7E-03	9.6	26.4
Aflatoxin B1	313.0707	17.52	1720	0	1.2E-03	Infinity	21.0					
Progesterone	315.2339	42.10						92174	5124	7.2E-04	18.0	16.8
paroxetine	330.15	18.34	1399	0	4.9E-03	Infinity	34.0	63835	1690	4.8E-03	37.8	32.8
Propiconazole	342.0771	41.73	1877	0	3.4E-03	Infinity	30.0	106056	1606	7.7E-04	66.1	17.6
Boscalid	343.0399	38.00	245	0	1.4E-01	Infinity	148.7	42716	1367	8.8E-04	31.2	18.1
Chlorpyrifos	349.9336	45.53										
Cortisone	361.2006	16.12	715	19	3.4E-04	37.7	16.9	48932	19052	3.4E-03	2.6	11.7
hydrocortisone	363.2166	15.86	6141	4607	1.2E-02	1.3	13.3	3232	43681	2.0E-01	0.1	30.6
Prochloraz	376.0381	38.74						26450	880	1.1E-03	30.1	19.3
Solanidine	398.342	24.54	3759	0	3.1E-03	Infinity	28.8					
Azoxystrobine	404.1241	38.03	1773	0	1.1E-02	Infinity	45.1	109048	1310	6.3E-04	83.2	16.5
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95						5745	490	4.1E-07	11.7	4.4
2-phenylphenol	169.0659	30.19										
Hydroxyindoleacetic acid	190.051	5.71										
Ibuprofen	205.1223	39.94						2803	2578	2.9E-01	1.1	26.2
Diclofenac	294.0094	39.59						8796	2055	1.6E-02	4.3	40.7
Arachidonic Acid	303.233	47.00	22072	17499	1.1E-01	1.3	6.7	795212	637697	1.2E-01	1.2	6.2
Leukotriene B4	335.2228	39.52	1059	0	4.6E-04	Infinity	15.1	38523	5277	2.9E-02	7.3	57.7
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50						14764	761	1.3E-02	19.4	46.0
Prostaglandin F2a	353.2333	25.60	1451	0	1.6E-04	Infinity	10.6					
Leukotriene D4	495.2534	33.04						18084	1338	2.9E-02	13.5	61.8

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			MzMine - CWT pipeline - Optimized settings - Noise 10					MzMine - ADAP pipeline - Optimized settings - Noise 50				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74	149671	106487	5.6E-03	1.4	11.8	30266	3167	9.5E-05	9.6	9.2
Paracetamol	152.0706	4.98	17027	13473	5.5E-04	1.3	5.4	17195	10666	9.6E-02	1.6	37.6
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31	22347	974	4.2E-05	23.0	6.6	19515	716	7.5E-05	27.3	8.5
Carbendazim	192.0768	5.69	34511	1712	1.2E-04	20.2	9.0	25071	1075	6.0E-03	23.3	35.1
Cyprodinil	226.1339	33.22	134023	7031	3.7E-12	19.1	0.8	131783	4385	9.2E-04	30.1	18.4
Carbamazepine	237.1022	18.01	32584	3369	5.7E-04	9.7	14.5	32150	2662	2.8E-03	12.1	29.2
Clothianidin	250.016	7.99	2216	870	3.1E-06	2.5	5.9	2140	470	1.9E-03	4.6	19.4
Thiacloprid	253.0309	12.24	12799	3475	3.4E-04	3.7	11.2	16508	578	1.6E-04	28.6	10.1
Imidacloprid	256.0596	8.57	7543	616	2.6E-04	12.2	12.0	6729	538	2.4E-04	12.5	11.3
Acetochlor	270.1255	40.57	2431	1492	4.2E-04	1.6	9.4	4152	3282	2.0E-01	1.3	41.7
estrone	271.1693	31.60	14276	2834	8.7E-06	5.0	4.2	12909	2009	1.1E-03	6.4	17.3
venlafaxine	278.2115	9.84	51775	57	3.3E-05	907.3	6.2	51850	43	1.4E-04	1202.1	10.2
Piperine	286.1444	36.42	162513	80321	5.2E-04	2.0	11.0	664947	426816	8.4E-03	1.6	16.6
Androstenedione	287.2006	31.50	50608	3262	3.8E-05	15.5	6.3	54171	2028	4.3E-04	26.7	14.2
Testosterone	289.2168	28.90	69964	2129	2.4E-04	32.9	11.8	60408	1479	6.7E-04	40.8	16.8
Thiamethoxam	292.0266	6.97	3061	401	6.1E-05	7.6	7.2	3993	504	5.6E-04	7.9	14.4
Codeine	300.1594	5.12	55856	6481	3.2E-04	8.6	12.0	64510	1368	1.4E-04	47.1	10.0
Diazinon	305.1083	43.38	153302	786	1.2E-04	195.0	9.7	144666	161	8.8E-05	898.8	8.6
sertraline	306.0811	24.34	12301	2347	3.7E-05	5.2	8.2	11976	1534	1.3E-03	7.8	18.7
Tebuconazole	308.1524	39.36	116564	3123	1.4E-04	37.3	9.8	118487	3212	7.4E-04	36.9	17.2
fluoxetine	310.1413	23.71	23954	5180	3.2E-04	4.6	12.1	32463	2950	1.2E-03	11.0	19.1
Aflatoxin B1	313.0707	17.52	1524	562	1.2E-03	2.7	14.5	1273	635	2.9E-03	2.0	15.3
Progesterone	315.2339	42.10	88544	3667	5.4E-04	24.1	15.3	89845	5074	7.6E-04	17.7	17.0
paroxetine	330.15	18.34	66077	1666	7.1E-06	39.7	3.6	56070	1210	1.6E-02	46.3	51.0
Propiconazole	342.0771	41.73	88793	1712	4.7E-05	51.9	6.9	105646	1540	7.6E-04	68.6	17.6
Boscalid	343.0399	38.00	33419	1398	2.5E-04	23.9	11.7	41300	906	8.9E-04	45.6	18.4
Chlorpyrifos	349.9336	45.53										
Cortisone	361.2006	16.12	42547	2016	4.9E-05	21.1	7.0	44318	16248	1.9E-03	2.7	13.5
hydrocortisone	363.2166	15.86	3134	139721	1.1E-03	0.0	16.0	232703	138042	6.5E-02	1.7	11.5
Prochloraz	376.0381	38.74	26004	886	7.2E-05	29.4	7.8	25631	694	1.2E-03	36.9	20.5
Solanidine	398.342	24.54						161031	2162	7.1E-05	74.5	8.1
Azoxystrobine	404.1241	38.03	80322	1151	1.4E-04	69.8	10.0	106729	762	5.7E-04	140.1	16.1
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95	7281	811	3.7E-05	9.0	6.1	6531	736	1.2E-06	8.9	5.4
2-phenylphenol	169.0659	30.19										
Hydroxyindoleacetic acid	190.051	5.71	5233	184	2.0E-04	28.5	11.0					
Ibuprofen	205.1223	39.94	2440	1703	6.3E-04	1.4	6.8	3048	2302	2.6E-03	1.3	8.1
Diclofenac	294.0094	39.59	12338	2027	2.5E-08	6.1	2.4	11031	2034	7.1E-05	5.4	7.4
Arachidonic Acid	303.233	47.00	1395321	747617	4.2E-05	1.9	5.4					
Leukotriene B4	335.2228	39.52	11630	388	4.3E-05	30.0	6.6	54961	937	1.0E-04	58.6	9.1
Prostaglandin D2	351.2177	27.60						1930	355	4.2E-02	5.4	64.0
Prostaglandin E2	351.2177	26.50	4864	974	2.8E-05	5.0	4.9	10200	685	2.3E-02	14.9	56.4
Prostaglandin F2a	353.2333	25.60						50282	181	2.8E-02	277.7	65.9
Leukotriene D4	495.2534	33.04	2212	1833	2.5E-03	1.2	3.1	21981	1460	2.6E-02	15.1	60.1

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			MzMine - ADAP pipeline - Optimized settings - Noise 100					Progenesis - More sensitivity				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74	30406	7306	6.7E-06	4.2	8.7	91516	234	6.5E-05	391.3	6.3
Paracetamol	152.0706	4.98										
Nicotine	163.123	3.37										
Cotinine	177.1022	4.31	19498	669	7.3E-05	29.2	8.4					
Carbendazim	192.0768	5.69	25071	1075	6.0E-03	23.3	35.1	96635	2714	6.9E-05	35.6	8.6
Cyprodinil	226.1339	33.22	131783	4385	9.2E-04	30.1	18.4	481950	1674	6.9E-04	287.9	13.6
Carbamazepine	237.1022	18.01	31518	2923	5.1E-03	10.8	33.8	145456	2125	3.0E-06	68.4	5.2
Clothianidin	250.016	7.99	2048	364	2.3E-03	5.6	21.7					
Thiacloprid	253.0309	12.24	16418	538	1.7E-04	30.5	10.4	47488	9	1.9E-04	5304.4	8.9
Imidacloprid	256.0596	8.57	6625	447	2.5E-04	14.8	11.5	16865	20	1.1E-03	831.8	16.0
Acetochlor	270.1255	40.57	4629	3074	6.6E-02	1.5	32.8					
estrone	271.1693	31.60	13050	2084	1.4E-03	6.3	18.4	1623	0	3.9E-01	Infinity	200.0
venlafaxine	278.2115	9.84	52088	43	1.3E-04	1207.6	9.7	161418	0	2.1E-04	Infinity	9.2
Piperine	286.1444	36.42	664947	426816	8.4E-03	1.6	16.6	2526592	1804926	3.6E-02	1.4	12.3
Androstenedione	287.2006	31.50	54171	2025	4.1E-04	26.7	14.2	177679	118	5.2E-04	1511.7	12.4
Testosterone	289.2168	28.90	60408	2091	7.0E-04	28.9	16.8	182774	0	6.5E-05	Infinity	6.2
Thiamethoxam	292.0266	6.97	3917	485	5.0E-04	8.1	13.9	5345	38	8.0E-02	139.0	76.2
Codeine	300.1594	5.12	64510	1368	1.4E-04	47.1	10.0	224949	303	1.9E-04	742.9	8.9
Diazinon	305.1083	43.38	144666	161	8.8E-05	898.8	8.6	478989	0	7.5E-05	Infinity	6.5
sertraline	306.0811	24.34	11976	1534	1.3E-03	7.8	18.7	7957	0	4.3E-03	Infinity	25.5
Tebuconazole	308.1524	39.36	118487	3212	7.4E-04	36.9	17.2	424554	792	9.5E-04	535.8	15.2
fluoxetine	310.1413	23.71	32463	2950	1.2E-03	11.0	19.1	75793	92	1.6E-03	823.9	18.3
Aflatoxin B1	313.0707	17.52	1167	513	1.3E-03	2.3	15.0					
Progesterone	315.2339	42.10	89818	5005	7.6E-04	17.9	17.0	302956	0	6.3E-04	Infinity	13.2
paroxetine	330.15	18.34	56070	1210	1.6E-02	46.3	51.0	271775	1	1.9E-03	187919.2	19.2
Propiconazole	342.0771	41.73	105642	1540	7.6E-04	68.6	17.6	339380	0	1.0E-03	Infinity	15.6
Boscalid	343.0399	38.00	41300	906	8.9E-04	45.6	18.4	109280	0	2.4E-03	Infinity	20.9
Chlorpyrifos	349.9336	45.53										
Cortisone	361.2006	16.12	44424	16453	2.1E-03	2.7	13.4	156782	52025	4.3E-03	3.0	11.2
hydrocortisone	363.2166	15.86	233117	138264	6.5E-02	1.7	11.4	887099	575399	2.0E-01	1.5	8.1
Prochloraz	376.0381	38.74	25627	694	1.2E-03	36.9	20.5	43336	0	3.0E-02	Infinity	51.5
Solanidine	398.342	24.54	172137	2347	9.2E-04	73.3	18.8	713233	0	1.2E-03	Infinity	16.6
Azoxystrobine	404.1241	38.03	106729	762	5.7E-04	140.1	16.1	392714	0	7.9E-04	Infinity	14.3
Pravastatin	425.2534	20.50										
Dimethyldithiophosphate	156.9541	2.95	6476	736	1.1E-06	8.8	5.8	1056	0	1.9E-03	Infinity	19.3
2-phenylphenol	169.0659	30.19						963	415	6.2E-02	2.3	37.9
Hydroxyindoleacetic acid	190.051	5.71						1821	3397	2.5E-01	0.5	73.4
Ibuprofen	205.1223	39.94	2751	1971	2.4E-03	1.4	10.1	780	443	2.8E-02	1.8	5.0
Diclofenac	294.0094	39.59	10908	1961	1.2E-04	5.6	8.4					
Arachidonic Acid	303.233	47.00						1712	212	1.5E-01	8.1	90.2
Leukotriene B4	335.2228	39.52	39778	937	2.8E-02	42.4	64.3	197740	0	3.7E-04	Infinity	11.1
Prostaglandin D2	351.2177	27.60	10186	675	2.3E-02	15.1	56.6					
Prostaglandin E2	351.2177	26.50						57441	0	6.0E-03	Infinity	28.5
Prostaglandin F2a	353.2333	25.60	50173	155	2.9E-02	324.2	66.3	252221	0	1.5E-04	Infinity	8.1
Leukotriene D4	495.2534	33.04	21981	1460	2.6E-02	15.1	60.1	62552	0	4.6E-05	Infinity	5.5

Table A5.1 – (continued) Results of data processing workflows on individual compounds in serum

			Progenesis - Default sensitivity					Manual integration				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.0713	4.74	88520	251	2.3E-05	352.6	4.5	119919	683	2.3E-04	175.6	9.4
Paracetamol	152.0706	4.98						8401	0	7.4E-05	Infinity	6.5
Nicotine	163.123	3.37						6605	0	2.7E-05	Infinity	4.6
Cotinine	177.1022	4.31						68792	2808	3.8E-05	24.5	7.5
Carbendazim	192.0768	5.69	93406	2769	5.4E-06	33.7	6.0	85070	0	3.1E-04	Infinity	10.5
Cyprodinil	226.1339	33.22	465799	1740	4.5E-04	267.6	11.8	383042	0	9.5E-05	Infinity	7.0
Carbamazepine	237.1022	18.01	130723	2226	5.4E-09	58.7	2.8	89375	10357	8.0E-05	8.6	6.7
Clothianidin	250.016	7.99						4952	0	6.1E-04	Infinity	13.1
Thiacloprid	253.0309	12.24	45899	10	6.8E-05	4619.8	6.3	35550	0	4.8E-04	Infinity	12.1
Imidacloprid	256.0596	8.57	16282	23	6.8E-04	723.6	13.6	13617	0	2.9E-04	Infinity	10.2
Acetochlor	270.1255	40.57						7102	0	3.7E-04	Infinity	11.0
estrone	271.1693	31.60						21456	0	2.2E-05	Infinity	4.3
venlafaxine	278.2115	9.84	156056	0	9.6E-05	Infinity	7.1	105740	0	3.1E-04	Infinity	10.4
Piperine	286.1444	36.42	2441946	1848043	3.5E-02	1.3	10.4	1322259	908440	2.3E-02	1.5	4.9
Androstenedione	287.2006	31.50	171669	122	2.8E-04	1407.8	10.0	107952	4105	8.0E-04	26.3	13.9
Testosterone	289.2168	28.90	176907	0	6.2E-05	Infinity	6.1	112541	687	1.1E-06	163.9	1.7
Thiamethoxam	292.0266	6.97						7596	0	4.1E-05	Infinity	5.3
Codeine	300.1594	5.12	217503	316	9.5E-05	688.7	7.0	125800	0	2.8E-04	Infinity	10.1
Diazinon	305.1083	43.38	463596	0	6.8E-05	Infinity	6.3	269902	435	1.5E-04	621.2	8.2
sertraline	306.0811	24.34	7696	0	4.0E-03	Infinity	24.9	19824	0	1.2E-05	Infinity	3.5
Tebuconazole	308.1524	39.36	410252	809	6.4E-04	506.9	13.3	195928	0	1.2E-05	Infinity	3.5
fluoxetine	310.1413	23.71	73224	95	1.2E-03	769.1	16.5	53370	0	3.6E-05	Infinity	5.1
Aflatoxin B1	313.0707	17.52						1807	0	2.5E-05	Infinity	4.5
Progesterone	315.2339	42.10	292783	0	3.9E-04	Infinity	11.3	146106	6952	3.2E-06	21.0	4.0
paroxetine	330.15	18.34	262447	2	1.3E-03	163519.1	17.1	124854	0	2.9E-05	Infinity	4.8
Propiconazole	342.0771	41.73	327888	0	6.9E-04	Infinity	13.6	75859	0	2.0E-05	Infinity	4.2
Boscalid	343.0399	38.00	105598	0	2.0E-03	Infinity	19.5	61505	0	6.9E-05	Infinity	6.3
Chlorpyrifos	349.9336	45.53						8636	499	1.5E-04	17.3	8.8
Cortisone	361.2006	16.12	151477	52494	7.1E-03	2.9	8.8	68182	25248	8.7E-03	2.7	4.1
hydrocortisone	363.2166	15.86	857573	580596	2.5E-01	1.5	5.5	380787	287326	1.8E-02	1.3	11.9
Prochloraz	376.0381	38.74	41713	0	2.7E-02	Infinity	49.2	31082	0	8.7E-05	Infinity	6.8
Solanidine	398.342	24.54	689074	0	8.6E-04	Infinity	14.7	225831	2242	2.6E-05	100.7	4.7
Azoxystrobine	404.1241	38.03	379428	0	4.9E-04	Infinity	12.2	133413	0	2.4E-05	Infinity	4.4
Pravastatin	425.2534	20.50						7801	0	4.4E-04	Infinity	11.8
Dimethyldithiophosphate	156.9541	2.95	942	0	1.2E-01	Infinity	93.0	28970	9888	2.3E-03	2.9	3.7
2-phenylphenol	169.0659	30.19						3857	0	8.3E-04	Infinity	14.5
Hydroxyindoleacetic acid	190.051	5.71						8032	6099	2.2E-02	1.3	12.4
Ibuprofen	205.1223	39.94						8387	6704	1.6E-03	1.3	5.4
Diclofenac	294.0094	39.59						18942	0	3.1E-04	Infinity	10.5
Arachidonic Acid	303.233	47.00	12097	149	1.1E-01	81.3	86.0	1587614	131346	2.0E-05	12.1	4.8
Leukotriene B4	335.2228	39.52	197740	0	3.7E-04	Infinity	11.1	88909	0	1.3E-04	Infinity	7.8
Prostaglandin D2	351.2177	27.60						12086	0	4.8E-04	Infinity	12.1
Prostaglandin E2	351.2177	26.50	57441	0	6.0E-03	Infinity	28.5	1882	0	3.7E-04	Infinity	11.1
Prostaglandin F2a	353.2333	25.60	252221	0	1.5E-04	Infinity	8.1	105160	0	4.9E-05	Infinity	5.6
Leukotriene D4	495.2534	33.04	62552	0	4.6E-05	Infinity	5.5	29833	0	3.3E-04	Infinity	10.6

3.6. Table A5.2 – Results of data processing workflows on individual compounds in plasma

Table A5.2 – Results of data processing workflows on individual compounds in plasma

			XCMS - Default settings - Noise 10					XCMS - Optimized settings - Noise 10				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	24543	13475	1.1E-01	1.8	57.7	24157	13338	1.0E-01	1.8	53.6
Paracetamol	152.0706	4.98						1350	879	1.1E-01	1.5	46.2
Nicotine	163.12297	3.37										
Cotinine	177.1022	4.31	293545	240626	6.2E-03	1.2	6.1	290574	234814	3.4E-03	1.2	3.9
Carbendazim	192.07675	5.69						2528	1885	3.7E-01	1.3	12.4
Cyprodinil	226.1339	33.22	210386	236	1.1E-04	893.1	9.3	201405	223	1.0E-04	901.5	9.0
Carbamazepine	237.10224	18.01	67930	2485	4.0E-05	27.3	6.6	69301	2474	9.0E-05	28.0	8.4
Clothianidin	250.016	7.99										
Thiacloprid	253.0309	12.24	27972	40	1.2E-05	707.4	4.4	27753	41	3.7E-05	684.4	6.5
Imidacloprid	256.0596	8.57						4943	1748	1.2E-01	2.8	86.7
Acetochlor	270.12553	40.57	5887	183	8.5E-04	32.1	18.1	5937	184	6.3E-04	32.3	16.3
estrone	271.1693	31.60	21645	630	4.0E-04	34.4	14.0	21309	651	3.2E-04	32.7	13.0
venlafaxine	278.2115	9.84	89788	211	2.3E-05	425.3	5.5	91286	204	3.7E-05	447.3	6.4
Piperine	286.1444	36.42	1281641	669526	6.3E-04	1.9	12.2	1269497	646795	9.9E-05	2.0	8.9
Androstenedione	287.20056	31.50	96867	9151	1.2E-04	10.6	1.0	97193	9059	1.5E-04	10.7	9.5
Testosterone	289.2168	28.90	26191	21073	4.6E-02	1.2	14.2	27328	21677	5.1E-03	1.3	8.4
Thiamethoxam	292.0266	6.97										
Codeine	300.15942	5.12	26223	19148	1.2E-01	1.4	9.0	26951	18998	9.2E-02	1.4	10.0
Diazinon	305.1083	43.38	234369	1222	1.4E-05	191.8	4.8	249394	1248	6.1E-06	199.8	3.7
sertraline	306.0811	24.34	21377	263	1.8E-04	81.3	10.8	22073	273	7.9E-05	80.7	8.3
Tebuconazole	308.1524	39.36	188846	2331	3.4E-04	81.0	13.6	186567	2287	1.7E-04	81.6	10.8
fluoxetine	310.1413	23.71	2686	1902	1.1E-01	1.4	26.4	2827	1889	6.7E-02	1.5	24.9
Aflatoxin B1	313.07066	17.52						4705	295	1.8E-04	15.9	14.1
Progesterone	315.2339	42.10	72116	207713	1.4E-01	0.3	7.7	74161	214529	1.4E-01	0.3	7.4
paroxetine	330.15	18.34	175779	1452	6.0E-04	121.1	16.3	178127	1506	8.1E-05	118.3	8.3
Propiconazole	342.0771	41.73	172892	143	7.7E-05	1208.9	8.2	167926	139	8.8E-05	1210.9	8.6
Boscalid	343.03994	38.00	65762	125	3.4E-06	527.4	2.9	66875	124	1.5E-04	541.0	10.3
Chlorpyrifos	349.93356	45.53	10068	917	2.0E-04	11.0	12.2	9976	928	7.4E-05	10.8	9.7
Cortisone	361.2006	16.12	97777	50971	1.1E-02	1.9	22.4	97319	48908	8.7E-03	2.0	21.6
hydrocortisone	363.2166	15.86	482937	319086	5.6E-03	1.5	12.9	489548	319592	2.3E-03	1.5	10.6
Prochloraz	376.0381	38.74										
Solanidine	398.342	24.54						203185	1849	1.2E-04	109.9	9.5
Azoxystrobine	404.1241	38.03	194361	502	4.5E-05	386.9	6.9	188971	535	9.6E-05	353.4	8.9
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95	4080	80	7.4E-03	51.3	30.2	1618	14	1.2E-02	111.6	35.9
2-phenylphenol	169.0659	30.19	2161	206	4.7E-06	10.5	7.1	209598	128152	1.2E-01	1.6	36.4
Hydroxyindoleacetic acid	190.051	5.71						3066	818	3.7E-04	3.7	14.6
Ibuprofen	205.1223	39.94	23182	20502	7.0E-02	1.1	0.9	23118	3832	1.6E-05	6.0	6.4
Diclofenac	294.0094	39.59						5809	348	2.7E-04	16.7	12.0
Arachidonic Acid	303.233	47.00						39163	5529	1.9E-04	7.1	8.7
Leukotriene B4	335.2228	39.52						3484103	44646	1.3E-01	78.0	96.1
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50	137290	120533	2.2E-01	1.1	14.5					
Prostaglandin F2a	353.2333	25.60	159424	58088	1.2E-06	2.7	4.8					
Leukotriene D4	495.2534	33.04						1895	246	3.6E-04	7.7	14.9

