Recherche de la signature de l'exposition aux xénobiotiques pendant la période périnatale : un reflet de l'exposome chimique périnatal

Dylan Saunier

Université Paris Cité

Université Paris Cité

École doctorale 563 Médicament, Toxicologie, Chimie, Imageries

UMR 0496 MTS, Laboratoire d'Immuno-Allergie Alimentaire (LIAA)

Recherche de la signature de l'exposition aux xénobiotiques pendant la période périnatale : un reflet de l'exposome chimique périnatal

Par **Dylan Saunier**

Thèse de doctorat en **Chimie analytique**

Dirigée par **Estelle Rathahao-Paris**

Présentée et soutenue publiquement

le 10/02/2025

Devant un jury composé de :

Dr. Estelle PUJOS-GUILLOT **Rapporteur**
Ingénieure de recherche HDR, INRAE, Plateforme d'Exploration du Métabolisme, Clermont

Dr. Ronan CARIOU **Rapporteur**
Chercheur HDR, LABERCA, Oniris, Nantes

Pr. Soizic PRADO **Examinatrice**
Professeure, UMR MCAM, CNRS, Muséum national d'Histoire naturelle, Paris

Pr. Arthur DAVID **Examinateur**
Professeur, Université de Rennes, Inserm, EHESP, Irset IRSET, Rennes

Pr. Olivier LAPREVOTE **Examinateur**
Professeur – Praticien hospitalier, CitCoM, Université Paris Cité, Paris

M. Éric VENOT **Invité**
Ingénieure de recherche, INRAE, UMR MTS, Université Paris-Saclay, CEA, INRAE, CEA-Saclay

Dr. Estelle RATHAHAO-PARIS **Directrice de thèse**
Ingénieure de recherche HDR, INRAE, UMR MTS, Université Paris-Saclay, CEA, INRAE, CEA-Saclay

Titre : Recherche de la signature de l'exposition aux xénobiotiques pendant la période périnatale : un reflet de l'exposome chimique périnatal

Mots clés : Exposome, Chimie, Bio-informatique, Xénobiotiques, LC-HRMS, Périnatal

Résumé : De nos jours, la population est largement exposée à des substances chimiques, ou xénobiotiques. Ces expositions, de plus en plus préoccupantes, notamment celles survenues pendant la période périnatale, où l'organisme est particulièrement vulnérable, pourraient avoir des effets à long terme sur la santé des individus. La caractérisation de ces expositions nécessite des méthodes adaptées. Les avancées technologiques ont permis le développement d'outils analytiques de plus en plus performants, comme la chromatographie liquide couplée à la spectrométrie de masse à haute résolution (LC-HRMS), qui contribue à l'amélioration de la détection et de la caractérisation de composés chimiques. Cependant, la taille et la complexité des données produites par ce type d'approche, notamment dans le cadre d'analyses de grandes cohortes, rendent leur exploitation manuelle difficile. Dans ce contexte, mes travaux de thèse ont consisté à développer des stratégies de traitement automatisé des données pour caractériser l'exposome chimique dans des jeux de données métabolomiques produits par LC-HRMS sur des échantillons de méconium (n = 308) et de lait maternel (n = 320) collectés au sein de la cohorte de naissance, EDEN. Deux stratégies ont été développées. La première approche dite de « suspect screening » est semi-ciblée et consiste à rechercher les signaux générés à partir d'une liste de molécules dont la présence dans les matrices analysées est suspectée. Cette stratégie comprend la prédiction des métabolites des molécules suspectées et la recherche automatique de leurs signaux dans les jeux de données LC-HRMS. L'évaluation de la méthode de prédiction des métabolites a été réalisée en comparant les structures des métabolites générés par notre approche avec les profils métaboliques référencés dans la base de données SMPDB (The Small Molecule Pathway Database) pour 60 composés exogènes. Environ 70 % des métabolites ont été correctement prédits dans notre étude et, tous les métabolites ont été correctement produits pour 50 % des composés. Des marqueurs potentiels d'exposition à un certain nombre de xénobiotiques ont pu être détectés dans le méconium et le lait maternel. L'examen de leurs défauts de masse a permis de renforcer l'identité des métabolites qui présentent des défauts de masse différents mais proches de celui du parent avec un écart plus important pour les métabolites conjugués que les non conjugués. La deuxième approche non ciblée a été développée pour extraire, sans a priori, les signaux potentiels de marqueurs d'exposition. Cela comprend : i) l'enrichissement isotopique de la matrice de données (signaux spécifique des isotopes $^{12}C/^{13}C$, $^{79}Br/^{81}Br$, $^{35}Cl/^{37}Cl$ ou $^{32}S/^{34}S$), ii) la filtration spécifique de paires de signaux correspondant à des couples de métabolites conjugués et non conjugués et, iii) l'attribution de la fréquence de détection de chaque signal sur l'ensemble des échantillons. L'application de notre stratégie au jeu de données issus d'échantillons de méconium a permis de réduire sa taille d'un facteur six (de 155 047 à 25 276 signaux). La présence d'espèces halogénées et de couples de métabolites conjugués/non conjugués a pu être révélée. Grâce à l'annotation automatique intégrant les informations issues de l'enrichissement isotopique, le nombre de candidats possibles a pu être réduit. Des marqueurs d'exposition aux xénobiotiques courants, tels que le paracétamol, la caféine et la nicotine, ont été détectés avec succès, démontrant le potentiel du méconium pour révéler une exposition in utero. Cette étude a permis de mettre en évidence une exposition très précoce, via l'exposition maternelle pendant la grossesse. Les stratégies proposées apparaissent très prometteuses pour l'étude de l'exposome chimique dans de grands jeux de données LC-HRMS non ciblés. Ces approches pourraient également être utilisées pour traiter les données issues de n'importe quelle matrice.

Title : Searching for the Signature of Xenobiotic Exposure During the Perinatal Period: A Reflection of the Perinatal Chemical Exposome

Keywords : Exposome, Chemistry, Bioinformatics, Xenobiotics, LC-HRMS, Perinatal

Abstract : Nowadays, the population is widely exposed to chemicals, or xenobiotics. These increasingly concerning exposures, particularly those occurring during the perinatal period when the organism is especially vulnerable, may have long-term effects on individual health. Appropriate methods are needed to characterizing these exposures. Technological advances have enabled the development of increasingly powerful analytical tools, such as liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS), which improves the detection and characterization of chemicals. However, the size and complexity of the data produced by this approach, especially in the context of large cohort analyses, make manual data mining challenging. In this context, my research focused on developing automated data processing strategies to characterize the chemical exposome in metabolomic datasets generated by LC-HRMS on meconium (n = 308) and breast milk (n = 320) samples collected in the EDEN birth cohort. Two strategies were developed. The first, a semi-targeted "suspect screening" approach, involves searching for signals generated from a list of suspected molecules likely to be present in the analyzed matrices. This strategy includes predicting metabolites of the suspected molecules and automatically searching for their signals in LC-HRMS datasets. The evaluation of the metabolite prediction method was conducted by comparing the structures of the metabolites generated by our approach with the metabolic profiles referenced in the Small Molecule Pathway Database (SMPDB) for 60 exogenous compounds. Approximately 70% of the metabolites were correctly predicted in our study, and all metabolites were accurately generated for 50% of the compounds. Potential markers of exposure to several xenobiotics were detected in meconium and breast milk. Examination of their mass defects reinforced the identity of the metabolites, which displayed mass defect values different from but close to those of the parent compounds, with larger deviations for conjugated metabolites compared to non-conjugated ones. The second, untargeted approach was developed to extract potential markers of exposure without prior assumptions. This included: i) isotopic enrichment of the data matrix (specific signals of the isotopes $^{12}C/^{13}C$, $^{79}Br/^{81}Br$, $^{35}Cl/^{37}Cl$, or $^{32}S/^{34}S$), ii) specific filtering of signal pairs corresponding to conjugated and non-conjugated metabolite pairs, and iii) assignment of a detection frequency to each signal across all samples. Applying this strategy to the dataset from meconium samples resulted in a sixfold reduction in data size (from 155,047 to 25,276 signals). The presence of halogenated species and of the conjugated and non-conjugated metabolite pairs was revealed. The number of possible candidates obtained in the annotation step was reduced by incorporating information from isotopic enrichment. Interestingly, markers of exposure to common xenobiotics, such as acetaminophen, caffeine, and nicotine, were successfully detected, demonstrating the potential of meconium to reveal in utero exposure. This study highlighted very early exposure via maternal exposure during pregnancy. The two proposed strategies are complementary and appear to be very promising for the study of the chemical exposome in large untargeted LC-HRMS datasets. These approaches could also be applied to process data from any type of matrix.

Remerciements

Ce travail n'aurait pas pu être effectué sans le soutien d'un grand nombre de personnes, tant dans les sphères professionnelles que personnelles, auxquelles je souhaite adresser mes remerciements pour leur implication dans sa réalisation.

Tout d'abord, je souhaite remercier les membres de mon jury de thèse pour avoir accepté de juger mon travail doctoral, et tout particulièrement **le Docteur Ronan Cariou et le Docteur Estelle Pujos-Guillot** d'avoir accepté d'être les rapporteurs de ma thèse. Je tiens également à remercier **le Professeur Soizic Prado**, **le Professeur Arthur David** et le **Professeur Olivier Laprévote** d'avoir accepté d'examiner mon travail de thèse

Je souhaite également exprimer ma profonde gratitude à ma directrice de thèse, **Estelle Rathahao-Paris**, pour m'avoir offert l'opportunité de travailler sur ce sujet. Tout au long de ces trois années, nos échanges hebdomadaires enrichissants, ainsi que sa pédagogie et son soutien, m'ont permis de mener à bien ce travail doctoral. De plus, ses relectures et son engagement lors de la rédaction du manuscrit ont été d'une aide précieuse.

Je souhaite également adresser mes remerciements à **Eric Venot** pour avoir co-encadré ma thèse, m'avoir initié au traitement des données et, plus largement, à la programmation. Sa disponibilité, son écoute et sa bienveillance m'ont été précieuses tout au long de ces trois années.

Au cours de mes trois années de thèse, j'ai travaillé au sein de l'équipe « Laboratoire d'Immuno-Allergie Alimentaire » (LIAA), au sein du département Médicament et Technologies pour la Santé du CEA à Saclay. Je souhaite exprimer ma profonde gratitude au **Docteur Karine Adel-Patient** pour m'avoir accueillie dans son laboratoire et pour son implication dans le cadre de mon projet de thèse. Je joins à ces remerciements le **Docteur François Fenaille**, directeur du laboratoire Innovations en Spectrométrie de Masse (LI-MS), pour sa participation à de nombreuses discussions enrichissantes autour de mon projet de thèse, sa disponibilité et son écoute durant ces trois années, mais également aux **Docteurs Annelaure Damont et Benoît Colsch** pour leur aide lors de la préparation de communications orales et le partage de

données me permettant de valider les méthodologies développées dans le cadre de mes travaux.

Je tiens également à remercier le **Docteur Etienne Thevenot** pour m'avoir donné l'opportunité d'intégrer l'équipe de science des données au sein du LI-MS, ce qui m'a permis d'enrichir mes compétences en traitement de données. Je lui suis aussi reconnaissant pour son engagement actif dans mes travaux. Par ailleurs, j'adresse un remerciement tout particulier au **Docteur Sylvain Dechaumet** pour nos nombreux échanges, dont la bienveillance et l'expérience m'ont apporté un regard neuf, tant sur le plan professionnel que personnel.

J'adresse également mes remerciements aux membres de mon comité de thèse : **le Professeur Carlos Afonso** et **le Docteur Héloïse Dossmann**, qui m'ont apporté un regard externe et expert sur mes travaux. Les échanges scientifiques m'ont permis de faire le point et de prendre du recul sur mon travail.

Je tiens à remercier la **Fondation pour la Recherche Médicale** (FRM) pour le financement de ma thèse, ainsi que pour la prise en charge des formations et la participation aux congrès.

Je souhaite également exprimer ma gratitude à mes collègues et amis Paloma, Eva, Nina, Justine, Romain, Marie Yann, Anaïs, Ylane, Vincent et Clément pour leur soutien et leur amitié tout au long de ces trois années, ainsi que pour celles à venir.

Mes plus profonds remerciements vont à ma mère, mon frère et mes amis les plus proches pour leur soutien tout au long de ces années. Un merci tout particulier à ma mère, qui m'a élevée avec amour et m'a transmis les valeurs qui me définissent aujourd'hui. Une pensée spéciale également pour Alexandre et Flavien, avec qui j'ai partagé dix années d'entraide et de complicité.

Enfin, je tiens à remercier mon meilleur ami, Quentin, sans qui je ne serais jamais devenu la personne que je suis aujourd'hui et sans qui cette aventure n'aurait jamais vu le jour. Merci d'être toujours là, de veiller sur moi et de me comprendre sans avoir besoin de mots. Je tiendrai notre promesse.