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			XCMS - Optimized settings - Noise 20					XCMS - Optimized settings - Noise 50				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	24068	13397	1.1E-01	1.8	54.0	23425	13030	1.0E-01	1.8	52.5
Paracetamol	152.0706	4.98	1338	880	1.1E-01	1.5	45.1	1356	877	1.1E-01	1.5	46.3
Nicotine	163.12297	3.37										
Cotinine	177.1022	4.31	291392	233803	4.3E-03	1.2	5.1	293027	233855	2.3E-03	1.3	7.0
Carbendazim	192.07675	5.69	2533	1921	3.8E-01	1.3	14.5	2487	1958	4.0E-01	1.3	11.5
Cyprodinil	226.1339	33.22	203882	223	9.4E-05	915.5	8.8	199907	217	3.1E-05	920.0	6.1
Carbamazepine	237.10224	18.01	68613	2509	9.9E-05	27.3	8.7	65646	2505	6.8E-05	26.2	7.7
Clothianidin	250.016	7.99										
Thiacloprid	253.0309	12.24	27532	41	2.5E-05	673.1	5.6	26724	42	9.5E-05	643.4	8.9
Imidacloprid	256.0596	8.57	4961	1738	1.2E-01	2.9	87.6	4695	1716	1.2E-01	2.7	84.9
Acetochlor	270.12553	40.57	5982	186	6.8E-04	32.2	16.7	5923	185	5.0E-04	32.0	15.2
estrone	271.1693	31.60	21286	650	4.6E-04	32.7	14.7	20467	633	3.9E-04	32.3	13.8
venlafaxine	278.2115	9.84	91732	202	3.4E-05	453.2	6.3	89631	198	2.8E-05	452.1	5.9
Piperine	286.1444	36.42	1259214	646378	1.5E-04	1.9	9.3	1254977	638026	2.2E-04	2.0	10.3
Androstenedione	287.20056	31.50	97464	8986	1.7E-04	10.8	9.9	99169	8777	1.8E-04	11.3	10.1
Testosterone	289.2168	28.90	27333	21486	4.5E-03	1.3	8.8	26419	21220	5.3E-03	1.2	8.3
Thiamethoxam	292.0266	6.97										
Codeine	300.15942	5.12	26955	18961	8.9E-02	1.4	10.3	27203	18574	7.4E-02	1.5	12.5
Diazinon	305.1083	43.38	248197	1260	1.3E-05	197.0	4.7	239035	1185	4.4E-05	201.7	6.9
sertraline	306.0811	24.34	22080	274	9.6E-05	80.7	8.8	21459	270	6.8E-05	79.6	7.9
Tebuconazole	308.1524	39.36	185318	2275	2.2E-04	81.5	11.9	181429	2163	1.7E-04	83.9	10.9
fluoxetine	310.1413	23.71	2803	1912	7.7E-02	1.5	24.4	2711	1884	8.3E-02	1.4	21.3
Aflatoxin B1	313.07066	17.52	4723	299	2.7E-04	15.8	15.4	4703	290	5.8E-04	16.2	18.0
Progesterone	315.2339	42.10	73964	214775	1.4E-01	0.3	7.2	70844	221714	1.4E-01	0.3	3.2
paroxetine	330.15	18.34	177275	1500	9.2E-05	118.2	8.7	172978	1465	1.4E-04	118.1	9.9
Propiconazole	342.0771	41.73	165661	139	4.9E-05	1196.1	7.1	164355	136	3.7E-05	1205.4	6.4
Boscalid	343.03994	38.00	66875	126	1.5E-04	532.0	10.3	66993	122	3.2E-04	547.4	13.3
Chlorpyrifos	349.93356	45.53	10047	921	6.1E-05	10.9	9.2	9842	917	4.7E-05	10.7	8.6
Cortisone	361.2006	16.12	96765	48400	8.3E-03	2.0	21.3	96542	46268	1.0E-02	2.1	23.8
hydrocortisone	363.2166	15.86	488910	318801	1.1E-03	1.5	9.0	475084	316969	4.2E-03	1.5	11.5
Prochloraz	376.0381	38.74										
Solanidine	398.342	24.54	205844	1843	1.6E-04	111.7	10.4	206665	1831	1.1E-04	112.9	9.2
Azoxystrobine	404.1241	38.03	187016	529	8.5E-05	353.5	8.5	186885	522	7.3E-05	358.1	8.1
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95	1613	15	1.2E-02	110.5	36.5	1649	20	1.3E-02	80.7	37.4
2-phenylphenol	169.0659	30.19	209782	127694	1.2E-01	1.6	35.9	211072	129358	1.3E-01	1.6	36.4
Hydroxyindoleacetic acid	190.051	5.71	3092	814	5.2E-04	3.8	15.4	3125	797	7.1E-04	3.9	16.0
Ibuprofen	205.1223	39.94	23052	3815	6.0E-06	6.0	5.8	22934	3833	6.0E-06	6.0	5.8
Diclofenac	294.0094	39.59	5780	348	2.7E-04	16.6	12.0	5776	346	2.2E-04	16.7	11.4
Arachidonic Acid	303.233	47.00	39055	5516	1.6E-04	7.1	8.2	39046	5608	1.5E-04	7.0	8.0
Leukotriene B4	335.2228	39.52	6151446	44540	1.3E-04	138.1	7.8	6155083	39540	2.0E-04	155.7	8.9
Prostaglandin D2	351.2177	27.60										
Prostaglandin E2	351.2177	26.50										
Prostaglandin F2a	353.2333	25.60										
Leukotriene D4	495.2534	33.04	1915	248	4.6E-04	7.7	15.5	1914	247	8.3E-04	7.7	17.3

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			XCMS - Optimized settings - Noise 100					Markerview - Noise 10				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	22686	12593	9.9E-02	1.8	51.9	1005	0	6.9E-05	Infinity	8.0
Paracetamol	152.0706	4.98	1348	867	1.2E-01	1.6	47.5	2026	2117	4.0E-01	1.0	25.0
Nicotine	163.12297	3.37						39	0	4.2E-02	159.0	78.4
Cotinine	177.1022	4.31	288850	226940	4.1E-03	1.3	8.5	8438	6702	2.7E-03	1.3	4.3
Carbendazim	192.07675	5.69	2411	1920	4.0E-01	1.3	7.9	1089	0	1.1E-04	Infinity	9.3
Cyprodinil	226.1339	33.22	200997	210	1.2E-05	955.9	4.4	5717	0	7.4E-05	Infinity	8.2
Carbamazepine	237.10224	18.01	65995	2488	1.0E-04	26.5	8.8	1928	0	1.3E-04	Infinity	9.8
Clothianidin	250.016	7.99						37	0	2.3E-02	Infinity	61.0
Thiacloprid	253.0309	12.24	25687	40	9.1E-05	639.2	8.7	738	0	5.8E-05	Infinity	7.5
Imidacloprid	256.0596	8.57	4670	1703	1.2E-01	2.7	86.6	235	0	1.2E-04	Infinity	9.5
Acetochlor	270.12553	40.57	5903	184	5.1E-04	32.1	15.3					
estrone	271.1693	31.60	20239	623	6.9E-04	32.5	16.8	386	0	2.8E-03	Infinity	27.9
venlafaxine	278.2115	9.84	88317	190	7.0E-05	465.2	8.0	2591	0	5.3E-05	Infinity	7.3
Piperine	286.1444	36.42	1234588	624278	3.7E-04	2.0	10.5	42728	20665	8.0E-05	2.1	8.8
Androstenedione	287.20056	31.50	95354	8524	3.7E-04	11.2	12.8	2564	0	1.8E-04	Infinity	10.9
Testosterone	289.2168	28.90	26181	20993	9.7E-03	1.2	9.2	3465	292	2.2E-04	11.9	11.1
Thiamethoxam	292.0266	6.97						121	0	2.9E-04	Infinity	12.9
Codeine	300.15942	5.12	26222	18313	8.4E-02	1.4	10.7	679	0	1.1E-03	Infinity	20.4
Diazinon	305.1083	43.38	238500	1198	4.7E-05	199.2	7.1	6870	0	1.0E-05	Infinity	4.2
sertraline	306.0811	24.34	21464	265	4.9E-05	80.9	7.0	455	0	1.4E-04	Infinity	10.2
Tebuconazole	308.1524	39.36	174956	2158	1.8E-04	81.1	11.1	5268	0	1.7E-04	Infinity	10.7
fluoxetine	310.1413	23.71	2681	1848	7.7E-02	1.5	24.3	1920	0	3.3E-05	Infinity	6.2
Aflatoxin B1	313.07066	17.52	4683	284	9.8E-04	16.5	20.4	38	0	2.5E-02	Infinity	62.5
Progesterone	315.2339	42.10	69600	214479	1.4E-01	0.3	5.2	4237	189	1.4E-04	22.4	10.5
paroxetine	330.15	18.34	171803	1481	1.3E-04	116.0	9.7	5296	0	1.2E-04	Infinity	9.5
Propiconazole	342.0771	41.73	163040	134	1.2E-04	1217.6	9.5	4576	0	1.3E-04	Infinity	9.8
Boscalid	343.03994	38.00	65753	124	1.9E-04	531.5	11.2	1716	0	1.7E-04	Infinity	10.8
Chlorpyrifos	349.93356	45.53	9643	870	8.4E-05	11.1	9.6					
Cortisone	361.2006	16.12	98736	44932	8.9E-03	2.2	23.8	2867	1248	8.9E-03	2.3	24.3
hydrocortisone	363.2166	15.86	468688	308739	2.0E-03	1.5	9.7	14911	9674	2.2E-03	1.5	10.6
Prochloraz	376.0381	38.74						976	0	4.3E-04	Infinity	14.8
Solanidine	398.342	24.54	206530	1820	8.4E-05	113.5	8.4	5987	0	1.1E-04	Infinity	9.4
Azoxystrobine	404.1241	38.03	182890	507	1.4E-04	360.6	9.9	5265	0	1.3E-04	Infinity	9.9
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95	1658	18	1.3E-02	94.0	37.5					
2-phenylphenol	169.0659	30.19	209467	130364	1.3E-01	1.6	35.2	26177	23327	1.1E-02	1.1	2.9
Hydroxyindoleacetic acid	190.051	5.71	3110	795	7.3E-04	3.9	16.1					
Ibuprofen	205.1223	39.94	23267	3898	1.4E-06	6.0	4.9	539	494	4.9E-02	1.3	2.3
Diclofenac	294.0094	39.59						138	0	5.0E-04	Infinity	15.5
Arachidonic Acid	303.233	47.00	39363	5631	2.6E-04	7.0	9.2	132395	94574	5.7E-03	1.4	12.1
Leukotriene B4	335.2228	39.52	6206305	39347	2.9E-04	157.7	10.1	37259	30145	3.5E-03	1.2	6.8
Prostaglandin D2	351.2177	27.60						4366	3388	6.5E-02	1.3	21.9
Prostaglandin E2	351.2177	26.50						3012	1333	1.9E-02	2.3	32.0
Prostaglandin F2a	353.2333	25.60						3851	579	3.0E-05	6.7	10.1
Leukotriene D4	495.2534	33.04						324	0	2.5E-04	Infinity	12.3

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			Markerview - Noise 20					Markerview - Noise 50				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	847	0	1.5E-04	Infinity	10.3	568	0	7.7E-05	Infinity	8.3
Paracetamol	152.0706	4.98	1601	1761	3.2E-01	0.9	25.8					
Nicotine	163.12297	3.37										
Cotinine	177.1022	4.31	7947	6246	2.6E-03	1.3	3.7	6927	5229	2.1E-03	1.3	4.4
Carbendazim	192.07675	5.69	957	0	1.7E-04	Infinity	10.8	468	0	2.9E-02	Infinity	66.9
Cyprodinil	226.1339	33.22	5400	0	8.5E-05	Infinity	8.5	4645	0	1.3E-04	Infinity	9.8
Carbamazepine	237.10224	18.01	1813	0	1.7E-04	Infinity	10.8	1616	0	2.9E-04	Infinity	12.8
Clothianidin	250.016	7.99	8	0	2.0E-01	Infinity	200.0					
Thiacloprid	253.0309	12.24	655	0	6.5E-05	Infinity	7.8	443	0	8.9E-04	Infinity	18.8
Imidacloprid	256.0596	8.57	141	0	3.0E-02	Infinity	67.6					
Acetochlor	270.12553	40.57										
estrone	271.1693	31.60	250	0	2.0E-02	Infinity	57.4					
venlafaxine	278.2115	9.84	2427	0	5.5E-05	Infinity	7.4	1972	0	3.4E-05	Infinity	6.3
Piperine	286.1444	36.42	42053	20214	8.7E-05	2.1	9.1	40599	19154	1.2E-04	2.1	9.6
Androstenedione	287.20056	31.50	2323	0	2.7E-04	Infinity	12.6	1857	0	5.3E-04	Infinity	15.8
Testosterone	289.2168	28.90	2375	226	3.0E-02	10.5	61.8	2984	0	3.3E-04	Infinity	13.5
Thiamethoxam	292.0266	6.97	65	0	3.1E-03	Infinity	29.0					
Codeine	300.15942	5.12	427	0	7.9E-03	Infinity	21.3	371	0	4.5E-03	Infinity	33.0
Diazinon	305.1083	43.38	6405	0	1.0E-05	Infinity	4.2	5519	0	6.8E-06	Infinity	3.7
sertraline	306.0811	24.34	288	0	2.7E-03	Infinity	27.4					
Tebuconazole	308.1524	39.36	5026	0	1.6E-04	Infinity	10.6	4434	0	3.1E-04	Infinity	13.2
fluoxetine	310.1413	23.71	1733	0	7.7E-05	Infinity	8.3	1255	0	4.5E-04	Infinity	14.9
Aflatoxin B1	313.07066	17.52	4002	36	1.8E-04	111.0	11.8	3556	0	2.1E-04	Infinity	11.5
Progesterone	315.2339	42.10										
paroxetine	330.15	18.34	5110	0	1.5E-04	Infinity	10.4	4606	0	2.4E-04	Infinity	12.1
Propiconazole	342.0771	41.73	4305	0	1.2E-04	Infinity	9.6	3808	0	1.6E-04	Infinity	10.6
Boscalid	343.03994	38.00	1475	0	4.0E-04	Infinity	14.4	696	0	1.8E-02	Infinity	54.7
Chlorpyrifos	349.93356	45.53										
Cortisone	361.2006	16.12	2743	1136	6.6E-03	2.4	22.8	2417	934	9.5E-03	2.6	27.1
hydrocortisone	363.2166	15.86	14718	9522	2.3E-03	1.5	10.7	14319	9127	2.2E-03	1.6	10.8
Prochloraz	376.0381	38.74	794	0	1.0E-03	Infinity	19.6	49	0	2.0E-01	Infinity	200.0
Solanidine	398.342	24.54	5747	0	1.2E-04	Infinity	9.6	5187	0	1.9E-04	Infinity	11.2
Azoxystrobine	404.1241	38.03	4969	0	1.3E-04	Infinity	9.9	4175	0	4.2E-04	Infinity	14.6
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95										
2-phenylphenol	169.0659	30.19										
Hydroxyindoleacetic acid	190.051	5.71										
Ibuprofen	205.1223	39.94										
Diclofenac	294.0094	39.59	31	0	9.2E-02	Infinity	116.4					
Arachidonic Acid	303.233	47.00	131178	93675	5.7E-03	1.4	12.1	128771	91344	5.8E-03	1.4	12.3
Leukotriene B4	335.2228	39.52	205473	170313	1.9E-03	1.2	5.1	202795	168089	1.9E-03	1.2	5.1
Prostaglandin D2	351.2177	27.60	4056	3191	7.7E-02	1.3	22.6	3163	2630	1.8E-01	1.2	29.8
Prostaglandin E2	351.2177	26.50	2791	1264	2.2E-02	2.2	32.8	2595	1153	1.0E-02	2.3	25.9
Prostaglandin F2a	353.2333	25.60	3645	398	6.0E-05	9.2	9.2	3122	137	5.4E-07	22.7	5.6
Leukotriene D4	495.2534	33.04	250	0	9.3E-04	Infinity	19.1					

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			Markerview - Noise 100					MzMine - CWT pipeline - Default settings - Noise 10				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	409	0	8.5E-04	Infinity	18.6	46798	11100	1.7E-10	4.2	1.4
Paracetamol	152.0706	4.98										
Nicotine	163.12297	3.37										
Cotinine	177.1022	4.31	5561	4117	3.7E-03	1.4	4.9	276117	239480	2.2E-02	1.2	4.2
Carbendazim	192.07675	5.69	302	0	2.9E-02	Infinity	66.8	41332	1237	2.7E-06	33.4	2.6
Cyprodinil	226.1339	33.22	3593	0	4.6E-04	Infinity	15.1	245350	5732	2.7E-05	42.8	5.7
Carbamazepine	237.10224	18.01	1333	0	5.6E-04	Infinity	16.1	91642	1500	1.7E-04	61.1	10.7
Clothianidin	250.016	7.99										
Thiacloprid	253.0309	12.24	241	0	3.1E-04	Infinity	13.2	29841	2957	1.9E-05	10.1	4.7
Imidacloprid	256.0596	8.57						10173	1262	1.1E-06	8.1	2.7
Acetochlor	270.12553	40.57										
estrone	271.1693	31.60						973	147	5.0E-05	6.6	6.2
venlafaxine	278.2115	9.84	182	0	1.1E-01	Infinity	129.4	139721	12	1.0E-04	11515.6	9.1
Piperine	286.1444	36.42	422	0	1.8E-02	Infinity	54.8	4406047	1704098	3.2E-05	2.6	4.7
Androstenedione	287.20056	31.50	922	0	3.9E-02	Infinity	75.7	96238	8676	2.0E-06	11.1	2.4
Testosterone	289.2168	28.90	2511	0	6.1E-04	Infinity	16.5	138408	26293	1.6E-05	5.3	4.8
Thiamethoxam	292.0266	6.97						6365	1369	7.7E-06	4.6	4.8
Codeine	300.15942	5.12						109818	5720	6.0E-05	19.2	7.3
Diazinon	305.1083	43.38	4074	0	1.0E-04	Infinity	9.0	162555	167	1.1E-04	972.4	9.2
sertraline	306.0811	24.34						17251	909	2.2E-04	19.0	11.1
Tebuconazole	308.1524	39.36	3639	0	6.5E-04	Infinity	16.9	234650	4118	2.0E-05	57.0	5.2
fluoxetine	310.1413	23.71	186	0	1.2E-01	Infinity	137.5					
Aflatoxin B1	313.07066	17.52	2996	0	5.6E-04	Infinity	16.1					
Progesterone	315.2339	42.10						446253	116368	1.3E-04	3.8	8.6
paroxetine	330.15	18.34	3968	0	4.3E-04	Infinity	14.6	192911	1905	3.8E-05	101.3	6.5
Propiconazole	342.0771	41.73	2876	0	2.0E-03	Infinity	24.9	179728	1177	7.7E-05	152.7	8.2
Boscalid	343.03994	38.00	49	0	2.0E-01	Infinity	200.0	82166	804	1.0E-04	102.2	9.1
Chlorpyrifos	349.93356	45.53						10670	1046	2.5E-05	10.2	6.0
Cortisone	361.2006	16.12	2113	632	5.6E-03	3.3	25.9	156124	40550	1.7E-09	3.9	2.0
hydrocortisone	363.2166	15.86	13784	8678	2.1E-03	1.6	11.0	21337	5200	1.0E-04	4.1	7.0
Prochloraz	376.0381	38.74						39028	183	4.5E-05	213.5	6.9
Solanidine	398.342	24.54	4355	0	3.5E-04	Infinity	13.7	257646	2556	4.6E-05	100.8	6.9
Azoxystrobine	404.1241	38.03	2602	0	1.7E-02	Infinity	53.5	169010	2330	6.0E-05	72.5	7.5
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95						1313	0	1.2E-02	Infinity	46.8
2-phenylphenol	169.0659	30.19						2216	1048	3.2E-05	2.1	7.3
Hydroxyindoleacetic acid	190.051	5.71						5083	3841	3.5E-03	1.3	7.6
Ibuprofen	205.1223	39.94						23787	21018	2.3E-02	1.1	1.7
Diclofenac	294.0094	39.59										
Arachidonic Acid	303.233	47.00	124323	88239	6.0E-03	1.4	12.5	4419419	3166128	5.5E-03	1.4	11.7
Leukotriene B4	335.2228	39.52	197903	164449	1.9E-03	1.2	4.7					
Prostaglandin D2	351.2177	27.60	2441	1871	1.0E-01	1.3	29.2					
Prostaglandin E2	351.2177	26.50	1825	932	4.7E-03	2.0	18.9					
Prostaglandin F2a	353.2333	25.60										
Leukotriene D4	495.2534	33.04						17857	5283	5.6E-02	3.4	64.0