Table des matières

Liste des abréviations

APCI : Atmospheric Pressure Chemical Ionization (Ionisation chimique à pression atmosphérique)

APPI : Atmospheric Pressure Photoionization (Photoionisation à pression atmosphérique)

CCS : Section Efficace de Collision

CEA-Saclay : Commissariat à l'Énergie Atomique et aux Énergies Alternatives, site de Saclay

DIMS : Direct Introduction Mass Spectrometry (Spectrométrie de masse par introduction directe)

DOHaD : Developmental Origins of Health and Disease (Origines développementales de la santé et des maladies)

EDEN : Étude des Déterminants pré et post natals du développement et de la santé de l'Enfant

ELFE : Étude Longitudinale Française depuis l'Enfance

ESI : Electrospray Ionization (Ionisation par électronébulisation)

EXHES : EXPOsOMICS Health Effects Study

EXPOsOMICS : Exposure and Health Effects of the European Exposome

FT-ICR-MS : Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (Spectrométrie de masse à résonance cyclonique ionique)

FTICR : Fourier Transform Ion Cyclotron Resonance

FRM : Fondation pour la Recherche Médicale

GC-MS : Gas chromatography–mass spectrometry (Chromatographie Gazeuse couplée Spectromètre de Masse)

GSH : Glutathion

HEALS : Health and Environment-wide Associations based on Large Population Surveys

HELIX : Human Early-Life Exposome

HILIC : Hydrophilic interaction chromatography (Chromatographie à interaction hydrophile)

HMDB : Human Metabolome Database

HRMS : High-Resolution Mass Spectrometry (Spectrométrie de masse à haute résolution)

IE : Ionisation électronique

IM : Ion Mobility (Mobilité Ionique)

IM-MS : Ion Mobility Mass Spectrometry (spectrométrie de masse à mobilité ionique

LC-HRMS : Liquid Chromatography-High Resolution Mass Spectrometry (Chromatographie liquide couplée à la spectrométrie de masse à haute résolution)

LC-MS : Liquid Chromatography-Mass Spectrometry (Chromatographie liquide couplée à la spectrométrie de masse)

LOESS : Locally Estimated Scatterplot Smoothing

MS : Mass Spectrometry (Spectrométrie de masse)

MS/MS : Spectrométrie de masse en tandem

PÉLAGIE : Perturbateurs Endocriniens : Étude Longitudinale sur les Anomalies de la Grossesse, l'Infertilité et l'Enfance

PARC : Partnership for the Assessment of Risks from Chemicals

ppm : Parties par million

QuEChUp : Variante combinée des méthodes QuEChERS et QuPPe

QuEChERS : Quick, Easy, Cheap, Effective, Rugged, and Safe)

QuPPe : Quick Polar Pesticides extraction

Rp : Pouvoir de résolution

RMN : Nuclear Magnetic Resonance (Résonance magnétique nucléaire)

SMPDB : The Small Molecule Pathway Database

SPE : Solid Phase Extraction (Extraction en phase solide)

SMARTS : SMiles ARbitrary Target Specification

SMILES : Simplified Molecular Input Line Entry System

SMIRKS : SMiles and Reaction Kinetics Smart

TOF : Time of Flight

UHPLC : Ultra High Performance Liquid Chromatography (Chromatographie liquide à ultra haute performance)

logP : Logarithme du coefficient de partage

logS : Logarithme de la solubilité dans l'eau

BMI : Body Mass Index (Indice de Masse Corporelle)

KMD : Défaut de masse de Kendrick

FWHM : Full Width at Half Maximum

rpm : Révolutions par minute (vitesse de rotation)

Q-ToF : Quadrupole Time of Flight

Publication

Dylan Saunier, Éric Venot, Sylvain Dechaumet, Blanche Guillon, Florence Castelli, Etienne Thevenot, François Fenaille, Blandine de Lauzon-Guillain, Karine Adel-Patient, Estelle Rathahao-Paris. Exploration of chemical exposome in a mother-child cohort using a non-targeted data filtering strategy applied to a large and complex LC-HRMS dataset. Article soumis au journal « Analytica Chimica Acta »

Communications orales

Dylan Saunier, Éric Venot, Sylvain Dechaumet, Blanche Guillon, Florence Castelli, Etienne Thevenot, François Fenaille, Blandine de Lauzon-Guillain, Karine Adel-Patient, Estelle Rathahao-Paris. Workflow de traitement de données automatique pour la détection et la caractérisation de xénobiotiques à partir de données produites par LC-HRMS. 16ieme journées scientifiques de la RFMF. **16èmes Journées Scientifiques du Réseau Francophone de Métabolomique et de Fluxomique**, Saint-Malo, 4-6 juin 2024.

Dylan Saunier, Eric Venot, Estelle Rathahao-Paris. Automated approaches for processing data acquired by liquid chromatography coupled with high resolution mass spectrometry (LC-HRMS) to characterize the chemical exposome. **Colloque Doc'J**, INRAE, Jouy-en-Josas, 16 mai 2024.

Dylan Saunier, Eric Venot, Estelle Rathahao-Paris. Search for the signature of exposure to food contaminants during the perinatal period: a reflect of the perinatal chemical exposome ? **Journée des doctorants DMTS**, CEA de Saclay, 24 janvier 2023.

Dylan Saunier, Search for the signature of exposure to food contaminants during perinatal period: a reflection of the perinatal chemical exposome ? **Séminaire invité** du Laboratoire de Chimie Structurale Organique et Biologique, IPCM, Sorbonne université, 31 mai 2022

Communications par affiche.

Dylan Saunier, Éric Venot, Sylvain Dechaumet, Blanche Guillon, Florence Castelli, Etienne Thevenot, François Fenaille, Blandine de Lauzon-Guillain, Karine Adel-Patient, Estelle Rathahao-Paris. Development of an automatic search tool for food chemical exposure markers to characterize the perinatal chemical exposome. **Journée des doctorants DMTS**, CEA de Saclay, 16 janvier 2024.

Dylan Saunier, Éric Venot, Sylvain Dechaumet, Blanche Guillon, Florence Castelli, Etienne Thevenot, François Fenaille, Blandine de Lauzon-Guillain, Karine Adel-Patient, Estelle Rathahao-Paris. Automatic search for chemical exposure markers in LC-HRMS metabolomic data. **15ème Journées Scientifiques du Réseau Francophone de Métabolomique et de Fluxomique**, Perpignan, 24-26 mai 2023.

Dylan Saunier, Éric Venot, Sylvain Dechaumet, Blanche Guillon, Florence Castelli, Etienne Thevenot, François Fenaille, Blandine de Lauzon-Guillain, Karine Adel-Patient, Estelle Rathahao-Paris. Development of an automatic search tool for food chemical exposure markers to characterize the perinatal chemical exposome. **Journées scientifiques de l'école doctorale MTCI**, Le Mée-sur-Seine, 16-17 mars 2023.

Dylan Saunier, Éric Venot, Sylvain Dechaumet, Blanche Guillon, Florence Castelli, Etienne Thevenot, François Fenaille, Blandine de Lauzon-Guillain, Karine Adel-Patient, Estelle Rathahao-Paris. Exploration of the perinatal exposome by automatic search for markers of exposure to chemical food contaminants in a biological matrix (meconium). **Analytics**, Nantes, 5-8 septembre 2022.

Liste des figures

Liste des figures des Chapitres 1 et 2

Liste des figures du Chapitre 3

Liste des tableaux

Liste des tableaux des Chapitres 1 et 2

Liste des tableaux du Chapitres 3

INTRODUCTION GENERALE

De nos jours, l'être humain est constamment exposé à des xénobiotiques (substances chimiques étrangères à l'organisme) provenant de sources diverses, qu'elles soient alimentaires, thérapeutiques, environnementales ou issues de produits stupéfiants ou récréatifs, comme le tabac. Cette exposition est de plus en plus préoccupante, en raison du nombre important de molécules et mélanges chimiques synthétiques enregistrés sur le marché des substances chimiques (plus de 350 000 en 2020)[1] et susceptibles de contaminer la population. Une étude récente de Geueke *et al*.[2] a montré des preuves d'exposition à 3 601 substances chimiques introduites dans notre organisme *via* l'alimentation, en particulier les emballages alimentaires.

La caractérisation de ces xénobiotiques ou indicateurs reflétant l'exposition à ces xénobiotiques chez l'Homme a suscité de nombreuses investigations de la part de la communauté scientifique ces dernières années.[3,4] Néanmoins, ces études sont bien souvent limitées à un nombre restreint de xénobiotiques pour des raisons analytiques propres à chaque technique. Parmi celles-ci, nous pouvons citer la chromatographie liquide couplée à la spectrométrie de masse (LC-MS) et la résonance magnétique nucléaire (RMN) comme les plus utilisées. Dans un contexte de caractérisation structurale, la RMN à travers ses propriétés présente un avantage certain en tant que technique non destructive et non invasive. Cependant, elle est moins performante en termes de sensibilité que la LC-MS, qui offre une couverture plus large permettant la détection d'une plus grande variété de composés chimiques.

Grâce à la haute sensibilité et à la haute résolution apportées par la spectrométrie de masse à haute résolution (HRMS) ces dernières années, des altérations du réseau métabolique ont pu être observées chez plusieurs groupes d'individus (maladies, géographie etc.), notamment à travers des études métabolomiques, et des efforts qui ont pu se concentrer sur la caractérisation des biomarqueurs, c'est-à-dire des variables ou caractéristiques significativement différentes selon les groupes d'individus.

Bien que cette stratégie analytique permette la caractérisation de certains biomarqueurs, il est difficile d'apprécier l'ensemble des composés chimiques détectés car le traitement statistique entraîne inévitablement une perte d'informations lors du filtrage des données.

Trouver des signaux relatifs à l'ensemble des molécules chimiques d'intérêt dans les données issues de la chromatographie liquide couplée à la spectrométrie de masse à haute résolution (LC-HRMS) s'avère complexe en raison de la présence d'un grand nombre de signaux. Cette tâche manuelle devient impossible dans le cas des études de cohortes comportant un grand nombre d'échantillons et ne peut être réalisée qu'à l'aide d'outils bio-informatiques dédiés. Cependant, plusieurs aspects restent encore difficiles à gérer, notamment les signaux artefactuels, qui sont présents en quantité non négligeable dans les données HRMS ainsi que la détection possible d'un grand nombre de faux-positifs. Dans ce type d'approches, il est donc nécessaire de mettre en place des outils spécifiques de filtration des données fondés sur des caractéristiques inhérentes aux espèces telles que leur nature organique.

De plus, les algorithmes de traitement du signal disponibles pour gérer un volume aussi important de données présentent eux aussi des difficultés qui peuvent impacter l'efficacité des outils dédiés à la fouille de données. Pour s'en affranchir, il est recommandé de travailler au plus près des données brutes. Dans un contexte plus général, la caractérisation des expositions aux xénobiotiques est également complexe en raison des nombreuses transformations que ces xénobiotiques peuvent subir depuis leur synthèse jusqu'à leur administration dans un organisme (activité microbienne, métabolique, environnementale). Il est donc nécessaire de prendre en compte ces transformations pour étayer la recherche des marqueurs d'exposition aux xénobiotiques.

Le choix des échantillons à étudier est également essentiel pour caractériser l'exposition aux xénobiotiques compte tenu du renouvellement continu des fluides biologiques et des transformations que peuvent subir des molécules chimiques dans l'organisme. En effet, chaque matrice biologique agrège, dans une fenêtre d'exposition qui lui est propre, la présence des xénobiotiques donnés et/ou de leurs métabolites, présence qui dépend de la biodistribution des molécules mères. Cette situation peut poser problème en pratique. Bien qu'une matrice biologique puisse fournir des informations sur l'exposition d'un individu à une ou plusieurs substances chimiques, elle est limitée par le nombre de composés qu'elle accumule et par la fenêtre d'exposition qu'elle enregistre.

La fenêtre d'exposition choisie pour comprendre l'impact de ces substances sur la santé humaine est également remise en question au sein de la communauté scientifique. En effet, des expositions durant la période périnatale, où de nombreux systèmes biologiques n'ont pas encore atteint leur maturité fonctionnelle, peuvent avoir un impact plus important sur la santé lors du développement ultérieur de l'individu.

C'est dans ce contexte de l'exposome périnatal, en particulier celui chimique, que s'inscrit mon travail de thèse. Cette thèse fait partie du projet PeriContAll financé par la Fondation pour la Recherche Médicale (FRM), qui vise à déterminer les facteurs favorisant l'apparition d'allergies durant l'enfance. Dans mon projet doctoral, deux stratégies, une de type « suspect screening » et une approche sans *a priori*, ont été envisagées pour la fouille automatique de grands jeux de données acquises de manière non ciblée par LC-HRMS afin de détecter des signaux de marqueurs d'expositions aux xénobiotiques au sein de matrices périnatales, telles que le méconium et le lait maternel, collectées dans la cohorte française de couples mères-enfants, l'étude EDEN (Étude des Déterminants pré et post natals du développement et de la santé de l'Enfant).

Mes travaux de thèse sont présentés dans ce manuscrit sous la forme de trois chapitres.

Le premier chapitre est une synthèse bibliographique consacrée à l'exposome. Il aborde les différents aspects de l'exposome ainsi que les approches analytiques envisagées depuis le choix de la matrice à étudier, la préparation des échantillons et l'acquisition des données jusqu'à l'analyse des données pour explorer l'exposome périnatal, notamment chimique.

Le deuxième chapitre est axé sur la stratégie « suspect screening », qui consiste à rechercher de façon automatique des marqueurs de molécules « suspectées ». Il présente le développement d'une stratégie de traitement des données automatisée pour la recherche et la détection automatique et semi-ciblée de marqueurs d'expositions à partir d'une liste prédéfinie de composés chimiques au sein d'un grand jeu de données métabolomiques acquis par LC-HRMS. Cette stratégie comprend également la prédiction de métabolites basée sur une sélection de réactions de biotransformation connues. L'efficacité de notre approche dans la prédiction des métabolites a pu être évaluée par comparaison avec les profils métaboliques de 60 composés exogènes référencés dans la base de données SMPDB (The Small Molecule Pathway Database). L'approche « suspect screening » appliquée à nos jeux de données métabolomiques issus de méconiums et de lait maternel de la cohorte EDEN a permis de

détecter de potentiels marqueurs d'exposition à certains xénobiotiques grâce à la présence de massifs isotopiques spécifiques qui contribuent fortement à l'attribution de leur identité structurale.