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			MzMine - CWT pipeline - Optimized settings - Noise 10					**MzMine - ADAP pipeline - Optimized settings - Noise 50**				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	40703	10406	1.4E-04	3.9	10.5	40816	11746	1.1E-03	3.5	5.8
Paracetamol	152.0706	4.98						67239	61337	4.3E-02	1.1	4.3
Nicotine	163.12297	3.37						5283	2339	1.5E-04	2.3	11.1
Cotinine	177.1022	4.31	297014	235009	6.3E-04	1.3	3.5	270705	214652	6.0E-04	1.3	5.4
Carbendazim	192.07675	5.69	46631	1376	5.4E-05	33.9	7.1	41481	1135	9.1E-05	36.5	8.5
Cyprodinil	226.1339	33.22	204458	4650	4.1E-05	44.0	6.6	199963	4372	5.7E-05	45.7	7.3
Carbamazepine	237.10224	18.01	72041	2886	6.3E-05	25.0	8.2	68529	1855	6.7E-05	36.9	8.6
Clothianidin	250.016	7.99	4505	695	2.2E-06	6.5	6.1	3667	542	2.2E-03	6.8	22.9
Thiacloprid	253.0309	12.24	28701	3619	2.1E-05	7.9	5.2	27179	376	3.3E-05	72.3	6.2
Imidacloprid	256.0596	8.57	9773	1427	2.5E-03	6.8	23.6	10684	1270	1.3E-05	8.4	6.9
Acetochlor	270.12553	40.57						6140	2956	6.9E-03	2.1	21.4
estrone	271.1693	31.60	653	197	1.8E-02	3.3	38.8	20921	1972	4.4E-04	10.6	13.6
venlafaxine	278.2115	9.84	92716	15	4.1E-05	6241.2	6.7	91129	11	3.1E-05	8596.6	6.1
Piperine	286.1444	36.42	3780083	1754841	6.7E-04	2.2	9.3	1562926	794563	5.6E-05	2.0	8.2
Androstenedione	287.20056	31.50	94800	9075	4.1E-04	10.4	13.1	96810	8791	1.7E-04	11.0	9.9
Testosterone	289.2168	28.90	138283	26205	1.9E-04	5.3	10.1	121077	14884	1.9E-04	8.1	10.2
Thiamethoxam	292.0266	6.97	7731	994	3.2E-05	7.8	8.9	7106	719	2.6E-04	9.9	13.8
Codeine	300.15942	5.12	107700	6095	3.3E-05	17.7	6.2	97785	3066	4.8E-05	31.9	6.8
Diazinon	305.1083	43.38	249838	394	6.8E-06	633.7	3.7	244116	244	7.0E-06	1000.1	3.7
sertraline	306.0811	24.34	20290	792	7.9E-06	25.6	4.0	18273	861	4.9E-04	21.2	14.7
Tebuconazole	308.1524	39.36	190027	4596	1.2E-04	41.4	9.2	186338	4098	1.4E-04	45.5	9.9
fluoxetine	310.1413	23.71						68304	2146	2.0E-05	31.8	5.2
Aflatoxin B1	313.07066	17.52	4484	463	3.7E-04	9.7	12.8	4025	574	4.1E-03	7.0	27.4
Progesterone	315.2339	42.10	378931	105247	3.6E-04	3.6	7.3	9805	5070	8.8E-05	1.9	8.4
paroxetine	330.15	18.34	178001	2346	1.3E-05	75.9	4.6	162493	1853	1.1E-05	87.7	4.4
Propiconazole	342.0771	41.73	168372	1023	1.0E-04	164.5	9.0	165249	853	1.3E-04	193.6	9.7
Boscalid	343.03994	38.00	67507	656	1.5E-04	103.0	10.2	66427	547	1.7E-04	121.4	10.8
Chlorpyrifos	349.93356	45.53	9590	1411	7.5E-08	6.8	4.5	9824	1536	3.9E-05	6.4	10.1
Cortisone	361.2006	16.12	107298	50275	6.5E-03	2.1	20.3	95632	42923	7.1E-03	2.2	21.7
hydrocortisone	363.2166	15.86	9805	7016	2.2E-01	1.4	40.7	491592	321024	2.5E-03	1.5	11.0
Prochloraz	376.0381	38.74	43944	169	5.8E-05	260.7	7.5	42666	164	7.6E-05	260.8	8.2
Solanidine	398.342	24.54	208957	2507	9.6E-05	83.4	8.9	196761	1569	1.1E-04	125.4	9.3
Azoxystrobine	404.1241	38.03	190862	2004	8.7E-05	95.2	8.6	185360	2002	1.1E-04	92.6	9.3
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95	1612		3.5E-06	Infinity	2.9	4024		5.7E-03	Infinity	35.9
2-phenylphenol	169.0659	30.19	2335	612	5.5E-08	3.8	2.1					
Hydroxyindoleacetic acid	190.051	5.71	2972	6167	1.2E-04	0.5	2.8	4599	3819	1.7E-02	1.2	5.2
Ibuprofen	205.1223	39.94	14918	16688	5.0E-03	0.9	4.7	22762	20063	2.0E-02	1.1	1.6
Diclofenac	294.0094	39.59						9333	772	1.2E-04	12.1	9.0
Arachidonic Acid	303.233	47.00	5517661	3945396	9.7E-05	1.4	3.1	4333359	3103838	6.1E-03	1.4	12.2
Leukotriene B4	335.2228	39.52						1904628	1594785	5.1E-03	1.2	5.5
Prostaglandin D2	351.2177	27.60	139935	76818	5.2E-04	1.8	9.8	107203	77593	1.8E-04	1.4	5.4
Prostaglandin E2	351.2177	26.50	10713	4947	6.3E-05	2.2	5.9	117304	94156	5.9E-04	1.2	4.3
Prostaglandin F2a	353.2333	25.60						163841	45785	1.5E-04	3.6	3.1
Leukotriene D4	495.2534	33.04	13385	10192	6.5E-04	1.3	5.8	19957	4147	5.3E-04	4.8	12.9

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			MzMine - ADAP pipeline - Optimized settings - Noise 100					**Progenesis - More sensitivity**				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74						102487	0	6.8E-05	Infinity	6.3
Paracetamol	152.0706	4.98	69118	61809	3.6E-02	1.1	4.5					
Nicotine	163.12297	3.37	5571	2688	1.6E-04	2.1	9.7	5691	1940	1.2E-02	2.9	26.8
Cotinine	177.1022	4.31	270697	214652	6.0E-04	1.3	5.4	905382	844804	2.2E-01	1.1	5.0
Carbendazim	192.07675	5.69	41481	1135	9.1E-05	36.5	8.5	117558	311	1.1E-04	377.7	7.4
Cyprodinil	226.1339	33.22	199963	4372	5.7E-05	45.7	7.3	683047	601	1.8E-04	1137.0	8.6
Carbamazepine	237.10224	18.01	68529	2604	7.6E-05	26.3	8.6	239090	496	2.0E-04	482.5	9.0
Clothianidin	250.016	7.99	4064	480	4.2E-05	8.5	8.7	4526	87	9.6E-03	52.0	33.5
Thiacloprid	253.0309	12.24	27162	376	3.4E-05	72.2	6.2	57978	0	2.2E-04	Infinity	9.3
Imidacloprid	256.0596	8.57	10475	775	2.0E-05	13.5	7.2	29370	19	8.7E-05	1583.7	6.8
Acetochlor	270.12553	40.57	6103	2956	7.3E-03	2.1	21.7					
estrone	271.1693	31.60	20919	1958	4.4E-04	10.7	13.6	4088	0	1.8E-01	Infinity	115.7
venlafaxine	278.2115	9.84	91126	8	3.1E-05	11065.0	6.1	294443	0	8.2E-05	Infinity	6.7
Piperine	286.1444	36.42	1562926	794563	5.6E-05	2.0	8.2	5942568	3562347	7.9E-04	1.7	8.9
Androstenedione	287.20056	31.50	96810	8836	1.7E-04	11.0	9.9	185469	1931	9.7E-04	96.1	15.3
Testosterone	289.2168	28.90	121077	15347	2.0E-04	7.9	10.2	427755	45085	3.6E-04	9.5	10.4
Thiamethoxam	292.0266	6.97	7102	698	2.4E-04	10.2	13.4	9031	0	3.6E-04	Infinity	10.9
Codeine	300.15942	5.12	97033	2932	3.7E-05	33.1	6.3	7768	5176	6.0E-02	1.5	6.8
Diazinon	305.1083	43.38	244116	244	7.0E-06	1000.1	3.7	758426	0	4.2E-05	Infinity	5.4
sertraline	306.0811	24.34	18273	861	4.9E-04	21.2	14.7	1563	0	9.3E-02	Infinity	82.0
Tebuconazole	308.1524	39.36	186338	4098	1.4E-04	45.5	9.9	639730	374	2.0E-04	1711.5	9.1
fluoxetine	310.1413	23.71	68304	2146	2.0E-05	31.8	5.2	160537	0	5.0E-02	Infinity	63.0
Aflatoxin B1	313.07066	17.52	3764	505	4.4E-03	7.5	28.3					
Progesterone	315.2339	42.10	147146	15555	1.2E-04	9.5	9.9	529065	28913	9.2E-05	18.3	10.4
paroxetine	330.15	18.34	162493	1853	1.1E-05	87.7	4.4	664387	842	2.3E-04	789.2	9.5
Propiconazole	342.0771	41.73	165249	853	1.3E-04	193.6	9.7	507268	0	3.6E-04	Infinity	11.0
Boscalid	343.03994	38.00	66427	547	1.7E-04	121.4	10.8	166437	82	1.3E-02	2033.9	37.4
Chlorpyrifos	349.93356	45.53	9821	1536	3.9E-05	6.4	10.1					
Cortisone	361.2006	16.12	95632	42923	7.1E-03	2.2	21.7	365823	182153	1.3E-02	2.0	20.0
hydrocortisone	363.2166	15.86	491556	321012	2.5E-03	1.5	11.0	1847637	1436477	1.2E-02	1.3	9.9
Prochloraz	376.0381	38.74	42662	164	7.7E-05	260.8	8.2	13918	0	1.7E-01	Infinity	111.3
Solanidine	398.342	24.54	196761	2085	1.1E-04	94.4	9.3	799204	2669	8.6E-05	299.4	6.8
Azoxystrobine	404.1241	38.03	185360	1642	1.1E-04	112.9	9.3	666771	0	2.4E-04	Infinity	9.6
Pravastatin	425.25337	20.50										
Dimethyldithiophosphate	156.95413	2.95	3929	0	5.6E-03	Infinity	35.8					
2-phenylphenol	169.0659	30.19						2636249	2585415	6.2E-01	1.0	2.9
Hydroxyindoleacetic acid	190.051	5.71	4642	3657	4.8E-03	1.3	3.9					
Ibuprofen	205.1223	39.94	22758	20041	2.1E-02	1.1	1.6	15374	13955	2.4E-01	1.1	9.0
Diclofenac	294.0094	39.59	9198	701	1.3E-04	13.1	9.2	2020	0	3.4E-02	Infinity	54.0
Arachidonic Acid	303.233	47.00	4333359	3103838	6.1E-03	1.4	12.2	16265510	12883305	6.6E-02	1.3	8.8
Leukotriene B4	335.2228	39.52	1904628	1594785	5.1E-03	1.2	5.5	8040115	7195749	3.3E-03	1.1	2.3
Prostaglandin D2	351.2177	27.60	107203	77593	1.8E-04	1.4	5.4					
Prostaglandin E2	351.2177	26.50	117304	94156	5.9E-04	1.2	4.3					
Prostaglandin F2a	353.2333	25.60	163841	45785	1.5E-04	3.6	3.1	628830	193899	2.6E-06	3.2	4.4
Leukotriene D4	495.2534	33.04	19957	4147	5.3E-04	4.8	12.9	14051	1549	7.3E-04	9.1	14.9

Table A5.2 – (continued) Results of data processing workflows on individual compounds in plasma

			Progenesis - Default sensitivity					Manual integration				
	m/z	Rt	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples	Average area in spiked samples	Average area in non-spiked samples	p-value (area in spiked vs. non-spiked samples)	Fold change (area in spiked/ area in non-spiked)	Area CV in spiked samples
AminoBenzimidazole	134.07127	4.74	102578	0	5.5E-05	Infinity	5.9	168885	557	1.0E-04	303.3	7.2
Paracetamol	152.0706	4.98						273862	238390	2.6E-02	1.1	3.7
Nicotine	163.12297	3.37						20083	6951	7.5E-04	2.9	14.7
Cotinine	177.1022	4.31						983674	790861	5.2E-03	1.2	4.1
Carbendazim	192.07675	5.69	2286	0	1.9E-04	Infinity	8.9	130835	306	1.2E-04	427.6	7.6
Cyprodinil	226.1339	33.22	680866	625	6.2E-05	1088.8	6.1	509013	499	1.5E-04	1021.1	8.2
Carbamazepine	237.10224	18.01	239099	510	1.0E-04	468.9	7.2	157472	0	1.9E-04	Infinity	8.8
Clothianidin	250.016	7.99	2880	1409	3.6E-01	2.0	65.6	8563	0	9.5E-05	Infinity	7.0
Thiacloprid	253.0309	12.24	57490	0	3.6E-04	Infinity	10.9	59923	0	1.2E-04	Infinity	7.5
Imidacloprid	256.0596	8.57	29385	20	4.5E-05	1501.1	5.5	23078	0	1.6E-04	Infinity	8.4
Acetochlor	270.12553	40.57						10563	0	7.4E-06	Infinity	3.0
estrone	271.1693	31.60	4009	0	1.8E-01	Infinity	113.6	38956	381	9.0E-04	102.2	14.8
venlafaxine	278.2115	9.84	294553	0	3.5E-05	Infinity	5.0	178786	0	3.0E-05	Infinity	4.8
Piperine	286.1444	36.42	5939776	3712990	4.6E-04	1.6	5.9	3437214	1680441	2.7E-04	2.0	9.6
Androstenedione	287.20056	31.50	186084	1984	1.3E-03	93.8	16.9	191923	13135	1.3E-04	14.6	8.8
Testosterone	289.2168	28.90	427530	47020	1.4E-04	9.1	7.7	244917	29170	4.3E-04	8.4	10.4
Thiamethoxam	292.0266	6.97						13063	0	1.1E-04	Infinity	7.5
Codeine	300.15942	5.12	311096	571	1.4E-05	544.5	3.7	191972	1094	7.1E-05	175.5	6.4
Diazinon	305.1083	43.38	760174	0	1.2E-04	Infinity	7.7	455855	0	5.6E-06	Infinity	2.7
sertraline	306.0811	24.34						38839	0	1.3E-04	Infinity	7.8
Tebuconazole	308.1524	39.36	639461	388	6.9E-05	1650.0	6.3	340486	1956	3.3E-04	174.1	10.6
fluoxetine	310.1413	23.71	159002	0	4.9E-02	Infinity	62.5	129336	0	3.6E-05	Infinity	5.1
Aflatoxin B1	313.07066	17.52						7761	0	3.8E-04	Infinity	11.2
Progesterone	315.2339	42.10	528710	29641	8.6E-06	17.8	7.6	261837	21243	3.2E-04	12.3	10.4
paroxetine	330.15	18.34	664026	875	8.1E-05	759.0	6.6	317360	982	2.1E-04	323.3	9.1
Propiconazole	342.0771	41.73	506899	0	1.6E-04	Infinity	8.3	269169	0	2.2E-04	Infinity	9.3
Boscalid	343.03994	38.00	165524	84	1.1E-02	1967.2	35.8	107469	0	3.5E-04	Infinity	10.8
Chlorpyrifos	349.93356	45.53						16680	2896	5.0E-04	5.8	12.2
Cortisone	361.2006	16.12	354122	187779	1.7E-02	1.9	20.0	161845	68332	8.2E-04	2.4	9.5
hydrocortisone	363.2166	15.86	1842377	1496669	1.1E-02	1.2	7.9	840557	533829	5.9E-03	1.6	11.9
Prochloraz	376.0381	38.74	4695	0	1.5E-01	Infinity	103.0	62878	0	1.6E-04	Infinity	8.3
Solanidine	398.342	24.54	799653	2779	4.7E-05	287.8	5.6	286540	2448	2.0E-04	117.0	9.0
Azoxystrobine	404.1241	38.03	666551	0	1.0E-04	Infinity	7.2	248938	0	1.5E-04	Infinity	8.2
Pravastatin	425.25337	20.50						6312	0	8.1E-04	Infinity	14.4
Dimethyldithiophosphate	156.95413	2.95						10242	0	1.7E-04	Infinity	8.6
2-phenylphenol	169.0659	30.19						2919579	2584915	2.5E-02	1.1	2.9
Hydroxyindoleacetic acid	190.051	5.71						51022	45338	4.4E-02	1.1	1.8
Ibuprofen	205.1223	39.94						62302	54968	5.0E-02	1.1	2.6
Diclofenac	294.0094	39.59						17233	0	2.4E-04	Infinity	9.6
Arachidonic Acid	303.233	47.00	16455687	13399467	1.2E-01	1.2	9.2	8909809	6276283	1.5E-02	1.4	13.6
Leukotriene B4	335.2228	39.52	8133285	7467473	3.0E-02	1.1	3.2	165375	72612	4.1E-04	2.3	9.1
Prostaglandin D2	351.2177	27.60						31655	10877	4.2E-06	2.9	5.9
Prostaglandin E2	351.2177	26.50						20553	7560	2.6E-05	2.7	1.6
Prostaglandin F2a	353.2333	25.60	636070	201102	5.5E-06	3.2	4.7	242325	22049	1.3E-06	11.0	5.0
Leukotriene D4	495.2534	33.04						17607	0	4.5E-04	Infinity	11.8

3.7. Table A5.3 – Summary of results of data processing workflows on individual compounds in plasma and serum

Table A5.3a – Summary of results of data processing workflows on individual compounds in plasma

Plasma	Noise threshold	Detection frequency (%)	Median p-value	Computing time	Median CV (spiked)	Compounds with CV < 30% (%)
XCMS DEF	10	64	3.98E-04	4	10	90
XCMS OPT	10	82	2.70E-04	3.5	10	84
	20	82	2.67E-04	3	10	86
	50	82	3.21E-04	3	10	86
	100	78	3.74E-04	3	10	86
Markerview	10	89	1.97E-04	0.5	10	90
	20	80	6.65E-04	0.5	11	83
	50	62	4.38E-04	0.5	13	86
	100	56	2.06E-03	0.5	17	72
Mzmine CWT	10	73	5.04E-05	14	7	94
MzMine CWT OPT	10	82	9.68E-05	25	7	95
MzMine ADAP	50	96	1.66E-04	18	9	98
	100	93	1.52E-04	17	9	98
Progenesis	More sensitivity	80	7.63E-04	1.5	9	81
	Default sensitivity	62	1.47E-04	1	8	82

Table A5.3b – Summary of results of data processing workflows on individual compounds in serum

Serum	Noise threshold	Detection frequency (%)	Median p-value	Computing time	Median CV (spiked)	Compounds with CV < 30% (%)
XCMS DEF	10	60	1.26E-03	4.5	16	74
XCMS OPT	10	71	4.68E-03	4	18	81
	20	69	2.07E-03	3	14	84
	50	64	2.05E-03	3	13	86
	100	64	4.16E-03	3	18	76
Markerview	10	82	8.62E-04	0.5	15	86
	20	80	1.30E-03	0.5	18	75
	50	69	2.63E-03	0.5	21	65
	100	49	3.13E-03	0.5	24	68
Mzmine CWT	10	78	3.13E-03	12	17	83
MzMine CWT OPT	10	84	7.61E-04	14	8	100
MzMine ADAP	50	87	9.19E-04	18	17	79
	100	82	9.19E-04	18	17	78
Progenesis	More sensitivity	78	1.10E-03	1.5	14	83
	Default sensitivity	67	5.65E-04	1	11	90