Le chapitre trois est focalisé sur la stratégie d'analyse de données sans *a priori* pour explorer de façon globale le xénométabolome (métabolites d'origine exogène) dans l'objectif de détecter des signaux de marqueurs d'exposition aux xénobiotiques non répertoriés dans la liste prédéfinie, comme dans l'approche « suspect screening ». Cette approche non ciblée implique le développement de plusieurs filtres de données fondés sur les propriétés isotopiques, la biotransformation des xénobiotiques (glucuroconjugaison, sulfoconjugaison, conjugaison au glutathion) et la présence de signaux récurrents dans les échantillons. L'annotation de certains marqueurs potentiels d'exposition, qui combine l'algorithme de MassTools[5] pour la prédiction de formule chimique et les caractéristiques des massifs isotopiques détectés, a permis de réduire considérablement le nombre de candidats possibles. Ces travaux sont présentés en anglais selon la structure d'un article scientifique.

Chapitre 1. Le bien-fondé de l'exposome à la caractérisation de l'exposome chimique périnatal

1.1. Exposome dans un contexte périnatal

L'exposome a été introduit pour la première fois en 2005 par Christopher Paul Wild[6] pour désigner la totalité des expositions, à savoir toutes les sources non génétiques, auxquelles un individu est soumis tout au long de sa vie, depuis sa conception jusqu'à la fin de sa vie.

1.1.1. Exposome et ses différentes catégories

En 2012, Wild a élargi sa définition[7] en subdivisant l'exposome en trois catégories (Figure 1-1) :

i) **L'exposome externe**, qui prend en compte l'exposition à des facteurs environnementaux spécifiques aux individus, incluant les éléments physiques, chimiques et biologiques (polluants, maladies infectieuses, interventions médicales, bruits, comportements individuel, etc.),

ii) **L'exposome externe général** associé aux facteurs environnementaux globaux et contextuels incluant les conditions socio-économiques, psychologiques et climatiques (éducation, profession, revenu, températures extrêmes etc.), et

iii) **L'exposome interne** en lien avec les caractéristiques biologiques propres à chaque individu (hormones, produits du métabolisme, etc.) et leurs réponses face aux facteurs environnementaux.

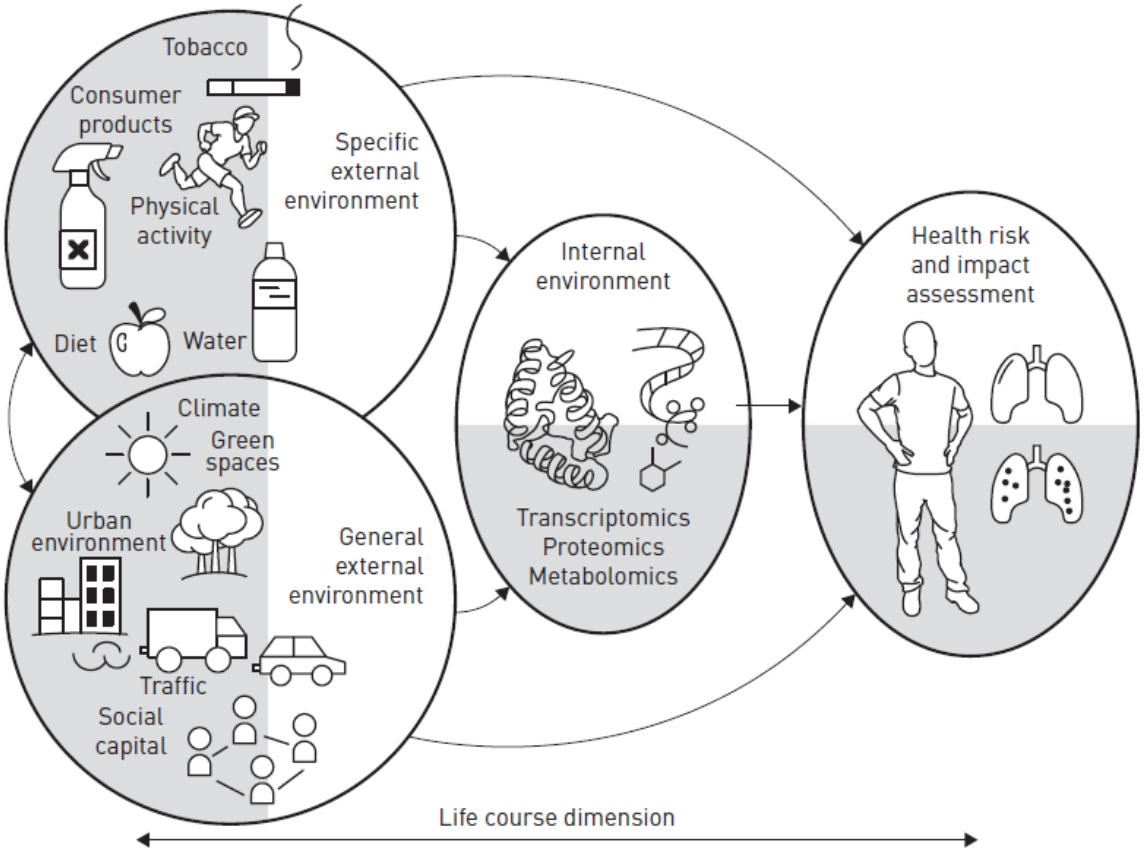


Figure 1-1 : Différents domaines constituant l'exposome : environnement externe spécifique ou général, environnement interne et le risque pour la santé.[8]

L'appropriation de cette définition par la communauté scientifique a conduit à une dernière évolution. En 2014, Miller & Jones ont notamment redéfini l'exposome comme « la mesure cumulative des influences environnementales et des réponses biologiques associées tout au long de la vie, y compris les expositions dues à l'environnement, à l'alimentation, au comportement des individus et aux processus endogènes ».[9]

Ainsi, l'exposome intègre des influences non génétiques sur les organismes vivants et pourrait être le maillon manquant qui aiderait à expliquer le développement de certaines maladies. Par exemple, un génotype spécifique, c'est-à-dire l'ensemble des gènes d'un organisme, peut présenter un risque basal pour développer un cancer du poumon, mais l'exposome lié à la pollution de l'air, au tabagisme, etc., pourrait sensiblement accroître ce risque.

Il est important de noter que même si le terme « exposome » n'est apparu qu'en 2005, de nombreux scientifiques avaient déjà observé les effets de l'environnement sur la santé bien avant l'apparition de ce concept. En effet, en 1968 lors de la de conférence internationale sur

le cancer en Israël, Higginson[10], directeur du centre international de recherche sur le cancer, rapportait déjà que les facteurs environnementaux (externes) pourraient être à l'origine de 80 % des cas de cancer. Par la suite, certaines études ont confirmé ces résultats. Czene *et al.*[11] ont ainsi observé en 2002 que l'apparition de nombreux cancers ne résultait pas seulement de la contribution génétique, mais également de facteurs environnementaux, dont une part représentative est supérieure à 70 %, à l'exception du cancer de la thyroïde.

La période d'exposition influence de manière significative l'impact sur la santé. Ainsi, dans les années 1980, Barker et Osmond ont émis l'hypothèse que des conditions de vie défavorables, notamment la malnutrition pendant la période de développement fœtal, pouvaient accroître le risque de développer des maladies cardiovasculaires tout au long de la vie.[12] L'évolution de cette hypothèse par la communauté scientifique a conduit, dans les années 2000, au concept DOHaD[13] (Developmental Origins of Health and Disease), qui se traduit par « origines développementales de la santé et des maladies ». Ce concept identifie la période des 1000 jours, c'est-à-dire les 1000 premiers jours de vie, depuis la conception jusqu'à l'âge de 2 ans, comme une fenêtre critique durant laquelle l'organisme humain en plein développement est particulièrement sensible aux influences environnementales qui ont un impact majeur sur sa santé future. De nombreuses études ont montré que l'exposition à certains facteurs environnementaux et comportementaux (l'alimentation, le stress et les expositions aux toxines) durant cette période sensible peut induire des changements épigénétiques, affectant ainsi l'expression des gènes tout au long de la vie. De plus, des maladies chroniques comme l'hypertension, le diabète et l'obésité peuvent avoir pour origine les expositions durant cette période périnatale.[14] Dans le contexte de l'exposome chimique, plusieurs études ont révélé des associations significatives entre l'exposition durant la période périnatale à certains xénobiotiques par l'intermédiaire de la mère (tabac, alcool et médicaments) et le développement futur de maladies liées à la consommation d'alcool[15] (dysmorphie cranio-faciale, anomalies sensorielles et neuropsychologiques, déficits neurologiques, etc.) ou de tabac[16] (conséquences neurocomportementales, métaboliques, cardiovasculaires et respiratoires).

1.1.2. Exposome et études de cohortes

L'intérêt pour le concept des 1000 jours a donné naissance à des études de cohortes mères-enfants (Tableau 1-1) visant à évaluer l'impact de différents facteurs auxquels un être humain est soumis durant cette période et qui peuvent avoir un impact sur son développement et sa santé à l'enfance et plus tard.

En France, on peut citer l'étude EDEN[17], première cohorte généraliste mères-enfants. Son objectif est d'établir les déterminants précoces de la santé et du développement de l'enfant, en prenant en compte plus particulièrement les facteurs environnementaux déclarés. Il s'agit d'une étude épidémiologique longitudinale qui consiste à suivre dans le temps des enfants depuis la fin du premier trimestre de grossesse jusqu'à l'âge de 5 ans. Ce suivi est fondé sur les informations recueillies auprès des parents sous forme de questionnaires, mais également sur les matrices biologiques collectées. Cette étude comprend 2002 femmes enceintes recrutées entre 2003 et 2006 sur deux sites, à Nancy et à Poitiers. La majorité des familles ont accepté un suivi au-delà de 5 ans. Ainsi l'étude se poursuit jusqu'à l'âge de 8 ou 10 ans des enfants. En 2002, l'étude de la cohorte PÉLAGIE[18] (Perturbateurs Endocriniens : Étude Longitudinale sur les Anomalies de la Grossesse, l'Infertilité et l'Enfance) a été menée pour suivre pendant 20 ans près de 3500 mères-enfants en Bretagne. Une autre étude plus récente et de plus grande envergure est l'Etude Longitudinale Française depuis l'Enfance (ELFE)[19]. Celle-ci a pour but de mieux connaître les facteurs (environnement, entourage familial, conditions de vie…) qui peuvent avoir une influence sur le développement physique et psychologique de l'enfant, ainsi que sur sa santé et sa socialisation. Elle est toujours en cours et porte sur le suivi de plus de 18 000 enfants français (nés en 2011), de la naissance à l'âge adulte, soit pendant 20 ans.

A l'échelle européenne, des initiatives similaires sont apparues, telles que les cohortes HELIX[20], EXPOsOMICS[21] et HEALS à travers son étude pilote EXHES[22], mais également des projets plus larges comme HBM4EU[23] et PARC[24] le projet prenant sa relève , ces projets visent notamment à renforcer la santé publique en Europe en établissant des systèmes de biosurveillance humaine pour évaluer l'exposition aux substances chimiques nocives et en développant des méthodologies pour une évaluation intégrée des risques chimiques.

Tableau 1-1 : Quelques études de cohortes visant à caractériser l'Impact de l'environnemental sur la santé des enfants en France ou en Europe.

Etudes	Année (durée)	Objectifs	Nombre de participants	Localisation
EDEN	2003-2022	Comprendre l'impact des premiers facteurs, notamment environnementaux, sur la santé des individus dès l'enfance et tout au long de la vie	2002	France (Poitiers et Nancy)
PELAGIE	2002-en cours	Répondre aux préoccupations de santé des enfants et des adolescents concernant la présence de composés toxiques dans notre environnement quotidien	3421	France (Ille-et-Vilaine, Côtes d'Armor, Finistère)
ELFE	2011-en cours	Déterminer l'exposition des femmes enceintes françaises aux polluants environnementaux durant la grossesse, identifier les modes d'exposition et évaluer les effets sur la santé et le développement de l'enfant	18 000	France (national)
HELIX	2013-2017	Combiner tous les facteurs environnementaux auxquels les mères et les enfants sont exposés durant les premières étapes de la vie, et analyser leur lien avec la santé, la croissance et le développement des enfants	31 472	Europe
EXPOsOMICS	2012-2016	Prédire le risque individuel de maladies liées à l'environnement en caractérisant l'exposome externe et interne pour les expositions courantes (contaminants de l'air et de l'eau potable) pendant les périodes critiques de la vie, y compris in utero	/	Europe
HEALS (EXHES)	2013-2019	Comprendre les facteurs environnementaux et génétiques qui influencent la santé tout au long de la vie à travers des études basées sur la population en Europe, y compris l'examen des enfants et des jumeaux	5000	Europe

Ces études épidémiologiques ont permis de démontrer l'existence d'associations entre le développement de certaines maladies et certains facteurs environnementaux sur la base des informations recueillies.[25][26][27]

Parmi les influences environnementales conditionnant l'exposome, les facteurs chimiques exogènes représentent une contribution importante. En effet, les substances chimiques sont omniprésentes dans notre existence (nourriture, vêtements, santé, environnement) et l'exposition à ces substances est inévitable. L'exposome chimique représente ainsi une sous-catégorie de l'exposome, qui est spécifique à l'exposition aux molécules chimiques exogènes, c'est-à-dire provenant de l'extérieur d'un organisme. Cela comprend l'exposition à la fois à des molécules de synthèse et à des molécules du vivant d'origine végétale ou animale (*via*

l'alimentation ou les produits d'usage quotidien) ; cette exposition dépend du comportement de chaque individu (alimentation, mode de vie, etc.).