3.8. Table A6.1 – Results of annotation after manual curation in serum

Table A6.1 – Results of annotation after manual curation in serum

Annotation	SMILES	CI m/z		CI Rt						CI isotopic fit		Global CI	
				Experimental		RTI-predicted		logP-predicted		CI overall			
		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
MEHP*	CCCCC(CC)COC(=O)C1=CC=CC=C1C(=O)O	0.86				0.83		0.89		0.95		G3_0.88	
Acesulfame	CC1=CC(=O)NS(=O)(=O)O1		0.93		**0.97**		0.66						G2_0.95
Alpha-tocopherol	CC1=C(C2=C(CCC(O2)(C)CCCC(C)CCCC(C)CCCC(C)C)C(=C1O)C)C	0.93				**0.90**						G2_0.91	
Eicosapentaenoic acid	CCC=CCC=CCC=CCC=CCC=CCCCC(=O)O	0.97	0.91			**0.90**	**0.95**	0.65	0.71		0.57	G2_0.94	G3_0.81
Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	0.82		**0.96**		0.43		0.54		0.50		G3_0.76	
Tryptophan	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N	0.95	0.98	**0.71**	**0.69**					0.83	0.81	G3_0.83	G3_0.83
4-indolecarbaldehyde	C1=CC(=C2C=CNC2=C1)C=O	0.86	0.93			0.89	0.89	0.85	0.85	0.63		G3_0.79	G2_0.91
Indoxyl sulfate	C1=CC=C2C(=C1)C(=CN2)OS(=O)(=O)O		0.94				0.81		0.77		0.80		G3_0.85
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O		0.91		**1.00**				0.91				G2_0.96
Mesterolone	CC1CC(=O)CC2C1(C3CCC4(C(C3CC2)CCC4O)C)C	0.92				**0.87**		1.00		0.87		G3_0.93	
Paracetamol	CC(=O)NC1=CC=C(C=C1)O	0.94	0.96	**0.77**	**0.80**	0.93	0.94	0.94	0.94			G2_0.85	G2_0.76
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	0.93		**0.99**		0.93		0.61		0.63		G3_0.85	
Paraxanthine	CN1C=NC2=C1C(=O)N(C(=O)N2)C	0.94	0.96			**0.79**	**0.41**			0.39		G3_0.71	G2_0.69
Theobromine	CN1C=NC2=C1C(=O)NC(=O)N2C	0.94				**0.86**				0.39		G3_0.73	
Theophylline	CN1C2=C(C(=O)N(C1=O)C)NC=N2	0.94				**0.84**				0.39		G3_0.73	
Coumaric acid	C1=CC(=CC(=C1)O)C=CC(=O)O	0.92	0.98			**0.83**	**0.83**	0.83	0.83			G2_0.87	G2_0.91
Cannabidiol	CCCCCC1=CC(=C(C(=C1)O)C2C=C(CCC2C(=C)C)C)O	0.82	0.96			**0.85**	**0.92**	0.81	0.90			G2_0.83	G2_0.94
Δ9-THC*	CCCCCC1=CC(=C2C3C=C(CCC3C(OC2=C1)(C)C)C)O	0.82	0.96			**0.97**	**0.95**	0.56	0.58			G2_0.89	G2_0.96
Cotinine	CN1C(CCC1=O)C2=CN=CC=C2	0.95		**0.99**		1.00		1.00				G2_0.97	
3-hydroxycotinine	CN1C(CC(C1=O)O)C2=CN=CC=C2	1.00						0.98				G2_0.99	
Allopregnanolone	CC(=O)C1CCC2C1(CCC3C2CCC4C3(CCC(C4)O)C)C	0.66	0.92	**0.94**	**0.98**	0.89	0.94	0.98	0.92	0.37		G3_0.66	G2_0.99
Androstanediol	CC12CCC(CC1CCC3C2CCC4(C3CCC4O)C)O	0.92				**0.65**		-0.09		0.82		G3_0.8	
Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	0.81	0.90	**0.98**	**0.99**	0.87	0.83	0.70	0.76			G2_0.89	G2_0.94
Arachidonic acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	0.92	0.92	**1.00**	**0.72**	1.00	0.67			0.87		G3_0.93	G2_0.82
Cortisol	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	0.91	0.98	**1.00**	**0.99**	0.70	0.70	0.13	0.14	0.78		G3_0.9	G2_0.98
Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)CO)O)C	0.82	0.98	**1.00**	**0.97**	0.68	0.70	1.00	0.97			G2_0.91	G2_0.97
DHA*	CCC=CCC=CCC=CCC=CCC=CCC=CCCC(=O)O	0.90	0.94			**0.92**	**0.92**	0.63	0.63		0.88	G2_0.95	G3_0.94
Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	0.73	0.99	**0.99**	**0.97**	0.98	0.97	0.99	0.92		0.61	G2_0.86	G3_0.86
Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N		0.73		**0.92**		0.45		0.18				G2_0.82
Progesterone	CC(=O)C1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	0.82	0.93	**0.97**	**0.99**	0.86	0.91	0.69	0.76			G2_0.89	G2_0.96
Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	0.61	0.94	**0.99**	**0.98**	0.89	0.88	0.92	0.90	0.46		G3_0.69	G2_0.96

*

Δ9-THC	Delta9-tetrahydrocannabinol
DHA	Docosahexaenoic acid
MEHP	2-(2-ethylhexoxycarbonyl)benzoic acid

Table A6.1 – (continued) Results of annotation after manual curation in serum

Annotation	MS/MS				Confidence level
	Theoretical fragments		Experimental fragments		
	(+)	(-)	(+)	(-)	
MEHP*	57.0699, 121.0284, 149.0239, 184.0731		57.0701, 121.0289, 149.023, 184.0741		2a
Acesulfame		77.9657, 82.0302		77.9660, 82.0300	1
Alpha-tocopherol	137.0981, 169.0922, 205.1194		137.0967, 169.0915, 205.1221		2a
Eicosapentaenoic acid	91.0534, 105.0703	149.1340, 203.1782, 257.2254	91.0543, 105.0704	149.1333, 203.1790, 257.2261	2a
Piperine	135.0450, 143.0499, 201.0551		135.0441, 143.0491, 201.0554		1
Tryptophan	130.0652, 142.0652, 170.0601, 188.0706	116.0507, 142.0651	130.0646, 142.0651, 170.0601, 188.0706	116.0509, 142.0657	1
4-indolecarbaldehyde	91.0553, 118.0669, 128.0614	90.0351, 116.0506	91.0547, 118.0675, 128.0621	90.0355, 116.0506	2b
Indoxyl sulfate		80.9665, 132.0460		80.9662, 132.0453	2a
Ibuprofen		154.9716, 161.1332		154.9722, 161.1333	1
Mesterolone	187.1486, 269.2269, 287.2364		187.1487, 269.2263, 287.2375		2a
Paracetamol	110.0598, 134.0587	107.0375	110.0600, 134.0588	107.0386	1
Caffeine	110.0715, 138.0659		110.0718, 138.0664		1
Paraxanthine	96.0572, 124.0522	108.0198, 122.0365, 164.0341	96.0572, 124.0515	108.0208, 122.0360, 164.0344	2a
Theobromine	108.0554, 122.0589, 163.0611		108.0559, 122.0590, 163.0618		2a
Theophylline	124.0497		124.0501		2a
Coumaric acid	91.0538, 119.0486, 147.0431	93.0349, 119.0505	91.0542, 119.0492, 147.0447	93.0343, 119.0498	2a
Cannabidiol	193.1223, 259.1686	179.1066, 229.1228, 245.1541	193.1229, 259.1684	179.1066, 229.1218, 245.1534	2a
Δ9-THC*	109.0648, 121.1012, 131.0856, 297.2214	191.1050, 245.1521	109.0648, 121.1019, 131.0861, 297.2205	191.1055, 245.1527	2a
Cotinine	118.0649, 146.0588		118.0656, 146.0592		1
3-hydroxycotinine	119.0603, 175.0665		119.0604, 175.0668		2a
Allopregnanolone	263.2007, 271.2058, 275.2009, 287.2371	297.1529, 311.1687, 325.1842	263.1996, 271.2058, 275.2010, 287.2371	297.1519, 311.1690, 325.1847	1
Androstanediol	109.0648, 121.1012, 131.0856, 297.2214		109.0648, 121.1019, 131.0861, 297.2205		2a
Androstenedione	173.1310, 211.1451, 269.1910	183.1128	173.1319, 211.1446, 269.1916	183.1126	1
Arachidonic acid	121.1025, 221.1559, 269.2300, 287.2397	205.1965, 231.2106, 259.2419	121.1021, 221.1550, 269.2289, 287.2389	205.1970, 231.2115, 259.2428	1
Cortisol	121.0647, 309.1858	297.1497, 315.1616, 331.1910	121.0651, 309.1859	297.1503, 315.1606, 331.1917	1
Cortisone	163.1115, 267.1729	301.1795, 329.1750	163.1125, 267.1728	301.1801, 329.1757	1
DHA*	119.08556, 159,1176, 173.1326, 329.25	229.1958, 284.2446	119.0863, 159.1169, 173.1334, 329.2482	229.1956, 284.2439	2a
Leukotriene B4	149.0966, 259.2066	71.0136, 195.1011, 317.2125	149.0968, 259.2072	71.0134, 195.1021, 317.2129	1
Leukotriene D4		177.0334, 477.2423		177.0329, 477.2431	1
Progesterone	109.0652, 123.0804, 297.2214	255.2323, 311.1689	109.0650, 123.0809, 297.2229	255.2321, 311.1680	1
Testosterone	253.1946, 271.2054	283.2640, 297.1529	253.1951, 271.2056	283.2642, 297.1525	1

*

Δ9-THC Delta9-tetrahydrocannabinol
DHA Docosahexaenoic acid
MEHP 2-(2-ethylhexoxycarbonyl)benzoic acid

3.9. Table A6.2 – Results of annotation after manual curation in plasma

Table A6.2 – Results of annotation after manual curation in plasma

Annotation	SMILES	CI m/z		CI Rt						CI isotopic fit		Global CI	
				Experimental		RTI-predicted		logP-predicted		CI overall			
		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
TDMPAB*	CC(C)(C)C(=O)NC1=CC(=CC(=C1)NC(=O)C(C)(C)C)NC(=O)C(C)(C)C	0.94	0.93			**0.79**	**0.79**	-0.22	-0.22	0.81		G3_0.85	G2_0.86
2-naphthylamine	C1=CC=C2C=C(C=CC2=C1)N	0.97	0.94			**0.64**	**0.85**	0.64	0.84			G2_0.8	G2_0.89
Bisphenol F	C1=CC(=CC=C1CC2=CC=C(C=C2)O)O		0.90				**0.98**		0.90				G2_0.94
Butylparaben	CCCCOC(=O)C1=CC=C(C=C1)O	0.90	0.93			**0.98**	**0.99**	0.83	0.82			G2_0.94	G2_0.96
Ethyl paraben	CCOC(=O)C1=CC=C(C=C1)O	0.82	0.89			**0.93**	**0.96**	0.75	0.87			G2_0.87	G2_0.44
Propylparaben	CCCOC(=O)C1=CC=C(C=C1)O	0.99		**0.96**		0.76		0.80		0.61		G3_0.85	
4-hydroxybenzoic acid	C1=CC(=CC=C1C(=O)O)O	0.88	0.95			**0.85**	**0.84**	0.84	0.84			G2_0.87	G2_0.9
TCPP*	CC(CCl)OP(=O)(OC(C)CCl)OC(C)CCl	0.67	0.81			**0.90**	**0.72**	0.67	0.72	0.92		G3_0.83	G2_0.76
Acesulfame	CC1=CC(=O)NS(=O)(=O)O1		0.98		**0.99**		0.66						G2_0.99
Caffeic acid	C1=CC(=C(C=C1C=CC(=O)O)O)O	0.86	0.94			**0.85**	**0.84**	0.83	0.83			G2_0.86	G2_0.89
Coumaric acid	C1=CC(=CC(=C1)O)C=CC(=O)O	1.00	0.98			**0.82**	**0.82**	0.82	0.82			G2_0.91	G2_0.9
Tryptophan	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N	0.95	0.93	**0.68**	**0.73**					0.60	0.61	G3_0.74	G3_0.76
4-indolecarbaldehyde	C1=CC(=C2C=CNC2=C1)C=O	0.95	0.93			**0.72**	**0.89**	0.66	0.85			G2_0.83	G2_0.91
Chlortalidone	C1=CC=C2C(=C1)C(=O)NC2(C3=CC(=C(C=C3)Cl)S(=O)(=O)N)O		0.82				**0.93**		0.84		0.71		G3_0.82
Hydrochlorothiazide	C1NC2=CC(=C(C=C2S(=O)(=O)N1)S(=O)(=O)N)Cl		0.85				**0.76**				0.92		G3_0.84
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O		0.99		**1.00**				0.90				G2_1
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	1.00		**0.95**		0.94		0.60				G2_0.97	
Paraxanthine	CN1C=NC2=C1C(=O)N(C(=O)N2)C	0.94	0.96			**0.78**	**0.39**					G2_0.86	G2_0.68
Theobromine	CN1C=NC2=C1C(=O)NC(=O)N2C	0.94				**0.85**						G2_0.9	
Allopregnanolone	CC(=O)C1CCC2C1(CCC3C2CCC4C3(CCC(C4)O)C)C	0.97	0.92	**0.89**	**0.98**	0.83	0.94	0.95	0.92	0.87		G3_0.91	G2_0.95
Androstenedione	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4=O)C	0.81	0.84	**0.90**	**0.92**	0.79	0.80	0.99	0.97			G2_0.85	G2_0.88
Arachidonic acid	CCCCC/C=C\C\C=C/C/C=C\C\C=C/CCCC(O)=O	0.89	0.84	**1.00**	**1.00**	1.00	1.00				0.80	G2_0.85	G3_0.88
Cortisol	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	0.83	0.92	**1.00**	**0.99**	0.70	0.70	0.13	0.14			G2_0.91	G2_0.96
Cortisone	CC12CCC(=O)C=C1CCC3C2C(=O)CC4(C3CCC4(C(=O)CO)O)C	0.91	0.84	**1.00**	**1.00**	0.67	0.67	0.99	0.99			G2_0.85	G2_0.92
DHA*	CCC=CCC=CCC=CCC=CCC=CCC=CCCC(=O)O	0.78	0.91			**0.92**	**0.92**	0.63	0.63	0.66	0.88	G3_0.81	G3_0.81
Leukotriene B4	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O		0.89		**0.97**		0.97		0.92		0.57		G3_0.81
Leukotriene D4	CCCCCC=CCC=CC=CC=CC(C(CCCC(=O)O)O)SCC(C(=O)NCC(=O)O)N		0.93		**1.00**		0.00		-0.86				G2_0.96

*

DHA	Docosahexaenoic acid
TCPP	Tris(1-chloro-2-propyl)phosphate
TDMPAB	1,3,5-tris(2,2-dimethylpropionylamino)benzene

Table A6.2 – (continued) Results of annotation after manual curation in plasma

Annotation	MS/MS				Confidence level
	Theoretical fragments		Experimental fragments		
	(+)	(-)	(+)	(-)	
TDMPAB*	191.1178, 275.1754, 292.2020	206.1299, 290.1874, 316.1667	191.1175, 275.1758, 292.2025	206.1297, 290.1875, 316.1666	2b
2-naphthylamine	77.0386, 117.0704, 127.0541	101.0391, 116.0500	77.0387, 117.0704, 127.0541	101.0391, 116.0498	2a
Bisphenol F		157.0649, 171.0815		157.0645, 171.0446	2a
Butylparaben	121.0290, 139.0395, 177.0916	121.0306, 137.0239	121.0287, 139.0390, 177.0913	121.0303, 137.0238	2a
Ethyl paraben	121.0290, 149.0603	136.0158, 137.0244	121.0285, 149.0598	136.0157, 137.0240	2a
Propylparaben	121.0283, 139.0384		121.0285, 139.0390		1
4-hydroxybenzoic acid	95.0488, 121.0293	93.0340, 119.0133	95.0491, 121.0291	93.0338, 119.0132	2a
TCPP*	174.9909, 250.9995	159.0645, 256.0313	174.9913, 251.0002	159.0644, 256.0311	2a
Acesulfame		77.9657, 82.0302		77.9653, 82.0305	1
Caffeic acid	135.0436, 163.0389	89.0391, 108.0230, 135.0429	135.0439, 163.0387	89.0388, 108.0229, 135.0426	2a
Coumaric acid	91.0538, 103.0540, 147.0431	93.0349, 119.0505	91.0542, 103.0544, 147.0447	93.0347, 119.0501	2a
Tryptophan	118.0644, 130.0653, 146.0597, 170.0575	116.0507, 142.0651	118.0647, 130.0658, 146.0598, 170.0578	116.0501, 142.0647	1
4-indolecarbaldehyde	91.0553, 118.0669, 128.0614	90.0351, 116.0506	91.0546, 118.0664, 128.0620	90.0351, 116.0506	2b
Chlortalidone		146.0247, 189.9739		146.0247, 189.9733	2a
Hydrochlorothiazide		126.0118, 204.9835, 268.9465		126.0111, 204.9837, 268.9456	2a
Ibuprofen		154.9722, 161.1333		154.9720, 161.1332	1
Caffeine	135.0436, 163.0389		135.0439, 163.0387		1
Paraxanthine	96.0572, 124.0522	108.0198, 122.0365, 164.0341	96.0579, 124.0525	108.0197, 122.0363, 164.0340	2a
Theobromine	108.0554, 122.0589, 163.0611		108.0551, 122.0580, 163.0612		2a
Allopregnanolone	263.2007, 271.2058, 275.2009, 287.2371	297.1529, 311.1687, 325.1842	263.2005, 271.2057, 275.2004, 287.2368	297.1527, 311.1688, 325.1840	1
Androstenedione	173.1310, 211.1451, 269.1910	183.1128	173.1308, 211.1449, 269.1908	183.1125	1
Arachidonic acid	121.1025, 221.1559, 269.2300, 287.2397	205.1965, 231.2106, 259.2419	121.1021, 221.1555, 269.2296, 287.2395	205.1964, 231.2102, 259.2419	1
Cortisol	121.0647, 309.1858	297.1497, 315.1616, 331.1910	121.0644, 309.1857	297.1494, 315.1614, 331.1905	1
Cortisone	163.1115, 267.1729	301.1795, 329.1750	163.1113, 267.1727	301.1797, 329.1749	1
DHA*	119.08556, 159,1176, 173.1326, 329.25	229.1958, 284.2446	119.08554, 159,1173, 173.1325, 329.2499	229.1957, 284.2445	2a
Leukotriene B4		71.0136, 195.1011, 317.2125		71.0136, 195.1010, 317.2126	1
Leukotriene D4		177.0334, 477.2423		177.0332, 477.2421	1

*

DHA Docosahexaenoic acid

TCPP Tris(1-chloro-2-propyl)phosphate

TDMPAB 1,3,5-tris(2,2-dimethylpropionylamino)benzene

3.10. Appendix S.1 – Chemicals and solvents

Standard compounds (native and isotopically labeled) were purchased from LGC, VWR, Sigma Aldrich, or Bertin and were stored at -20°C. Details can be found in Supporting Information (SI, Table A1). UHPLC-MS-grade acetonitrile and formic acid were purchased from Biosolve (Dieuze, France) while UHPLC-MS-grade methanol was purchased from Carlo Erba (Val-de-Reuil, France). Ultrapure water was obtained with a Millipore Milli-Q Gradient system.

3.11. Appendix S.2 – Data acquisition

Samples were analyzed on AB SCIEX X500R QTOF interfaced with an AB SCIEX ExionLC AD UPLC. Compound chromatographic separation was achieved using an Acquity UPLC HSS T3 C18 column (1.8µm, 1.0 x150mm) maintained at 40°C. Injection volume was set at 2 µL. Flow rate was set at 100 µL/min with mobile phases of ultrapure water/0.01% formic acid (A) and acetonitrile/0.01% formic acid (B). The gradient was as follows: 0-2.5 min, 10-20% B ; 2.5-20 min, 20-30% B ; 20-38 min, 30-45% B ; 38-45 min, 45-100% B ; 45-55 min, 100% B ; 55-60 min, 10% B. Full-scan mass spectra was acquired between 50-1100 m/z using ESI source settings: temperature 550°C, ionspray voltage 4,5kV (-4,5kV in negative mode), declustering potential 80V (-80V in negative mode), accumulation time 300 ms, spray N2 gas 35 arbitrary units, heat conduction gas 35 arbitrary units; curtain gas 7 arbitrary units, collisionally activated dissociation gas 7 arbitrary units, run time 60min. Samples were analyzed in full scan experiment in both – and + ESI modes. MS/MS mass fragmentation information for chemical elucidation was obtained by further analysis of selected samples in sequential window acquisition of theoretical mass spectrum (SWATH).

3.12. Appendix S.3 – Quality control

One workup sample (i.e. extraction with HPLC grade water instead of sample) per analytical batch was prepared to monitor background contaminants. Quality control (QC) samples comprising a composite sample were prepared in order to monitor for UHPLC-ESI-TOFMS repeatability and sensitivity during analysis of a sample run. Solvent samples (acetonitrile/H2O (10:90)) were also injected to ensure that there was no carryover in the UHPLC system that might affect adjacent results in analytical runs. Each run commenced with the injection of blank samples (workup and solvent) followed by injection of a QC sample. The samples were injected randomly with QC samples analyzed after every 7 samples.

3.13. Appendix S.4 – In-house annotation workflow

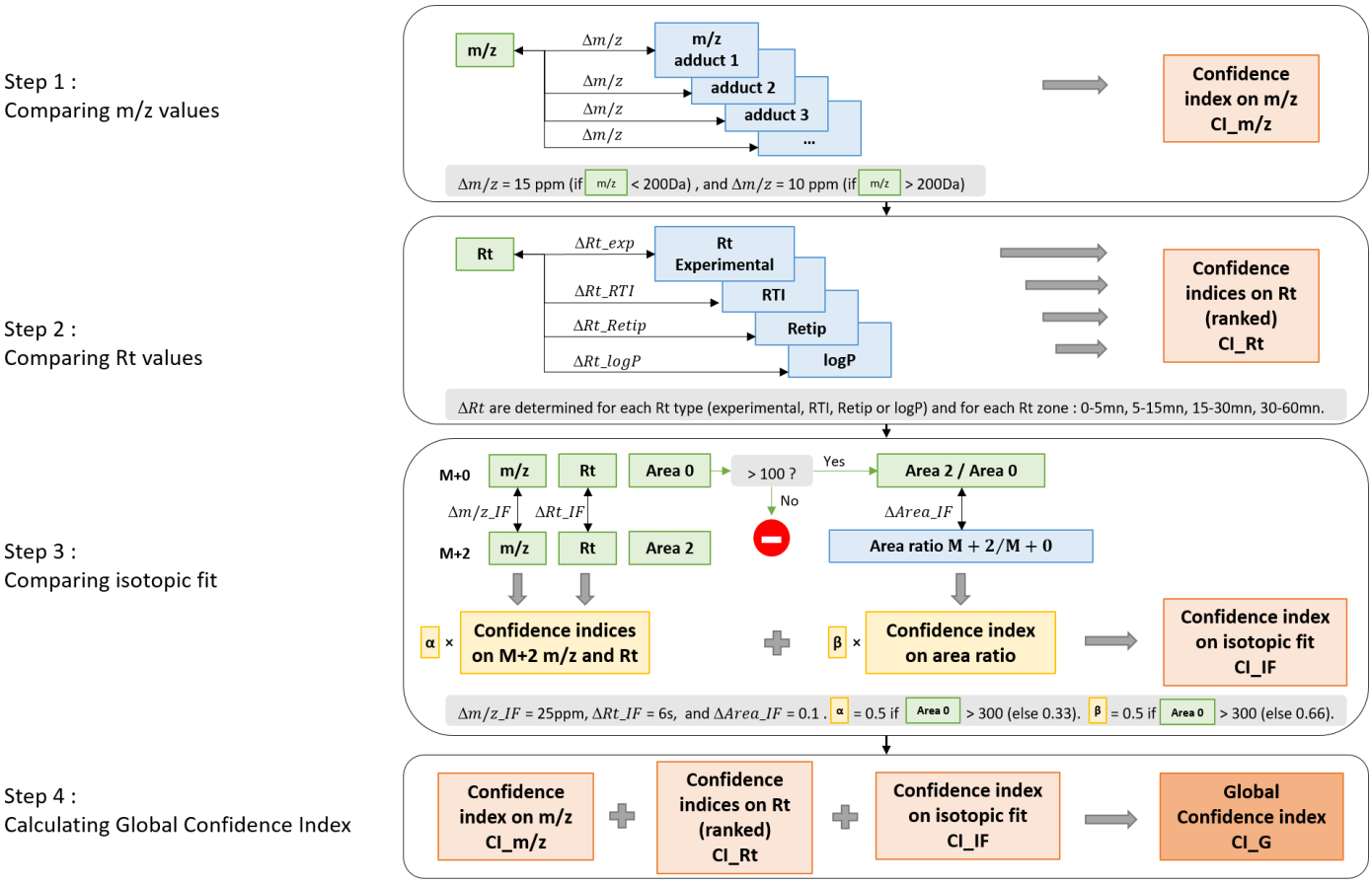

Figure B1 - In-house annotation workflow in four steps: comparing successively m/z, Rt and isotopic fit, then generating a global scoring. Calculation of CI for m/z and Rt is fairly simple and is only based on a comparison between the feature's and suspect's predictors. For isotopic fit, a multi-step approach is needed, involving first a detection of M+2, then a ratio abundance comparison. More specifically, as a first step for a given pre-annotated feature, the software computes a temporary CI based on m/z for the M+2. Then, another temporary CI is computed based on M_0 and M_2 Rts proximity. The two temporary CI values are averaged to give an intermediate M+2 identification CI. The second step is the M+0 abundance check. Since data processing software might generate less accurate integrations for low-abundant compounds, abundance ratios are only compared if the pre-annotated feature's area is higher than a threshold value of 100 (linked to the experimental data). If this is not verified, only the intermediate M+2 identification CI is displayed. Else, the area ratios are compared and a second intermediate CI is computed for abundances with a $\Delta_{A2/A0}$ value of 0.1. The last step is to compute an overall CI value for isotopic fit, which is calculated as a weighed sum of the intermediate CIs for M+2 identification and abundance, with a ponderation linked to the M+0 area. Finally, global CI is computed as a mean of the three predictors' CI.

3.14. Appendix S.5 – Modelling the retention time predictor

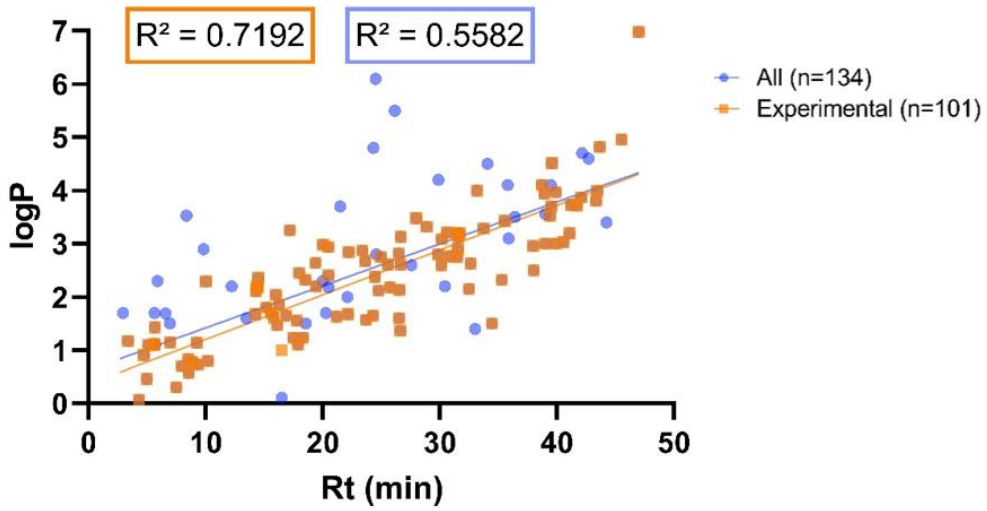

Figure B2 – Modelling retention time using modelled (n=134) or exclusively experimental (n=101) octanol-water partition coefficients as predictors.

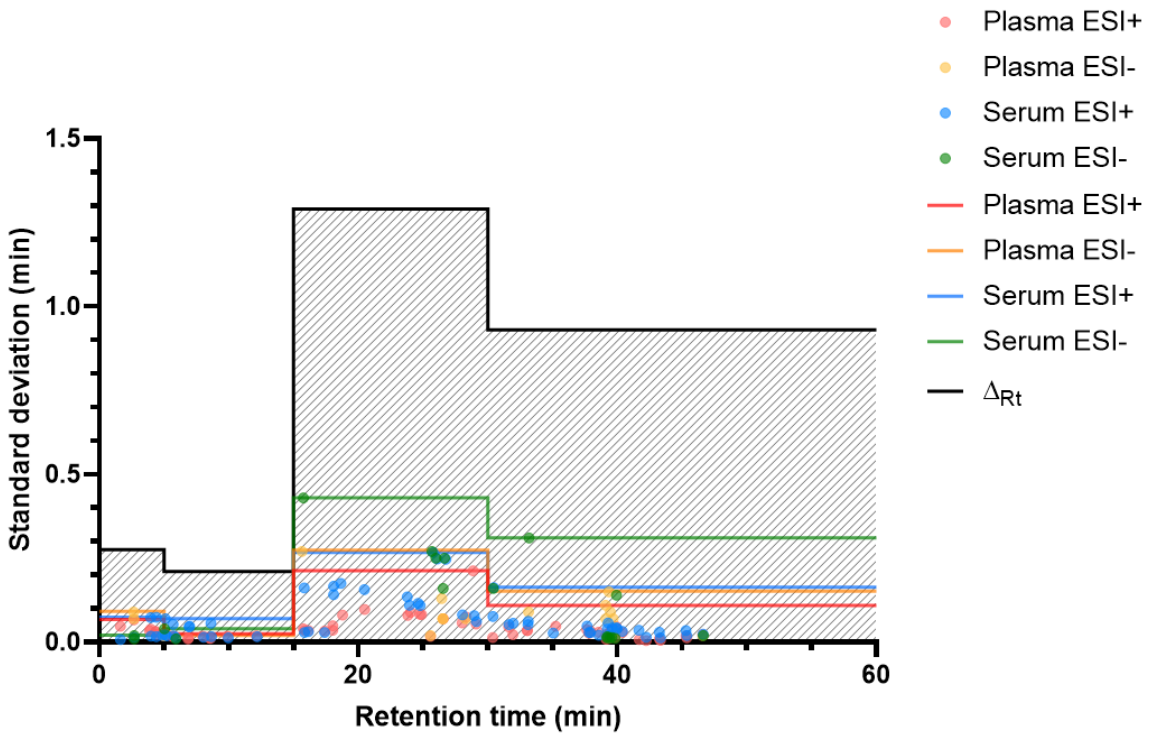

Figure B3 - Determination of experimental retention time tolerance Δ_{Rt}. Standard deviation on Rt were determined for compounds from the spiking set on the four spiked plasma and four spiked serum samples, and for internal standards (ISTD) on the eight plasma and eight serum samples. This value was multiplied by three as to theoretically obtain 99.7% of values under the curve.

3.15. Appendix S.6 – Optimization of individual data processing tools

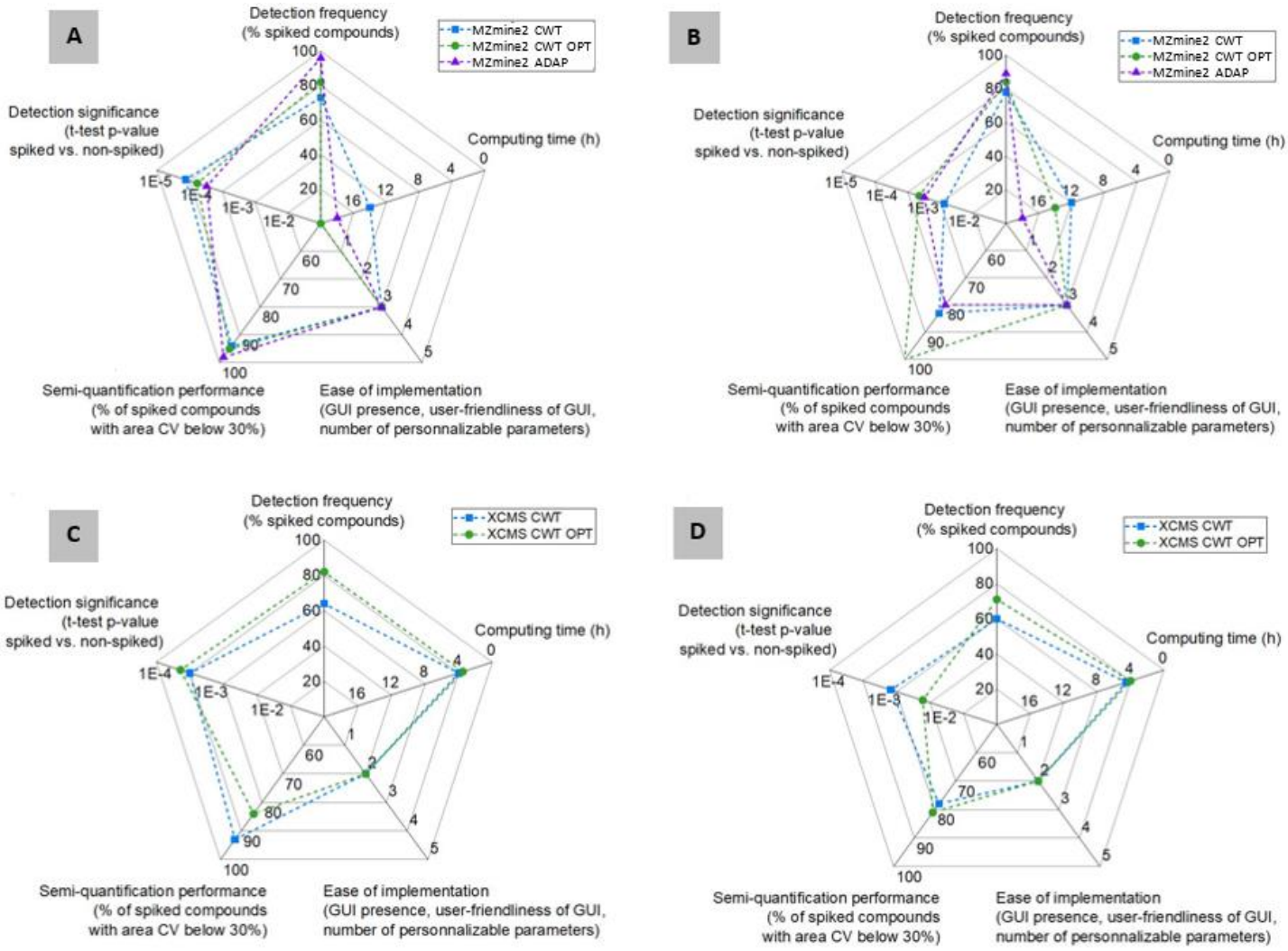

Figure B4 - Data processing (i.e. peak picking, deconvolution, alignment, gap filling) evaluation for detection and semi-quantification of environmental-level spiked compounds with different sets of parameters using MZmine2 in plasma (A) and serum (B), and using XCMS in plasma (C) and serum (D) (n=4 samples each).

3.16. Appendix S.7 – Application of the in-house software in real-life conditions

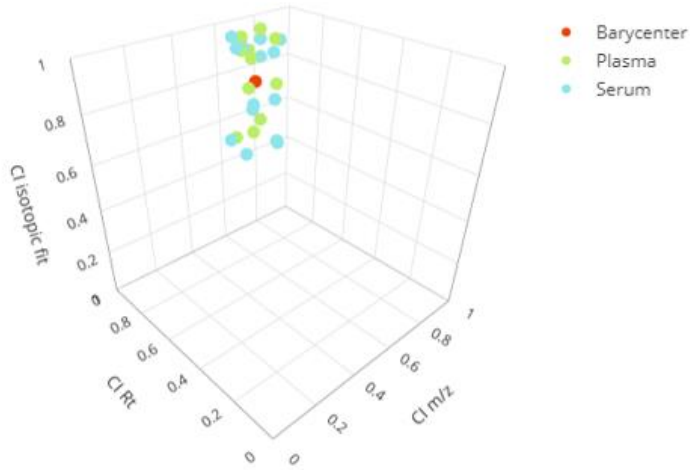

Figure B5 - Projection of spiking compounds for which all three CI were available in plasma and serum, as well as a barycenter.

	In-house tool	MZmine2	msPurity	MS-DIAL	xMSannotator
Using in-house libraries	5 (4+1)	5 (4+1)	4 (4+0)	4 (4+0)	5 (4+1)
Using existing databases	0	5	4	4	5
Using experimental and/or predicted Rt	5 (2+1+2)	3 (2+1+0)	0	3 (2+1+0)	0
Using MS/MS	0	3 (3+0)	5 (3+2)	3 (3+0)	0
Speed of implementation	5 (2+3)	4 (2+2)	3 (1+2)	4 (2+2)	2 (1+1)
Scoring	5 (2+3)	0	5 (2+3)	2 (2+0)	3 (2+1)
Prioritization of spiked chemicals	5 (2+2+1)	2 (2+0+0)	4 (2+1+1)	5 (2+2+1)	3 (2+1+0)

Figure B6 - Scoring of five suspect screening tools: xMSannotator, MS-DIAL, msPurity (green), MZmine2 and in-house tool. Comparison was made on use of in-house databases, use of predicted or experimental Rt and MS/MS, speed of implementation, scoring and prioritization. **Use of in-house libraries** was rated based on availability (/4), with a bonus given to tools which allow the use of an easily formatted database such as .csv (/1). **Use of existing databases** was rated based on availability of none, one to three, or more external database (0/5, 4/5, or 5/5). **Use of experimental and/or predicted Rt** was rated based on availability (/2), use of experimental Rt only through in-house library (/1), and use of experimental and predicted Rt (/2). **Use of MS/MS** was rated based on availability (/3), and scoring on this predictor (/2). **Speed of implementation** considers ease of set up (/2) and computational speed (/3). **Scoring** is rated based on availability (/2), and basis of said score on within-dataset correlation or on correlation with the suspect list (/3). Lastly, **prioritization of spiked chemicals** is rated based on availability of criteria for prioritization (e.g. detection frequency, or scoring, etc.) (/2), usability of scoring (if available) to estimate fit between suspect and feature (/2), and efficiency of ranking (/1).

4. Appendix 4. Chapter V

4.1. Effect of total ion current correction on mean feature area and principal component analysis results

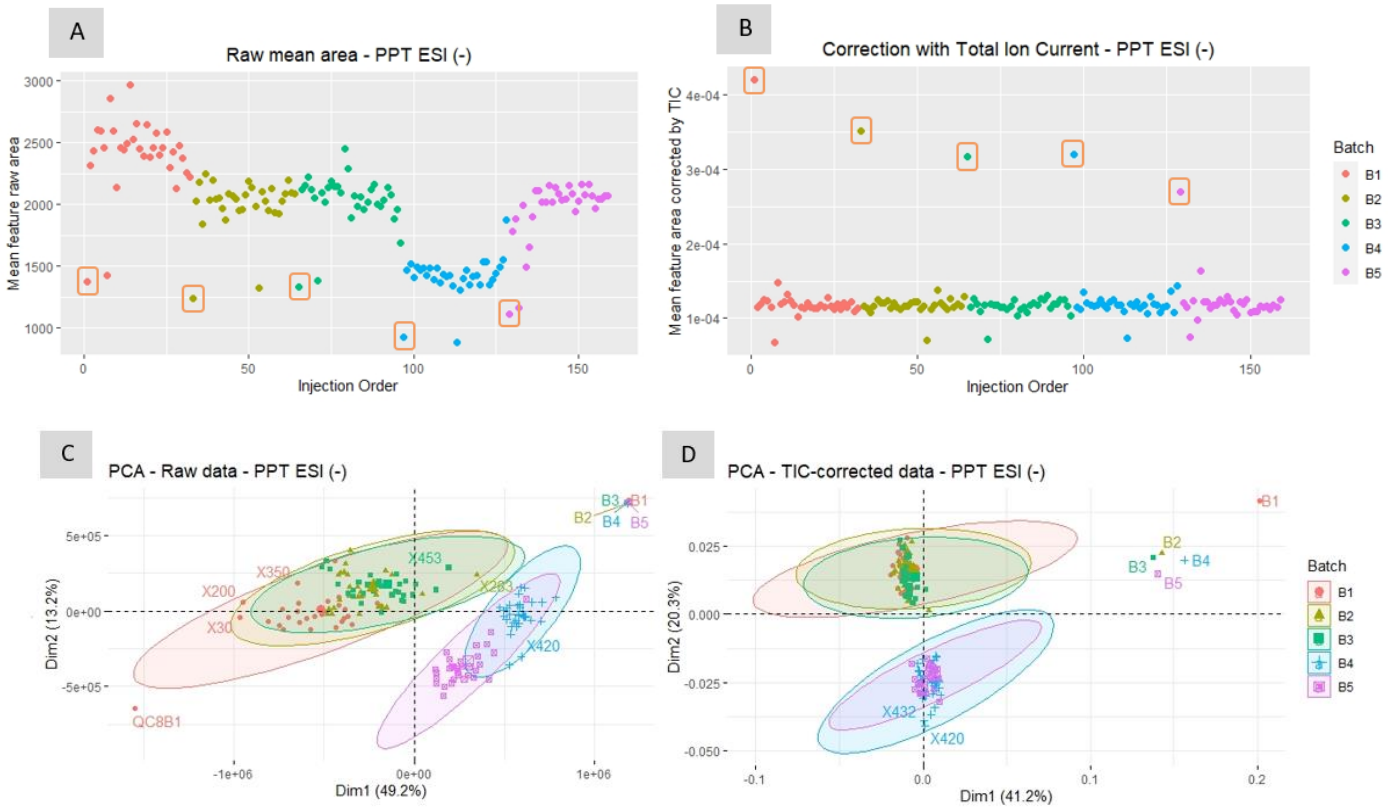

Figure B1 – Mean feature raw area (A), mean feature area after total ion current correction (B), PCA using raw area (C) and PCA using area after total ion current correction (D) shown on samples (including the composite quality control samples) prepared by protein precipitation (PPT) injected in ESI (−) mode. Blank samples for each batch are identified by orange squares.