Les études sur cohortes s'intéressent également à la détection et la caractérisation des xénobiotiques retrouvés au sein des matrices biologiques collectées. Plusieurs tentatives ont essayé d'évaluer des associations entre les modifications cognitives ou les maladies chez l'enfant et l'exposition aux xénobiotiques. Ces travaux, résumés dans le Tableau 1-2, montrent qu'il est possible de mettre en évidence ces liens grâce à des approches épidémiologiques. Des marqueurs d'exposition ont ainsi pu être directement observés dans les matrices biologiques des individus de la cohorte. Cependant, ces marqueurs sont généralement détectés à l'aide d'analyses ciblées, qui se concentrent sur des familles spécifiques de composés chimiques. Malheureusement, dans certains cas, aucune méthodologie analytique n'est décrite avec précision car les analyses ont probablement été réalisées par un sous-traitant ; ceci constitue une limite rédhibitoire à l'exploitation de ce type d'information.

En raison de la spécificité de ces analyses ciblées, une perte d'informations relative à des composés non recherchés peut survenir lors de la préparation des échantillons. Cela est dû au fait que ces méthodes privilégient la qualité des résultats, notamment la sensibilité de détection pour uniquement des composés d'intérêt spécifique, au détriment de la quantité et de la diversité des autres substances présentes dans la matrice biologique étudiée. Ainsi, seule une partie limitée des xénobiotiques ciblés peut être identifiée.

De plus, dans les projets visant à caractériser de manière exhaustive les xénobiotiques environnementaux, le recours à des analyses ciblées nécessite une grande quantité d'échantillons et l'exécution de multiples analyses. Cela exige souvent le développement de nouvelles méthodes analytiques en interne ou la collaboration avec des laboratoires ayant déjà effectué ce type de recherche méthodologique, ce qui décuple les coûts, à la fois humains et budgétaires.

Tableau 1-2: Analyse par approche ciblée des expositions environnementales et de leur impact sur la santé dans les études de cohorte.

Etudes	Matrices	Molécules ciblées	Lien santé/maladie	Référence
PELAGIE	Urine	Pesticides	Scores d'intelligence à 6 ans	Cartier *et al* [28]
		Insecticides pyréthroïdes	Développement cognitif	Viel *et al* [29]
		Sous-produits de désinfectants	Croissance fœtale et durée de la gestation	Costet *et al* [30]
ELFE	Sang	Polluants organiques, métaux et métalloïdes	/	Berat *et al* [31]
EDEN	Urine	Phénols et phtalates	Santé respiratoire à cinq ans / Poids fœtal et placentaire	Vernet *et al* [32]/phillipat *et al* [33]
HELIX	Sang	Per- et polyfluoroalkylées	Susceptibilité aux lésions hépatiques chez les enfants	Stratakis *et al* [34]
EXPOsOMICS	Eau	Trihalométhanes	Cancer de la vessie	Villanueva *et al* [35]
HEALS (EXHES)	Sang (cordon ombilical et veineux)	Organochloré	Développement physiologique	Grimalt *et al* [36]

Dans la suite de ce manuscrit, seuls les aspects liés à l'exposome chimique seront abordés.

1.2. Marqueurs reflétant l'exposome chimique

1.2.1. Marqueurs d'exposition

Les marqueurs d'exposition sont des indicateurs ou des mesures utilisés pour évaluer l'exposition d'un individu à des facteurs environnementaux ou internes spécifiques. Dans le cas de l'exposome chimique, plus spécifiquement celui des xénobiotiques, un marqueur peut correspondre à une molécule chimique sous sa forme d'origine (molécule mère) ou bien sous l'une de ses formes modifiées (métabolites) par une série de processus biologiques réalisés au sein des organismes vivants. Il faut souligner que ces molécules xénobiotiques peuvent être présentes sous des formes métabolisées et dans des espèces consommées *via* l'alimentation ou dans des produits d'usage courant. Leur concentration dans l'organisme peut ainsi varier en fonction de l'alimentation ou de l'utilisation de certains produits.

Dans ce contexte, il est difficile de déterminer si une molécule est exogène en se basant uniquement sur la preuve de sa détection dans l'organisme, sans quantification précise. Il est nécessaire de comparer les niveaux de concentration de la molécule d'intérêt chez un ensemble d'individus afin de distinguer les personnes particulièrement exposées à des sources externes.

La caractérisation de xénobiotiques au sein d'une matrice biologique constitue un réel défi analytique pour la communauté scientifique. En effet, leur présence en quantité infime dans un mélange complexe nécessite des méthodes analytiques performantes. De plus, une fois introduits dans un organisme, ces xénobiotiques peuvent soit être éliminés par des protéines comme la P-glycoprotéine, soit subir des processus biologiques intrinsèques à chaque organisme (métabolisme) et peuvent ne plus être détectés sous sa forme initiale.

1.2.2. Réactions de biotransformation des xénobiotiques

Le métabolisme des xénobiotiques est un processus de détoxication qui implique diverses réactions chimiques appelées « biotransformations » pour transformer les xénobiotiques en produits plus hydrosolubles pour être plus facilement excrétés par différentes voies (urinaire, fèces, air expiré).[37] Il se produit principalement dans le foie, mais peut avoir lieu dans d'autres organes comme les reins, les poumons et le cerveau chez l'humain. Les biotransformations sont en fait des réactions enzymatiques qui peuvent être classées en deux phases :

i) **Réactions de phase I** dites « de fonctionnalisation » impliquant majoritairement le cytochrome P450, une famille de protéines enzymatiques présentes majoritairement dans le foie. La réaction de fonctionnalisation conduit à l'incorporation des fonctions chimiques *via* des réactions telles que l'oxydation, la réduction ou l'hydrolyse. Les métabolites fonctionnalisés peuvent ensuite subir des processus de métabolisation supplémentaires, tels que les conjugaisons.

ii) **Réactions de phase II** dites « de conjugaison » catalysées principalement par des enzymes cytosoliques présentes dans le cytosol (partie liquide du cytoplasme cellulaire). La molécule-mère et les métabolites produits lors de la phase I peuvent subir des réactions de conjugaison. Les réactions les plus fréquentes sont la glucuroconjugaison et la conjugaison au groupement sulfate et au glutathion.

Ces réactions étant catalysées par des enzymes spécifiques, il est important de souligner que les variations interindividuelles dans l'activité des enzymes du métabolisme et des transports des xénobiotiques peuvent entrainer des différences dans le métabolisme des xénobiotiques et dans leur élimination. En effet, une étude de Nakajima *et al* , a montré qu'un polymorphisme du CYP2A6, une enzyme appartenant à la famille des cytochromes P450, impliquée dans la biotransformation de la nicotine en cotinine, un métabolite majeur de ce xénobiotique, entraînait des différences interindividuelles sur le métabolisme et l'excrétion de la nicotine par l'organisme.[38]

Ces réactions de biotransformations et des différences interindividuelles posent un défi supplémentaire pour la détection et la caractérisation de composés exogènes au sein de matrices biologiques. Il est nécessaire de connaître le métabolisme des xénobiotiques d'intérêt ou au moins leurs métabolites majeurs afin de pouvoir rechercher leurs signaux dans les données analytiques acquises.