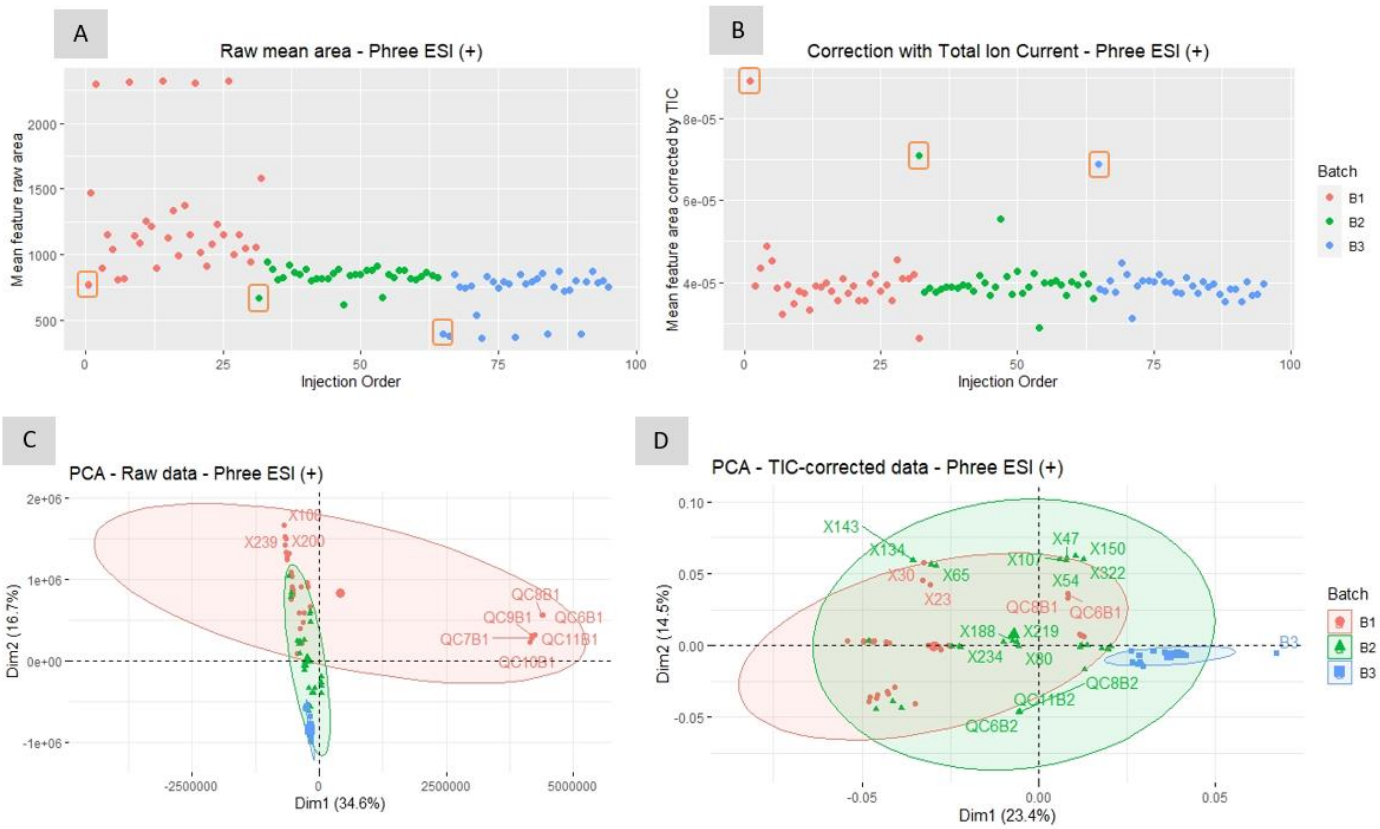

Figure B2 – Mean feature raw area (A), mean feature area after total ion current correction (B), PCA using raw area (C) and PCA using area after total ion current correction (D) shown on samples (including the composite quality control samples) prepared by phospholipid removal plates Phree injected in ESI (+) mode. Blank samples for each batch are identified by orange squares.

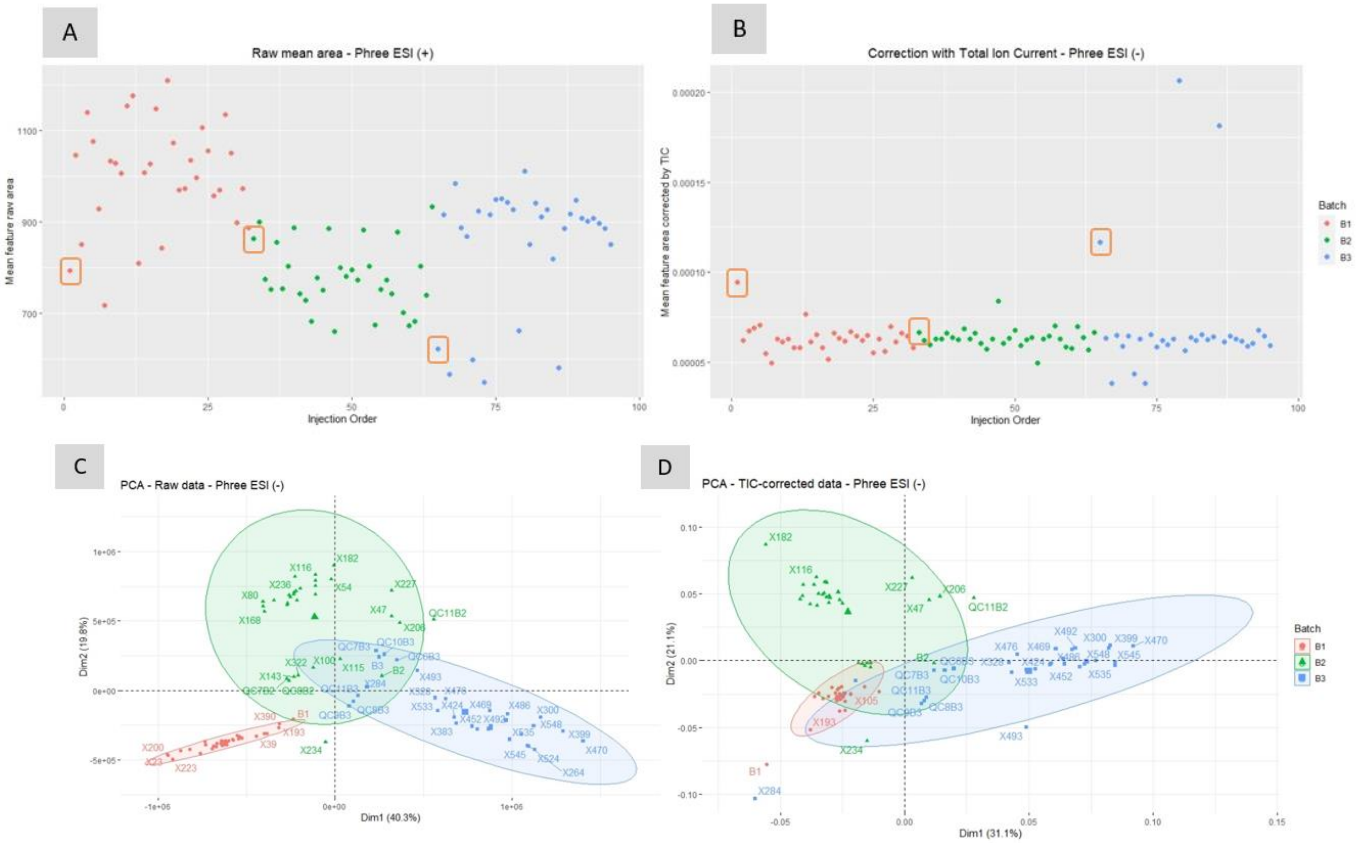

Figure B3 – Mean feature raw area (A), mean feature area after total ion current correction (B), PCA using raw area (C) and PCA using area after total ion current correction (D) shown on samples (including the composite quality control samples) prepared by phospholipid removal plates Phree injected in ESI (−) mode. Blank samples for each batch are identified by orange squares.

4.2. Annotations on Pélagie samples

Table A1 – Annotated compounds in Pélagie samples, with confidence indices (CI) on mass-to-charge (m/z) ratio, retention time (Rt), isotopic fit, and global confidence index. Compounds are either detected in the $[M+H]^+$ form ("(+)" columns) or the $[M-H]^-$ form ("(-)" columns)

Annotation	SMILES	m/z		Rt (min)	CI m/z		CI Rt								CI isotopic fit				Global CI	
							Experimental		RTI-predicted		Retip-predicted		logP-predicted		Considered Mn		CI overall			
		(+)	(-)		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
(2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid	C1=CC=C2C(=C1)C(C(=O)N2)CC(=O)O	192.0640	n.a.	2.7	0.95	n.a.	n.a.	n.a.	n.a.	n.a.	**0.99**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.97	n.a.
1,3,5-tris(2,2-dimethylpropionylamino)benzene	CC(C)(C)C(=O)NC1=CC(=CC(=C1)NC(=O)C(C)(C)C)NC(=O)C(C)(C)C	n.a.	410.2220	34.11	n.a.	0.86	n.a.	n.a.	n.a.	**0.66**	n.a.	0.66	n.a.	0	n.a.	M2	n.a.	0.54	n.a.	G3_0.76
10,11-trans-Dihydroxy-10,11-dihydrocarbamazepine	C1=CC=C2C(=C1)C(C(C3=CC=CC=C3N2C(=O)N)O)O	271.1100	n.a.	8.06	0.99	n.a.	n.a.	n.a.	n.a.	n.a.	**0.93**	n.a.	n.a.	n.a.	M2	n.a.	0.75	n.a.	G3_0.89	n.a.
13-Hydroxy-7,14-labdadien-6-one	CC1=CC(=O)C2C(CCCC2(C1CCC(C)(C=C)O)C)(C)C	305.2456	n.a.	46.63	0.87	n.a.	n.a.	n.a.	n.a.	n.a.	**0.67**	n.a.	0.42	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.77	n.a.
2-((3-dodecanamidopropyl)dimethylammonio)acetate	CCCCCCCCCCCC(=O)NCCC[N+](C)(C)CC(=O)[O-]	n.a.	377.2579	29.19	n.a.	0.88	n.a.	n.a.	n.a.	**0.77**	n.a.	0.65	n.a.	n.a.	n.a.	M2	n.a.	0.61	n.a.	G3_0.75
2-chlorophenol	C1=CC=C(C(=C1)O)Cl	n.a.	126.9957	14.01	n.a.	0.92	n.a.	n.a.	n.a.	**0.5**	n.a.	0.19	n.a.	0.5	n.a.	M2	n.a.	0.91	n.a.	G3_0.78
2-hydroxybenzoic acid	C1=CC=C(C(=C1)C(=O)O)O	n.a.	137.0243	21.9	n.a.	0.95	n.a.	n.a.	n.a.	**0.6**	n.a.	n.a.	n.a.	0.97	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.78
2-hydroxycarbamazepine	C1=CC=C2C(=C1)C=CC3=C(N2C(=O)N)C=CC(=C3)O	253.0966	n.a.	12.72	0.96	n.a.	**0.67**	n.a.	n.a.	n.a.	0.46	n.a.	0.55	n.a.	M1	n.a.	0.88	n.a.	G3_0.84	n.a.
2-Naphthalenesulfonic acid	C1=CC=C2C=C(C=CC2=C1)S(=O)(=O)O	n.a.	207.0124	7.25	n.a.	0.95	n.a.	n.a.	n.a.	**0.69**	n.a.	0.35	n.a.	0.37	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.82
2-Naphthol	C1=CC=C2C=C(C=CC2=C1)O	n.a.	143.0503	22.93	n.a.	0.94	n.a.	n.a.	n.a.	**0.87**	n.a.	0.4	n.a.	0.81	n.a.	M1	n.a.	0.65	n.a.	G3_0.82
2-naphthylamine	C1=CC=C2C=C(C=CC2=C1)N	144.0808	n.a.	14.41	0.97	n.a.	n.a.	n.a.	**0.75**	n.a.	0.23	n.a.	0.41	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.86	n.a.
3-(4-Hydroxyphenyl)lactic acid	C1=CC(=CC=C1CC(C(=O)O)O)O	n.a.	181.0499	4.07	n.a.	0.91	n.a.	n.a.	n.a.	**0.55**	n.a.	0.83	n.a.	0.63	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.73
3,5-dibromo-4-hydroxybenzoic acid	C1=C(C=C(C(=C1Br)O)Br)C(=O)O	n.a.	292.8458	19.22	n.a.	0.88	n.a.	n.a.	n.a.	**0.96**	n.a.	0	n.a.	0.59	n.a.	M2	n.a.	0.78	n.a.	G3_0.87
3-Formylindole	C1=CC=C2C(=C1)C(=CN2)C=O	146.0596	n.a.	4.24	0.95	n.a.	n.a.	n.a.	**0.68**	n.a.	0.11	n.a.	0.14	n.a.	M1	n.a.	0.76	n.a.	G3_0.8	n.a.
3-hydroxybenzoic acid	C1=CC(=CC(=C1)O)C(=O)O	n.a.	137.0245	5.38	n.a.	0.95	n.a.	n.a.	n.a.	**0.55**	n.a.	0.55	n.a.	0.41	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.75
4-chlorophenol	C1=CC(=CC=C1O)Cl	n.a.	126.9958	11.81	n.a.	0.94	n.a.	n.a.	n.a.	**0.76**	n.a.	0.05	n.a.	0.35	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.85
4-hydroxy-2,5,6-trichloroisophthalonitrile	C(#N)C1=C(C(=C(C(=C1Cl)C#N)Cl)Cl)O	n.a.	244.9085	27.74	n.a.	0.86	n.a.	n.a.	n.a.	n.a.	n.a.	**0.92**	n.a.	0.88	n.a.	M2	n.a.	0.96	n.a.	G3_0.91
4-hydroxybenzoic acid	C1=CC(=CC=C1C(=O)O)O	n.a.	137.0244	8.69	n.a.	1	n.a.	n.a.	n.a.	**0.8**	n.a.	0.72	n.a.	0.64	n.a.	M1	n.a.	0.88	n.a.	G3_0.89
4-hydroxyquinoline	C1=CC=C2C(=C1)C(=O)C=CN2	146.0599	n.a.	6.27	0.95	n.a.	n.a.	n.a.	**0.71**	n.a.	0.29	n.a.	0.22	n.a.	M1	n.a.	0.76	n.a.	G3_0.81	n.a.
5-acetylsalicylamide	CC(=O)C1=CC(=C(C=C1)O)C(=O)N	180.0657	n.a.	6.21	0.76	n.a.	**0.93**	n.a.	0.5	n.a.	0	n.a.	0.81	n.a.	M1	n.a.	0.94	n.a.	G3_0.88	n.a.
5-hydroxytryptophan	C1=CC2=C(C=C1O)C(=CN2)CC(C(=O)O)N	n.a.	219.0788	4.52	n.a.	0.91	n.a.	n.a.	n.a.	**0.89**	n.a.	0.27	n.a.	0.16	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.9
Acesulfame	CC1=CC(=O)NS(=O)(=O)O1	n.a.	161.9863	3.12	n.a.	0.90	n.a.	**0.96**	n.a.	0.28	n.a.	0.97	n.a.	0.64	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.95
Acetaminophen glucuronide	CC(=O)NC1=CC=C(C=C1)OC2C(C(C(C(O2)C(=O)O)O)O)O	n.a.	362.0640	4.09	n.a.	0.79	n.a.	**0.61**	n.a.	0.99	n.a.	0.97	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.7

Table A1 – (continued) Annotated compounds in Pélagie samples, with confidence indices (CI) on mass-to-charge (m/z) ratio, retention time (Rt), isotopic fit, and global confidence index. Compounds are detected in the $[M+H]^+$ form ("(+)" columns) or the $[M-H]^-$ form ("(-)" columns)

Annotation	SMILES	m/z		Rt (min)	CI m/z		CI Rt								CI Isotopic fit				Global CI	
							Experimental		RTI-predicted		Retip-predicted		logP-predicted		Considered Mn		CI overall			
		(+)	(-)		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
Acetaminophen sulfate	CC(=O)NC1=CC=C(C=C1)OS(=O)(=O)O	n.a.	230.0127	3.61	n.a.	0.97	n.a.	**0.91**	n.a.	0.84	n.a.	0.7	n.a.	0.8	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.94
Aminoacetophenone	C1=CC=C(C=C1)C(=O)CN	136.0752	n.a.	2.07	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Arabinosylhypoxanthine	C1=NC2=C(C(=O)N1)N=CN2C3C(C(C(O3)CO)O)O	n.a.	303.0501	2.98	n.a.	0.99	n.a.	n.a.	n.a.	n.a.	n.a.	**0.77**	n.a.	n.a.	n.a.	M2	n.a.	0.31	n.a.	G3_0.69
Aspartame	COC(=O)C(CC1=CC=CC=C1)NC(=O)C(CC(=O)O)N	295.13	n.a.	5.8	0.97	n.a.	**0.96**	n.a.	0.22	n.a.	0.95	n.a.	0.28	n.a.	M1	n.a.	0.94	n.a.	G3_0.96	n.a.
Auraptene	CC(=CCCC(=CCOC1=CC2=C(C=C1)C=CC(=O)O2)C)C	n.a.	297.1523	44.08	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Azelaic acid	C(CCCC(=O)O)CCCC(=O)O	n.a.	187.0977	9.41	n.a.	0.94	n.a.	n.a.	n.a.	**0.91**	n.a.	0.79	n.a.	0.78	n.a.	M1	n.a.	0.91	n.a.	G3_0.92
Benzothiazole	C1=CC=C2C(=C1)N=CS2	n.a.	136.0204	46.77	n.a.	0.91	n.a.	n.a.	n.a.	**0.83**	n.a.	n.a.	n.a.	0.68	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.87
Benzothiazole sulfonic acid	C1=CC=C2C(=C1)N=C(S2)S(=O)(=O)O	n.a.	213.9639	6.68	n.a.	0.99	n.a.	n.a.	n.a.	**0.83**	n.a.	0.62	n.a.	0.89	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.91
Benzylbutylphthalate	CCCCOC(=O)C1=CC=CC=C1C(=O)OCC2=CC=CC=C2	313.1439	n.a.	44.94	0.87	n.a.	**0.95**	n.a.	0.36	n.a.	0.72	n.a.	0.04	n.a.	M1	n.a.	0.77	n.a.	G3_0.86	n.a.
Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate	CC1=CC(=C(C(=C1)C(C)(C)C)O)C2=C(C(=CC(=C2)C)C(C)(C)C)OC(=O)C3=CC=C(C=C3)C(=O)OC4=C(C=C(C=C4C(C)(C)C)C)CC5=C(C(=CC(=C5)C)C(C)(C)C)O	811.4913	n.a.	12.55	0.85	n.a.	n.a.	n.a.	**0.78**	n.a.	0.32	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.82	n.a.
Bromoxynil	C1=C(C=C(C(=C1Br)O)Br)C#N	n.a.	275.9043	20.2	n.a.	0.84	n.a.	n.a.	n.a.	**0.42**	n.a.	0	n.a.	0.6	n.a.	M2	n.a.	0.79	n.a.	G3_0.68
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	195.09	n.a.	5.53	0.95	n.a.	n.a.	n.a.	**0.67**	n.a.	0.53	n.a.	0.77	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.81	n.a.
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	237.1016	n.a.	19.7	0.99	n.a.	n.a.	n.a.	**0.66**	n.a.	0	n.a.	0	n.a.	M1	n.a.	0.82	n.a.	G3_0.82	n.a.
Carveol	CC1=CCC(CC1O)C(=C)C	153.1272	n.a.	38.1	0.79	n.a.	n.a.	n.a.	**0.81**	n.a.	0	n.a.	n.a.	n.a.	M1	n.a.	0.77	n.a.	G3_0.79	n.a.
Carylophyllene oxide	CC1(CC2C1CCC3(C(O3)CC2=C)C)C	207.1735	n.a.	44.42	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chavicol sulfate	C=CCC1=CC=C(C=C1)OS(=O)(=O)O	n.a.	213.0226	11.38	n.a.	0.97	n.a.	**0.88**	n.a.	0.61	n.a.	0.54	n.a.	0.44	n.a.	M2	n.a.	0.91	n.a.	G3_0.92
Cinchonidine	C=CC1CN2CCC1CC2C(C3=CC=NC4=CC=CC=C34)O	n.a.	293.1752	32.49	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Cinnamaldehyde	C1=CC=C(C=C1)C=CC=O	133.0645	n.a.	14.33	0.82	n.a.	n.a.	n.a.	n.a.	n.a.	**0.68**	n.a.	0.76	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.75	n.a.
CMPF	CCCC1=C(C(=C(O1)CCC(=O)O)C(=O)O)C	n.a.	239.0925	22.08	n.a.	0.98	n.a.	n.a.	n.a.	**0.97**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.98
Cocamidopropyl Betaine	CCCCCCCCCCCC(=O)NCCC[N+](C)(C)CC(=O)[O-]	343.2939	n.a.	35.78	0.79	n.a.	n.a.	n.a.	**0.9**	n.a.	0.47	n.a.	n.a.	n.a.	M1	n.a.	0.9	n.a.	G3_0.86	n.a.
Coumaraldehyde	C1=CC(=CC=C1C=CC=O)O	149.0597	n.a.	3.36	0.92	n.a.	n.a.	n.a.	**0.96**	n.a.	0.86	n.a.	0.59	n.a.	M1	n.a.	0.91	n.a.	G3_0.93	n.a.
Coumaric acid	C1=CC(=CC=C1C=CC(=O)O)O	n.a.	163.0398	5.1	n.a.	0.9	n.a.	n.a.	n.a.	**0.74**	n.a.	0.23	n.a.	0.25	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.82
Cresol sulfate	CC1=CC=CC=C1OS(=O)(=O)O	n.a.	187.0066	7.35	n.a.	0.85	n.a.	**0.63**	n.a.	1	n.a.	0.62	n.a.	0.26	n.a.	M2	n.a.	0.86	n.a.	G3_0.78
Di(ethylhexyl) phthalate	CCCCC(CC)COC(=O)C1=CC=CC=C1C(=O)OCC(CC)CCCC	391.28	n.a.	45.67	0.91	n.a.	n.a.	n.a.	**0.79**	n.a.	0.69	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.85	n.a.

Table A1 – (continued) Annotated compounds in Pélagie samples, with confidence indices (CI) on mass-to-charge (m/z) ratio, retention time (Rt), isotopic fit, and global confidence index. Compounds are detected in the [M+H]+ form ("(+)" columns) or the [M-H]- form ("(-)" columns)

Annotation	SMILES	m/z		Rt (min)	CI m/z		CI Rt								CI Isotopic fit				Global CI	
							Experimental		RTI-predicted		Retip-predicted		logP-predicted		Considered Mn		CI overall			
		(+)	(-)		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
Dioctyl phthalate	CCCCCCCCOC(=O)C1=CC=CC=C1C(=O)OCCCCCCCC	391.2846	n.a.	49.09	0.95	n.a.	n.a.	n.a.	**0.86**	n.a.	0.78	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.91	n.a.
Diphenylphosphate	C1=CC=C(C=C1)OP(=O)([O-])OC2=CC=CC=C2	251.0466	n.a.	46.2	0.99	n.a.	**0.94**	n.a.	0.8	n.a.	0.02	n.a.	0.01	n.a.	M1	n.a.	0.93	n.a.	G3_0.95	n.a.
Diphenylsulfone	C1=CC=C(C=C1)S(=O)(=O)C2=CC=CC=C2	219.0473	n.a.	26.77	0.89	n.a.	n.a.	n.a.	**0.74**	n.a.	n.a.	n.a.	0.87	n.a.	M1	n.a.	0.82	n.a.	G3_0.82	n.a.
Docosahexaenoic acid	CCC=CCC=CCC=CCC=CCC=CCC=CCCC(=O)O	n.a.	327.2327	46.43	n.a.	0.97	n.a.	**1**	n.a.	0.79	n.a.	0.89	n.a.	0.6	n.a.	M1	n.a.	0.98	n.a.	G3_0.98
Dodecylbenzenesulfonic acid	CCCCCCCCCCCCC1=CC=CC=C1S(=O)(=O)O	n.a.	325.1851	48.41	n.a.	0.93	n.a.	n.a.	n.a.	**0.98**	n.a.	0.22	n.a.	0.69	n.a.	M1	n.a.	0.9	n.a.	G3_0.94
Eicosapentaenoic acid	CCC=CCC=CCC=CCC=CCC=CCCCC(=O)O	n.a.	301.2169	45.99	n.a.	0.91	n.a.	n.a.	n.a.	**0.88**	n.a.	0.96	n.a.	0.82	n.a.	M1	n.a.	0.92	n.a.	G3a_0.9
Ferulic acid	COC1=C(C=CC(=C1)C=CC(=O)O)O	n.a.	193.0505	6.5	n.a.	0.97	n.a.	n.a.	n.a.	**0.63**	n.a.	0.4	n.a.	0.39	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.8
Fipronil sulfone	C1=C(C=C(C(=C1Cl)N2C(=C(C(=N2)C#N)S(=O)(=O)C(F)(F)F)N)Cl)C(F)(F)F	n.a.	450.9301	44.56	n.a.	0.89	n.a.	n.a.	n.a.	**0.73**	n.a.	0.79	n.a.	n.a.	n.a.	M2	n.a.	0.95	n.a.	G3_0.86
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	n.a.	205.1233	41.19	n.a.	0.99	n.a.	**0.97**	n.a.	n.a.	n.a.	0.36	n.a.	0.46	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.98
Indole-3-acetaldehyde	C1=CC=C2C(=C1)C(=CN2)CC=O	160.0746	n.a.	9.52	0.86	n.a.	n.a.	n.a.	**0.94**	n.a.	0.62	n.a.	0.91	n.a.	M1	n.a.	0.89	n.a.	G3_0.90	n.a.
Indole-3-carbinol	C1=CC=C2C(=C1)C(=CN2)CO	130.0648	n.a.	10.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Indoxyl sulfate	C1=CC=C2C(=C1)C(=CN2)OS(=O)(=O)O	n.a.	212.0022	5.92	n.a.	0.96	n.a.	**0.86**	n.a.	n.a.	n.a.	n.a.	n.a.	0.41	n.a.	M2	n.a.	0.85	n.a.	G3_0.89
Ioxynil	C1=C(C=C(C(=C1I)O)I)C#N	n.a.	369.8217	28.48	n.a.	0.95	n.a.	n.a.	n.a.	**0.6**	n.a.	0.43	n.a.	0.33	n.a.	M1	n.a.	0.9	n.a.	G3_0.82
Isobutylparaben	CC(C)COC(=O)C1=CC=C(C=C1)O	195.1015	n.a.	40.82	0.88	n.a.	n.a.	n.a.	**0.72**	n.a.	0.38	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.8	n.a.
Isopropylparaben	CC(C)OC(=O)C1=CC=C(C=C1)O	181.0857	n.a.	8.43	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Lenticin	C[N+](C)(C)C(CC1=CNC2=CC=CC=C21)C(=O)[O-]	247.15	n.a.	5.26	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Lidocaine	CCN(CC)CC(=O)NC1=C(C=CC=C1C)C	235.1804	n.a.	5.94	0.93	n.a.	n.a.	n.a.	**0.81**	n.a.	0.72	n.a.	0.8	n.a.	M1	n.a.	0.87	n.a.	G3_0.87	n.a.
Lumichrome	CC1=CC2=C(C=C1C)N=C3C(=N2)C(=O)NC(=O)N3	243.0877	n.a.	10.08	0.93	n.a.	n.a.	n.a.	**0.52**	n.a.	0	n.a.	0.01	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.73	n.a.
Mercaptobenzothiazole	C1=CC=C2C(=C1)NC(=S)S2	n.a.	165.9795	15.22	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Methionine	CSCCC(C(=O)O)N	150.06	n.a.	1.53	0.78	n.a.	n.a.	n.a.	**0.78**	n.a.	0.51	n.a.	0.1	n.a.	M1	n.a.	0.87	n.a.	G3_0.81	n.a.
Methylperfluorooctanesulfonamido)acetic acid	CN(CC(=O)O)S(=O)(=O)C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F	n.a.	569.9673	44.87	n.a.	0.95	n.a.	n.a.	n.a.	n.a.	n.a.	**0.79**	n.a.	0.8	n.a.	M2	n.a.	0.70	n.a.	G3_0.81
Paracetamol	CC(=O)NC1=CC=C(C=C1)O	152.071	n.a.	4.24	0.83	n.a.	n.a.	n.a.	n.a.	n.a.	**0.74**	n.a.	0.69	n.a.	M1	n.a.	0.95	n.a.	G3_0.84	n.a.
Paraxanthine	CN1C=NC2=C1C(=O)N(C(=O)N2)C	n.a.	179.0575	5.23	n.a.	0.96	n.a.	n.a.	n.a.	**0.82**	n.a.	0.78	n.a.	n.a.	n.a.	M1	n.a.	0.83	n.a.	G3_0.87
Pentachlorophenol	C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)O	n.a.	262.8401	43.01	n.a.	0.85	n.a.	n.a.	n.a.	**0.84**	n.a.	0	n.a.	0.72	n.a.	M2	n.a.	0.91	n.a.	G3_0.87

Table A1 – (continued) Annotated compounds in Pélagie samples, with confidence indices (CI) on mass-to-charge (m/z) ratio, retention time (Rt), isotopic fit, and global confidence index. Compounds are detected in the $[M+H]^+$ form ("(+)" columns) or the $[M-H]^-$ form ("(-)" columns)

Annotation	SMILES	m/z		Rt (min)	CI m/z		CI Rt								CI Isotopic fit				Global CI	
							Experimental		RTI-predicted		Retip-predicted		logP-predicted		Considered Mn		CI overall			
		(+)	(-)		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
Perfluoroheptanesulfonic acid	C(C(C(C(F)(F)F)(F)F)(F)F)(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(F)F	n.a.	448.9334	33.49	n.a.	0.84	n.a.	n.a.	n.a.	**0.73**	n.a.	0.13	n.a.	0.93	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.78
Perfluorohexanesulfonic acid	C(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(C(C(F)(F)F)(F)F)(F)F	n.a.	398.9366	35.31	n.a.	0.89	n.a.	n.a.	n.a.	n.a.	n.a.	**0.84**	n.a.	0.84	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.87
Perfluorooctanesulfonic acid	C(C(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(F)F)(C(C(C(F)(F)F)(F)F)(F)F)(F)F	n.a.	498.9305	43.18	n.a.	0.93	n.a.	n.a.	n.a.	n.a.	n.a.	**0.76**	n.a.	0.79	n.a.	M2	n.a.	0.77	n.a.	G3_0.82
Phenol sulfate	C1=CC=C(C=C1)OS(=O)(=O)O	n.a.	172.9913	4.74	n.a.	0.97	n.a.	**0.62**	n.a.	n.a.	n.a.	0.55	n.a.	0.36	n.a.	M2	n.a.	0.87	n.a.	G3_0.82
Piperidone	C1CCNC(=O)C1	100.0751	n.a.	3.78	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Piperine	C1CCN(CC1)C(=O)C=CC=CC2=CC3=C(C=C2)OCO3	286.1436	n.a.	37.81	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Propylparaben	CCCOC(=O)C1=CC=C(C=C1)O	n.a.	179.0711	21.58	n.a.	0.95	n.a.	**0.76**	n.a.	0.46	n.a.	0.04	n.a.	0.19	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.86
Propylparaben sulfate	CCCOC(=O)C1=CC=C(C=C1)OS(=O)(=O)O	n.a.	259.0282	17.12	n.a.	0.98	n.a.	**0.81**	n.a.	0.23	n.a.	0.73	n.a.	0.65	n.a.	M2	n.a.	0.9	n.a.	G3_0.9
Reserpine	COC1C(CC2CN3CCC4=C(C3CC2C1C(=O)OC)NC5=C4C=CC(=C5)OC)OC(=O)C6=CC(=C(C(=C6)OC)OC)OC	609.2801	n.a.	15.39	0.96	n.a.	n.a.	n.a.	**0.99**	n.a.	0.48	n.a.	0	n.a.	M1	n.a.	0.9	n.a.	G3_0.95	n.a.
Solanidine	CC1CCC2C(C3C(N2C1)CC4C3(CCC5C4CC=C6C5(CCC(C6)O)C)C)C	398.34	n.a.	26.15	0.91	n.a.	n.a.	n.a.	**0.63**	n.a.	0.78	n.a.	0.75	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.77	n.a.
Sucralose	C(C1C(C(C(C(O1)OC2(C(C(C(O2)CCl)O)O)CCl)O)O)Cl)O	n.a.	430.9843	6.78	n.a.	0.90	n.a.	n.a.	n.a.	**0.59**	n.a.	0.97	n.a.	n.a.	n.a.	M2	n.a.	0.98	n.a.	G3_0.82
Theobromine	CN1C=NC2=C1C(=O)NC(=O)N2C	181.07	n.a.	3.86	0.95	n.a.	n.a.	n.a.	**0.45**	n.a.	n.a.	n.a.	0	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.7	n.a.
Theophylline	CN1C2=C(C(=O)N(C1=O)C)NC=N2	181.07	n.a.	5.91	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Thymol	CC1=CC(=C(C=C1)C(C)C)O	151.1111	n.a.	19.52	0.93	n.a.	n.a.	n.a.	n.a.	n.a.	**0.93**	n.a.	0.91	n.a.	M1	n.a.	0.98	n.a.	G3_0.95	n.a.
Triclosan glucuronide	C1=CC(=C(C=C1Cl)OC2C(C(C(C(O2)C(=O)O)O)O)O)OC3=C(C=C(C=C3)Cl)Cl	n.a.	462.9753	27.17	n.a.	0.86	n.a.	n.a.	n.a.	**0.91**	n.a.	n.a.	n.a.	0.85	n.a.	M2	n.a.	1	n.a.	G3_0.92
Triclosan sulfate	C1=CC(=C(C=C1Cl)OS(=O)(=O)O)OC2=C(C=C(C=C2)Cl)Cl	n.a.	366.9009	36.32	n.a.	0.94	n.a.	**0.97**	n.a.	0	n.a.	n.a.	n.a.	0.98	n.a.	M2	n.a.	0.88	n.a.	G3_0.93
Triethylphosphate	CCOP(=O)(OCC)OCC	183.0783	n.a.	12.57	0.89	n.a.	n.a.	n.a.	**0.77**	n.a.	0.82	n.a.	0.92	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.83	n.a.
Triphenylphosphine oxide	C1=CC=C(C=C1)P(=O)(C2=CC=CC=C2)C3=CC=CC=C3	279.0925	n.a.	2.76	0.8	n.a.	n.a.	n.a.	**0.55**	n.a.	0.7	n.a.	0.59	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.72	n.a.
Tris(2-butoxyethyl)phosphate	CCCCOCCOP(=O)(OCCOCCCC)OCCOCCCC	399.25	n.a.	44.69	0.93	n.a.	n.a.	n.a.	n.a.	n.a.	**0.66**	n.a.	0.82	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.8	n.a.
Tritosulfuron	COC1=NC(=NC(=N1)NC(=O)NS(=O)(=O)C2=CC=CC=C2C(F)(F)F)C(F)(F)F	n.a.	444.0210	34.47	n.a.	0.91	n.a.	**0.96**	n.a.	0.86	n.a.	0.69	n.a.	0.33	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.94
Tryptophan	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N	205.0968	n.a.	4.61	0.83	n.a.	n.a.	n.a.	**0.9**	n.a.	0.83	n.a.	0.51	n.a.	n.a.	n.a.	n.a.	n.a.	G2_0.87	n.a.

4.3. MS2 data for annotated compounds

Table A2 – MS2 theoretical and experimental fragmentation data for annotated compounds

Annotation	MS/MS			
	Theoretical fragments		Experimental fragments	
	(+)	(-)	(+)	(-)
(2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid	77.0391, 91.0555, 103.0545	n.a.	77.0390, 91.0554, 103.0548	n.a.
1,3,5-tris(2,2-dimethylpropionylamino)benzene	n.a.	232.1091, 290.1874, 316.1667	n.a.	232.1089, 290.1875, 316.1671
10,11-trans-Dihydroxy-10,11-dihydrocarbamazepine	87.0256, 104.0529, 133.0318	n.a.	87.0256, 104.0536, 133.0318	n.a.
13-Hydroxy-7,14-labdadien-6-one	195.0648, 397.2115	n.a.	195.0655, 397.2110	n.a.
2-((3-dodecanamidopropyl)dimethylammonio)acetate	n.a.	102.0561, 238.2177, 283.2756	n.a.	102.0558, 238.2182, 283.2751
2-chlorophenol	n.a.	91.02	n.a.	91.02
2-hydroxybenzoic acid	n.a.	93.03	n.a.	93.03
2-hydroxycarbamazepine	145.0764, 172.0870, 198.0663	n.a.	145.0768, 172.0864, 198.0659	n.a.
2-Naphthalenesulfonic acid	n.a.	79.9574, 114.0553, 143.0503	n.a.	79.9577, 115.0546, 143.0500
2-Naphthol	n.a.	98.9055, 115.0553	n.a.	98.9057, 115.0557
2-naphthylamine	91.0556, 115.0542, 117.0699, 127.0542	n.a.	91.0549, 115.0543, 117.0690, 127.0545	n.a.
3-(4-Hydroxyphenyl)lactic acid	n.a.	72.9921, 93.0301, 107.0480, 119.0489, 134.0377, 135.0456	n.a.	72.9922, 93.0304, 107.0482, 119.0488, 134.0372, 135.0455
3,5-dibromo-4-hydroxybenzoic acid	n.a.	248.86	n.a.	248.85
3-Formylindole	65.0400, 77.0389, 91.0560, 117.0590	n.a.	65.0396, 77.0384, 91.0551, 117.0583	n.a.
3-hydroxybenzoic acid	n.a.	93.03	n.a.	93.03
4-chlorophenol	n.a.	n.a.	n.a.	n.a.
4-hydroxy-2,5,6-trichloroisophthalonitrile	n.a.	146.9765, 174.9704, 181.9447, 209.9401	n.a.	146.9765, 174.9710, 181.9447, 209.9406
4-hydroxybenzoic acid	n.a.	93.0343	n.a.	93.0346
4-hydroxyquinoline	77.0415, 91.0555, 104.0494, 128.0476	n.a.	77.0415, 91.0548, 104.0490, 128.0479	n.a.
5-acetylsalicylamide	165.0694, 179.0726, 194.0958	n.a.	165.0697, 179.0720, 194.0963	n.a.
5-hydroxytryptophan	n.a.	132.0447, 144.0456, 158.0622	n.a.	132.0453, 144.0450, 158.0630
Acesulfame	n.a.	67.0065, 77.9655, 82.0298	n.a.	67.0063, 77.9654, 82.0299
Acetaminophen glucuronide	n.a.	113.0252, 150.0561, 175.0252	n.a.	113.0245, 150.0562, 175.0247
Acetaminophen sulfate	n.a.	107.0365, 108.0445, 150.0551	n.a.	107.0367, 108.0452, 150.0556
Aminoacetophenone	167.0729, 180.0807, 182.0963, 193.0886, 210.0914	n.a.	167.0730, 180.0803, 182.0966, 193.0885, 210.0917	n.a.

Table A2 – (continued) MS2 theoretical and experimental fragmentation data for annotated compounds

Annotation	MS/MS			
	Theoretical fragments		Experimental fragments	
	(+)	(-)	(+)	(-)
Arabinosylhypoxanthine	n.a.	92.0241, 108.0190, 135.0301	n.a.	92.0247, 108.0199, 135.0305
Aspartame	115.0543, 135.0446, 143.0485, 171.0453, 201.0548	n.a.	115.0551, 135.0450, 143.0495, 171.0454, 201.0556	n.a.
Auraptene	n.a.	170.0038, 183.0014, 197.0272	n.a.	170.0040, 183.0114, 197.0244
Azelaic acid	n.a.	57.0331, 95.0488, 97.0645, 123.0803, 125.0959	n.a.	57.0339, 95.0486, 97.0652, 123.0806, 125.0963
Benzothiazole	n.a.	65.0382, 105.0448, 109.0108	n.a.	65.0382, 105.0448, 109.0109
Benzothiazole sulfonic acid	n.a.	57.9751, 134.0069, 150.0019	n.a.	57.9749, 134.0071, 150.0023
Benzylbutylphthalate	380.332	n.a.	380.3319	n.a.
Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate	91.0546, 115.0548,130.0655, 143.0727, 146.0592, 159.0921, 170.0600	n.a.	91.0543, 115.0543,130.0654, 143.0728, 146.0598, 159.0925, 170.0605	n.a.
Bromoxynil	n.a.	78.92	n.a.	78.92
Caffeine	149.02	n.a.	149.02	n.a.
Carbamazepine	95.049, 121.0282, 139.0388	n.a.	95.0491, 121.0283, 139.0384	n.a.
Carveol	283.1693, 431.1844, 589.2939	n.a.	283.1700, 431.1849, 589.2947	n.a.
Carylophyllene oxide	51.0233, 53.0389, 77.0382, 95.0493, 105.0447, 125.0055, 141.0004	n.a.	51.0229, 53.0388, 77.0375, 95.0491, 105.0445, 125.0060, 140.9999	n.a.
Chavicol sulfate	n.a.	105.0710, 133.0659	n.a.	105.0710, 133.0657
Cinchonidine	n.a.	96.9588, 221.1544, 236.1056	n.a.	96.9590, 221.1546, 236.1055
Cinnamaldehyde	79.0548, 81.0701, 91.0546, 95.0854, 105.0702, 107.0849, 133.1028, 147.1185, 161.1331	n.a.	79.0539, 81.0705, 91.0537, 95.0844, 105.0699, 107.0848, 133.1017, 147.1179, 161.1330	n.a.
CMPF	n.a.	96.9588, 135.0810, 151.1119, 177.0913, 195.1021	n.a.	96.9584, 135.0818, 151.1121, 177.0925, 195.1027
Cocamidopropyl Betaine	67.0282, 108.0554, 110.0713, 122.0589, 138.0668, 163.0611	n.a.	67.0288, 108.0556, 110.0711, 122.0594, 138.0661, 163.0613	n.a.
Coumaraldehyde	77.0386, 105.0335	n.a.	77.0388, 105.0337	n.a.
Coumaric acid	n.a.	145.9019	n.a.	145.9011
Cresol sulfate	n.a.	92.0279, 107.0493	n.a.	92.0275, 107.0500
Di(ethylhexyl) phthalate	69.0454, ,96.0561, 124.0507, 142.0611	n.a.	69.0451, 96.0563, 124.0502, 142.0612	n.a.
Dioctyl phthalate	77.0380, 79.0550, 9.0540, 95.0490, 108.0200, 121.0650, 123.0440	n.a.	77.0386, 79.0552, 9.0544, 95.0485, 108.0199, 121.0651, 123.0441	n.a.

Table A2 – (continued) MS2 theoretical and experimental fragmentation data for annotated compounds

Annotation	MS/MS			
	Theoretical fragments		Experimental fragments	
	(+)	(-)	(+)	(-)
Diphenylphosphate	58.0652, 86.0968	n.a.	58.0655, 86.0965	n.a.
Diphenylsulfone	91.0553, 117.0564, 118.0657, 130.0648	n.a.	91.0549, 117.0568, 118.0658, 130.0655	n.a.
Docosahexaenoic acid	n.a.	229.1958, 283.2446	n.a.	229.1953, 283.2439
Dodecylbenzenesulfonic acid	n.a.	170.0042, 183.0121, 197.0277, 255.1376	n.a.	170.0041, 183.0128, 197.0287, 255.1377
Eicosapentaenoic acid	n.a.	203.1802, 229.1957, 257.2274	n.a.	203.1807, 229.1951, 257.2275
Ferulic acid	n.a.	133.0299, 149.0608	n.a.	133.0305, 149.0609
Fipronil sulfone	n.a.	246.0120, 281.9913, 414.9496	n.a.	246.0117, 281.9920, 414.9500
Ibuprofen	n.a.	91.0549, 105.0701, 119.0855	n.a.	91.0555, 105.0708, 119.0848
Indole-3-acetaldehyde	69.0442, 83.0606, 110.0718, 123.0425, 138.0659	n.a.	69.0448, 83.0603, 110.0712, 123.0427, 138.0662	n.a.
Indole-3-carbinol	167.0730, 180.0808, 182.0964, 210.0914	n.a.	167.0721, 180.0812, 182.0966, 210.0920	n.a.
Indoxyl sulfate	n.a.	79.9578, 132.0460	n.a.	79.9570, 132.0457
Ioxynil	n.a.	126.9051, 230.9182	n.a.	126.9041, 230.9178
Isobutylparaben	71.0851, 149.0232, 261.1485	n.a.	71.0858, 149.0234, 261.1490	n.a.
Isopropylparaben	55.0195, 77.0392, 91.0541, 103.0549, 105.0707, 115.0545	n.a.	55.0199, 77.0386, 91.0544, 103.0550, 105.0707, 115.0541	n.a.
Lenticin	60.0815, 118.0653, 146.0600, 170.0599, 188.0705	n.a.	60.0810, 118.0651, 146.0597, 170.0596, 188.0713	n.a.
Lidocaine	77.0380, 103.0560, 128.0500, 130.0638	n.a.	77.0388, 103.0555, 128.0501, 130.0642	n.a.
Lumichrome	121.0284, 139.0389, 163.0754	n.a.	121.0282, 139.0393, 163.0751	n.a.
Mercaptobenzothiazole	n.a.	57.9752, 134.0069	n.a.	57.9750, 134.0063
Methionine	51.0237, 65.0364, 91.0539, 117.0576, 118.0646	n.a.	51.0236, 65.0365, 91.0541, 117.0570, 118.0651	n.a.
Methylperfluorooctanesulfonamido)acetic acid	n.a.	418.9773, 482.9356, 511.9607	n.a.	418.9771, 482.9356, 511.9617
Paracetamol	91.0543, 149.02335, 239.0708	n.a.	91.0542, 149.0234, 239.0707	n.a.
Paraxanthine	n.a.	122.0365, 164.0341	n.a.	122.0361, 164.0341
Pentachlorophenol	n.a.	n.a.	n.a.	n.a.
Perfluoroheptanesulfonic acid	n.a.	168.9892	n.a.	168.9903
Perfluorohexanesulfonic acid	n.a.	98.9538, 118.9930, 168.9892	n.a.	98.9535, 118.9937, 168.9882
Perfluorooctanesulfonic acid	n.a.	98.9538, 118.9930, 168.9892	n.a.	98.9533, 118.9931, 168.9985
Phenol sulfate	n.a.	79.9551, 93.0325	n.a.	79.9558, 93.0331

Table A2 – (continued) MS2 theoretical and experimental fragmentation data for annotated compounds

Annotation	MS/MS			
	Theoretical fragments		Experimental fragments	
	(+)	(-)	(+)	(-)
Piperidone	91.0543, 118.0662, 128.0511, 132.0424, 146.0614	n.a.	91.0546, 118.0660, 128.0519, 132.0429, 146.0607	n.a.
Piperine	56.0493, 72.0444, 82.0651, 94.0650	n.a.	56.0494, 72.0451, 82.0651, 94.0649	n.a.
Propylparaben	n.a.	92.0266, 121.0300, 136.0167	n.a.	92.0263, 121.0302, 136.0166
Propylparaben sulfate	n.a.	121.0297, 137.0239, 179.0716	n.a.	121.0296, 137.0244, 179.0712
Reserpine	57.0701, 83.0855, 101.0971, 143.0104, 199.0730, 299.1618	n.a.	57.0702, 83.0850, 101.0972, 143.0102, 199.0738, 299.1623	n.a.
Solanidine	127.0164, 155.0480	n.a.	127.0155, 155.0472	n.a.
Sucralose	n.a.	146.9399, 359.0325	n.a.	146.9391, 359.0319
Theobromine	79.0544, 91.0543, 107.0856, 119.0856	n.a.	79.0551, 91.0544, 107.0853, 119.0859	n.a.
Theophylline	67.0544, 81.0699, 91.0543, 105.0714, 119.0847	n.a.	67.0545, 81.0701, 91.0547, 105.0721, 119.0841	n.a.
Thymol	103.0542, 120.0808, 130.0651, 131.0497	n.a.	103.0542, 120.0811, 130.0655, 131.0499	n.a.
Triclosan glucuronide	n.a.	n.a.	n.a.	n.a.
Triclosan sulfate	n.a.	286.9448	n.a.	286.9448
Triethylphosphate	183.1745, 240.2315	n.a.	183.1751, 240.2326	n.a.
Triphenylphosphine oxide	110.0598, 134.0593	n.a.	110.0602, 134.0596	n.a.
Tris(2-butoxyethyl)phosphate	149.0219, 173.0513, 201.0465, 219.0570	n.a.	149.0220, 173.0515, 201.0472, 219.0568	n.a.
Tritosulfuron	n.a.	193.0347, 223.9999	n.a.	193.0344, 223.9998
Tryptophan	77.0386, 95.0492, 152.0633, 175.0156, 215.0257	n.a.	77.0387, 95.0484, 152.0627, 175.0165, 215.0252	n.a.

4.4. Confidence levels, detection frequency and toxicological data

Table A3 – Confidence levels according to Schymanski et al. (2014) and to the updated classification, detection frequency, and availability of toxicological data from the CompTox dashboard.

Annotation	Confidence level (Schymanski 2014)	Confidence level (Updated)	Detection frequency (%)		Available toxicological data
			Phree	PPT	
(2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid	2a	2a	0	100	
1,3,5-tris(2,2-dimethylpropionylamino)benzene	2b	2b	45	45	
10,11-trans-Dihydroxy-10,11-dihydrocarbamazepine	2a	2a	0	1	
13-Hydroxy-7,14-labdadien-6-one	2a	2a	0	68	
2-((3-dodecanamidopropyl)dimethylammonio)acetate	2b	2b	0	1	
2-chlorophenol	2a	2a	8	7	X
2-hydroxybenzoic acid	2a	2a	0	88	
2-hydroxycarbamazepine	2a	2a	84	1	
2-Naphthalenesulfonic acid	2a	2a	0	39	
2-Naphthol	2a	2a	85	2	X
2-naphthylamine	2a	2a	0	8	X
3-(4-Hydroxyphenyl)lactic acid	2a	2a	87	85	X
3,5-dibromo-4-hydroxybenzoic acid	2a	2a	97	0	X
3-Formylindole	2a	2a	3	100	X
3-hydroxybenzoic acid	2a	2a	50	50	X
4-chlorophenol	4	MS1-3	80	2	X
4-hydroxy-2,5,6-trichloroisophthalonitrile	2a	2a	74	68	X
4-hydroxybenzoic acid	2a	2a	0	88	X
4-hydroxyquinoline	2a	2a	84	88	X
5-acetylsalicylamide	1	1	42	95	X
5-hydroxytryptophan	2a	2a	43	96	X
Acesulfame	2a	2a	97	15	X
Acetaminophen glucuronide	1	1	20	1	X
Acetaminophen sulfate	1	1	0	14	
Aminoacetophenone	2a	2a	45	100	
Arabinosylhypoxanthine	2a	2a	30	98	X
Aspartame	1	1	35	100	X
Auraptene	2a	2a	100	1	X
Azelaic acid	2a	2a	35	28	X
Benzothiazole	2a	2a	97	6	X
Benzothiazole sulfonic acid	2a	2a	3	2	
Benzylbutylphthalate	1	1	3	0	X
Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate	2b	2b	23	86	
Bromoxynil	2a	2a	61	2	X
Caffeine	2a	2a	1	94	X
Carbamazepine	2a	2a	95	1	X
Carveol	2a	2a	20	2	X
Carylophyllene oxide	2a	2a	91	14	
Chavicol sulfate	2b	2b	28	100	
Cinchonidine	2a	2a	81	99	X
Cinnamaldehyde	2a	2a	33	75	
CMPF	2a	2a	20	59	X
Cocamidopropyl Betaine	2a	2a	97	29	X
Coumaraldehyde	2b	2b	64	100	
Coumaric acid	2a	2a	100	3	
Cresol sulfate	1	1	16	100	X
Di(ethylhexyl) phthalate	2a	2a	97	56	X
Dioctyl phthalate	2a	2a	10	10	X
Diphenylphosphate	2a	2a	17	0	
Diphenylsulfone	2a	2a	45	9	X
Docosahexaenoic acid	1	1	85	100	X
Dodecylbenzenesulfonic acid	2a	2a	23	81	X
Eicosapentaenoic acid	2a	2a	5	100	X

Table A3 – (continued) Confidence levels according to Schymanski et al. (2014) and to the updated classification, detection frequency, and availability of toxicological data from the CompTox dashboard.

Annotation	Confidence level (Schymanski 2014)	Confidence level (Updated)	Detection frequency (%)		Available toxicological data
			Phree	PPT	
Ferulic acid	2b	2b	65	1	X
Fipronil sulfone	2a	2a	12	29	X
Ibuprofen	1	1	25	1	X
Indole-3-acetaldehyde	2a	2a	0	98	X
Indole-3-carbinol	2a	2a	20	100	X
Indoxyl sulfate	1	1	16	100	X
Ioxynil	2a	2a	17	92	X
Isobutylparaben	2a	2a	0	72	X
Isopropylparaben	2a	2a	96	2	X
Lenticin	2a	2a	3	94	
Lidocaine	2a	2a	29	93	X
Lumichrome	2b	2b	15	95	X
Mercaptobenzothiazole	2a	2a	38	98	X
Methionine	2a	2a	97	49	X
Methylperfluorooctanesulfonamido)acetic acid	2a	2a	100	3	X
Paracetamol	2a	2a	5	5	X
Paraxanthine	2a	2a	7	91	X
Pentachlorophenol	4	MS1-3	92	5	X
Perfluoroheptanesulfonic acid	2a	2a	1	2	X
Perfluorohexanesulfonic acid	2a	2a	0	57	X
Perfluorooctanesulfonic acid	2a	2a	97	98	X
Phenol sulfate	2a	2a	97	100	X
Piperidone	2a	2a	27	82	X
Piperine	2a	2a	13	86	X
Propylparaben	2a	2a	97	99	X
Propylparaben sulfate	1	1	95	90	
Reserpine	2a	2a	67	50	X
Solanidine	2a	2a	8	7	X
Sucralose	2a	2a	13	12	X
Theobromine	2a	2a	13	100	X
Theophylline	2a	2a	100	6	X
Thymol	2a	2a	0	2	X
Triclosan glucuronide	4	MS1-2	25	10	X
Triclosan sulfate	1	1	13	10	
Triethylphosphate	2a	2a	16	10	X
Triphenylphosphine oxide	2a	2a	45	77	X
Tris(2-butoxyethyl)phosphate	2a	2a	26	26	X
Tritosulfuron	1	1	91	2	X
Tryptophan	2a	2a	81	100	X

4.5. Classification of compounds annotated in Pélagie samples

Table A4 – Classification of compounds annotated in the Pélagie samples. P refer to primary uses, and S to secondary uses

	Gut-derived	Food				Health and personal care			Environmental pollutants				
Molecule	**Gut microbiota metabolites**	**Natural compound**	**Flavoring agent**	**Preservatives and other stabilizers**	**Indirect food additive**	**Medication**	**Personal care and cosmetics products**	**Preservatives and other stabilizers**	**Pesticides**	**Plasticizers**	**Organophosphate flame retardant**	**Synthesis intermediate**	**Preservatives and other stabilizers**
(2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid		P											
1,3,5-tris(2,2-dimethylpropionylamino)benzene					P								
10,11-trans-Dihydroxy-10,11-dihydrocarbamazepine						P							
13-Hydroxy-7,14-labdadien-6-one		P	S										
2-((3-dodecanamidopropyl)dimethylammonio)acetate		S					P						
2-chlorophenol												P	
2-hydroxybenzoic acid				P		S						S	
2-hydroxycarbamazepine						P							
2-Naphthalenesulfonic acid												P	
2-Naphthol						S			S			P	
2-naphthylamine												P	
3-(4-Hydroxyphenyl)lactic acid	P	S											
3,5-dibromo-4-hydroxybenzoic acid									P				
3-Formylindole	P	S											
3-hydroxybenzoic acid			P										
4-chlorophenol						S			S			P	
4-hydroxy-2,5,6-trichloroisophthalonitrile									P				
4-hydroxybenzoic acid				S				P					
4-quinolone		P											
5-acetylsalicylamide						P						S	
5-hydroxytryptophan	P												
Acesulfame			P				S						
Acetaminophen glucuronide						P							
Acetaminophen sulfate						P							
Aminoacetophenone		P											
Arabinosylhypoxanthine		P											
Aspartame			P										
Auraptene		P											
Azelaic acid		P				S	S						S
Benzothiazole			P										
Benzothiazole sulfonic acid			P										
Benzylbutylphthalate					S		S			P			
Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate					P								
Bromoxynil									P				
Caffeine		P					S						
Carbamazepine						P							

Table A4 – (continued) Classification of compounds annotated in the Pélagie samples. P refer to primary uses, and S to secondary uses

	Gut-derived	Food				Health and personal care			Environmental pollutants				
Molecule	**Gut microbiota metabolites**	**Natural compound**	**Flavoring agent**	**Preservatives and other stabilizers**	**Indirect food additive**	**Medication**	**Personal care and cosmetics products**	**Preservatives and other stabilizers**	**Pesticides**	**Plasticizers**	**Organophosphate flame retardant**	**Synthesis intermediate**	**Preservatives and other stabilizers**
Carveol		P	S				S						
Carylophyllene oxide		S	P				S						
Chavicol sulfate		P	S										
Cinchonidine		P											
Cinnamaldehyde		P					S						
CMPF		P											
Cocamidopropyl Betaine							P						
Coumaraldehyde		P											
Coumaric acid							P						
Cresol sulfate	P	S											
Di(ethylhexyl) phthalate					S					P			
Dioctyl phthalate					S					P			
Diphenylphosphate												P	S
Diphenylsulfone												P	
Docosahexaenoic acid		P											
Dodecylbenzenesulfonic acid							P						
Eicosapentaenoic acid		P											
Ferulic acid			S	P									
Fipronil sulfone									P				
Ibuprofen						P							
Indole-3-acetaldehyde		P											
Indole-3-carbinol		P											
Indoxyl sulfate	P												
Ioxynil									P				
Isobutylparaben				S				P					
Isopropylparaben								P					
Lenticin		P											
Lidocaine						P							
Lumichrome		P											
Mercaptobenzothiazole			P										
Methionine		P											
Methylperfluorooctanesulfonamidoacetic acid					S					P		S	
Paracetamol						P							
Paraxanthine		P											
Pentachlorophenol									P				S
Perfluoroheptanesulfonic acid					S					P		S	
Perfluorohexanesulfonic acid					S					P		S	
Perfluorooctanesulfonic acid					S					P		S	
Phenol sulfate	P												

Table A4 – (continued) Classification of compounds annotated in the Pélagie samples. P refer to primary uses, and S to secondary uses

	Gut-derived	Food				Health and personal care			Environmental pollutants				
Molecule	**Gut microbiota metabolites**	**Natural compound**	**Flavoring agent**	**Preservatives and other stabilizers**	**Indirect food additive**	**Medication**	**Personal care and cosmetics products**	**Preservatives and other stabilizers**	**Pesticides**	**Plasticizers**	**Organophosphate flame retardant**	**Synthesis intermediate**	**Preservatives and other stabilizers**
Piperidone		P										S	
Piperine		P											
Propylparaben					S								P
Propylparaben sulfate					S								P
Reserpine						P							
Solanidine		P											
Sucralose			P										
Theobromine		P											
Theophylline		P											
Thymol		P						S					
Triclosan glucuronide								S					P
Triclosan sulfate								S					P
Triethylphosphate										S	P	S	
Triphenylphosphine oxide												P	
Tris(2-butoxyethyl)phosphate											P		
Tritosulfuron									P				
Tryptophan		P											

4.6. Presence of annotated compounds on shared suspect lists

Table A5 – Presence of annotated compounds on shared suspect lists

Molecule	Present on shared suspect lists		
	CECscreen (HBM4EU)	Exposome Explorer	NORMAN's SUSDat list
(2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid			X
1,3,5-tris(2,2-dimethylpropionylamino)benzene	X		X
10,11-trans-Dihydroxy-10,11-dihydrocarbamazepine			
13-Hydroxy-7,14-labdadien-6-one			X
2-((3-dodecanamidopropyl)dimethylammonio)acetate			X
2-chlorophenol			
2-hydroxybenzoic acid		X	X
2-hydroxycarbamazepine			
2-Naphthalenesulfonic acid			
2-Naphthol		X	X
2-naphthylamine			X
3-(4-Hydroxyphenyl)lactic acid	X		X
3,5-dibromo-4-hydroxybenzoic acid			X
3-Formylindole	X		X
3-hydroxybenzoic acid			
4-chlorophenol		X	X
4-hydroxy-2,5,6-trichloroisophthalonitrile			X
4-hydroxybenzoic acid			
4-quinolone			X
5-acetylsalicylamide	X		X
5-hydroxytryptophan			X
Acesulfame			
Acetaminophen glucuronide	X		X
Acetaminophen sulfate	X		X
Aminoacetophenone	X		
Arabinosylhypoxanthine			X
Aspartame			
Auraptene			X
Azelaic acid			X
Benzothiazole			X
Benzothiazole sulfonic acid			X
Benzylbutylphthalate		X	X
Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate			X
Bromoxynil			X
Caffeine		X	X
Carbamazepine			
Carveol			X
Carylophyllene oxide			X
Chavicol sulfate			X
Cinchonidine	X		X
Cinnamaldehyde			
CMPF	X		X
Cocamidopropyl Betaine			X
Coumaraldehyde			X
Coumaric acid			X
Cresol sulfate			X
Di(ethylhexyl) phthalate			X
Dioctyl phthalate		X	X
Diphenylphosphate			X
Diphenylsulfone			X
Docosahexaenoic acid		X	X
Dodecylbenzenesulfonic acid			X
Eicosapentaenoic acid		X	X
Ferulic acid			X
Fipronil sulfone			X
Ibuprofen			X
Indole-3-acetaldehyde			X

Table A5 – (continued) Presence of annotated compounds on shared suspect lists

Molecule	Present on shared suspect lists		
	CECscreen (HBM4EU)	Exposome Explorer	NORMAN's SUSDat list
Indole-3-carbinol			X
Indoxyl sulfate	X		X
Ioxynil			X
Isobutylparaben			
Isopropylparaben			X
Lenticin			X
Lidocaine			X
Lumichrome			
Mercaptobenzothiazole			X
Methionine			X
Methylperfluorooctanesulfonamidoacetic acid			
Paracetamol			X
Paraxanthine		X	X
Pentachlorophenol			
Perfluoroheptanesulfonic acid			
Perfluorohexanesulfonic acid			X
Perfluorooctanesulfonic acid			X
Phenol sulfate			X
Piperidone			X
Piperine			X
Propylparaben		X	X
Propylparaben sulfate			X
Reserpine			X
Solanidine			X
Sucralose			X
Theobromine			X
Theophylline			X
Thymol			X
Triclosan glucuronide			X
Triclosan sulfate			X
Triethylphosphate			X
Triphenylphosphine oxide			X
Tris(2-butoxyethyl)phosphate			X
Tritosulfuron			X
Tryptophan			X

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

EHESP

Titre : Développements méthodologiques pour la caractérisation non-ciblée de l'exposome chimique interne humain dans des études épidémiologiques

Mots clés : Exposome, Analyse non-ciblée, Profilage de suspects, Spectrométrie de masse à haute résolution

Résumé : L'exposition chronique à des mélanges complexes de contaminants chimiques (xénobiotiques) est suspectée de contribuer à la survenue de certaines maladies chroniques. Encouragées par le développement de la spectrométrie de masse à haute résolution (SMHR) et l'émergence du concept d'exposome, des méthodes analytiques non-ciblées commencent à voir le jour pour caractériser l'exposition humaine aux xénobiotiques sans *a priori*. Ces méthodes innovantes pourraient ainsi permettre un changement d'échelle pour identifier de nouveaux facteurs de risque chimiques dans des études épidémiologiques. Ces approches présentent néanmoins plusieurs verrous, en lien, entre autres, avec la présence des contaminants à l'état de trace dans des matrices biologiques. Une optimisation de chaque étape analytique (préparation d'échantillon) et bio-informatique (prétraitement des données, annotation) est donc indispensable pour surmonter ces limites. L'objectif principal de ce travail est d'implémenter un workflow non-ciblé applicable aux études épidémiologiques pour apporter une solution opérationnelle à la caractérisation de l'exposome chimique interne à large échelle. Les développements effectués ont permis de proposer un workflow de préparation d'échantillon simple à mettre en œuvre et s'appuyant sur deux méthodes complémentaires pour élargir significativement l'espace chimique visible (jusqu'à 80% de marqueurs spécifiques à une méthode). L'optimisation de logiciels de prétraitement des données, réalisée pour la première fois dans un contexte exposomique, a permis de démontrer la nécessité d'ajuster certains paramètres pour assurer la détection des xénobiotiques à l'état de trace. Le développement d'un logiciel pour automatiser les approches de profilage de suspects avec des prédicteurs MS1, ainsi que le développement d'indices de confiance a permis de prioriser les marqueurs pertinents pour la curation manuelle. Une application à large échelle sur 125 échantillons de sérum de la cohorte Pélagie a permis de démontrer la robustesse et la sensibilité de ce nouveau workflow, ainsi que d'enrichir l'exposome chimique documenté avec la mise en évidence de nouveaux biomarqueurs d'exposition.

Title : Methodological developments for the non-targeted characterization of the human internal chemical exposome in epidemiological studies

Keywords : Exposome, Non-targeted screening, Suspect screening, High-resolution mass spectrometry

Abstract: Chronic exposure to complex mixtures of chemical contaminants (xenobiotics) is suspected to contribute to the onset of chronic diseases. The technological advances high-resolution mass spectrometry (HRMS), as well as the concept of exposome, have set the stage for the development of new non-targeted methods to characterize human exposure to xenobiotics without *a priori*. These innovative approaches may therefore allow changing scale to identify chemical risk factors in epidemiological studies. However, non-targeted approaches are still subject to a number of barriers, partly linked to the presence of these xenobiotics at trace levels in biological matrices. An optimization of every analytical (i.e. sample preparation) and bioinformatical (i.e. data processing, annotation) step of the workflow is thus required. The main objective of this work is to implement an HRMS-based non-targeted workflow applicable to epidemiological studies, to provide an operational solution to characterize the internal chemical exposome at a large scale. The undertaken developments allowed proposing a simple sample preparation workflow based on two complementary methods to expand the visible chemical space (up to 80% of features specific to one method). The optimization of various data processing tools, performed for the first time in an exposomics context, allowed demonstrating the necessity to adjust key parameters to accurately detect xenobiotics. Moreover, the development of a software to automatize suspect screening approaches using MS1 predictors, and of algorithms to compute confidence indices, allowed efficiently prioritizing features for manual curation. A large-scale application of this optimized workflow on 125 serum samples from the Pélagie cohort allowed demonstrating the robustness and sensitivity of this new workflow, and enriching the documented chemical exposome with the uncovering of new biomarkers of exposure.