La recherche de ces marqueurs d'exposition dans le contexte périnatal est bien plus complexe. En effet, si l'exposition d'un nouveau-né passe par l'interface avec son environnement et, le cas échéant, par la mère *via* l'allaitement, celle du fœtus est limitée à un environnement restreint (liquide amniotique, placenta et cordon ombilical) et par conséquent liée à l'exposition de la mère pendant la gestation. De plus, l'exposition du fœtus ne peut pas être identique à celle de la mère. Plusieurs variables doivent être prises en compte, notamment les processus de biotransformation des xénobiotiques auxquels la mère est exposée mais aussi le rôle respectif du placenta et du cordon ombilical dans les échanges mère-fœtus, qui conditionnent l'apport de nutriments et d'oxygène pour le bon développement du fœtus. Il est particulièrement important de souligner le rôle du placenta, qui agit comme une barrière protectrice pour le fœtus vis-à-vis des substances potentiellement nocives.

Cette fonction protectrice du placenta évolue tout au long de la gestation, ce qui peut altérer les échanges entre la mère et le fœtus, notamment le transfert de xénobiotiques et/ou de leurs métabolites. Cela complique la mise en évidence de l'exposition fœtale à partir de l'exposition maternelle et donc de la définition des « fenêtres d'exposition » réelles. Une étude sur l'exposition à l'ochratoxine A chez des souris à différentes périodes de gestation (de -2 à 16 jours) a montré l'apparition de malformations chez les progénitures de souris exposées au 9^{e} jour.[39] Une étude complémentaire a révélé une forte concentration en ochratoxine A

dans l'utérus de souris gestantes aux jours 8 et 9, suivie d'une diminution au jour 10.[40] Ces études soulignent non seulement l'évolution et l'impact de la fonction barrière du placenta, mais également la complexité quant à la relation entre exposition et développement de pathologies.

Bien que le transfert entre la mère et le fœtus puisse être altéré par l'évolution du placenta notamment, il est de plus en plus admis que les contaminants auxquels la mère est exposée peuvent contaminer le fœtus.[41] En effet, plusieurs études ont déjà mis en évidence le transfert de substances chimiques à travers la barrière placentaire.[40,42]

1.3. Approche(s) analytique(s) pour caractériser l'exposome chimique périnatal

Diverses approches méthodologiques permettent d'aborder les différents aspects de l'exposome (externe et interne). L'exposome chimique peut être évalué de manière indirecte au sein d'études épidémiologiques[43,44] associées ou non à l'analyse de milieux environnementaux, mais ces approches ne permettent pas d'affirmer qu'un individu a été effectivement exposé à un ou des contaminants. En revanche, l'analyse de matrices biologiques humaines[45] peut apporter une preuve directe de la présence ou non de marqueurs d'exposition aux xénobiotiques et ainsi mettre en évidence un lien entre la santé ou la maladie d'un individu et ses expositions.

Dans la suite de ce manuscrit, seules les approches analytiques seront abordées et, plus particulièrement, celles impliquant le couplage LC-MS qui est la méthode la plus utilisée pour effectuer l'analyse de molécules chimiques dans des matrices complexes. Parmi ces approches, on distingue celles dites « ciblées » de celles « non ciblées », la différence résidant dans la caractérisation plus ou moins exhaustive de la présence de composés chimiques « connus » ou « non connus » respectivement, dans une matrice donnée. Quelle que soit l'approche utilisée, les protocoles analytiques sont constitués principalement des étapes suivantes : i) choix de la matrice à étudier, ii) préparation des échantillons, iii) acquisition des données analytiques et iv) traitement des données.

L'approche ciblée consiste à détecter et quantifier une ou plusieurs familles de molécules spécifiques.[46] Dans ce cas, la sélection des matrices dépend de la présence de composés d'intérêt dont la liste peut être définie à partir de la littérature ou à partir d'hypothèses raisonnables, et la procédure de préparation des échantillons est optimisée en fonction des

familles de molécules étudiées. Cette approche implique forcément une réduction du nombre de molécules chimiques de la matrice étudiée puisqu'elle ne cible que les composés d'intérêt afin de limiter l'influence des autres molécules. Cette réduction du nombre de molécules retenues correspond à une pré-concentration, qui facilite la détection des composés d'intérêt et plus souvent à faible concentration. Ce type d'approche ne nécessite pas de développement particulier d'outils d'analyse de données, puisque la régression linéaire est la méthode utilisée pour traiter les données de quantification. Historiquement, cette approche a été utilisée pour des raisons de contraintes techniques mais son intérêt perdure aujourd'hui en raison de sa sélectivité et de sa sensibilité.

L'autre approche, dite « non ciblée », consiste en une analyse sans *a priori*, qui vise à détecter de manière la plus exhaustive possible différentes molécules chimiques au sein des matrices étudiées.[45,47,48] Elle peut servir à « capturer » une empreinte moléculaire caractéristique de chaque échantillon, permettant de remonter à son origine suspectée[49] ou à des conditions atmosphériques spécifiques.[50] Le choix de la matrice à étudier dépend de l'ensemble des molécules chimiques qui y sont présentes et de son lien causal possible avec les objectifs de l'étude. La préparation des échantillons doit minimiser les pertes d'informations afin de couvrir la plus grande diversité moléculaire possible. La différence majeure avec l'approche ciblée réside dans l'acquisition et le traitement des données. Cette approche nécessite une optimisation particulière lors de ces deux étapes pour couvrir l'ensemble des molécules chimiques et détecter les signaux les plus pertinents.

Il existe également **une approche semi-ciblée**, située à mi-chemin entre l'approche ciblée et celle non ciblée.[45] Celle-ci vise à détecter et/ou quantifier une ou plusieurs molécules suspectées, tout en conservant un maximum de l'ensemble des molécules chimiques au sein de la matrice.

La suite du manuscrit détaille les étapes de l'approche non-ciblée ou de celle du « *suspect screening* », dans le cadre de la caractérisation de l'exposome chimique pendant la période périnatale dans des matrices biologiques. Elle aborde plus spécifiquement les problèmes rencontrés lors du traitement des données issues d'un grand nombre de données produites par LC-HRMS.

1.3.1. Matrices utilisées pour caractériser l'exposome chimique périnatal

1.3.1.1. Choix de la matrice à étudier

Les trois principales sources par lesquelles l'organisme peut être exposé aux contaminants chimiques sont l'air, l'eau et l'alimentation.[51] La mesure des xénobiotiques dans ces matrices environnementales permet de caractériser *a priori* l'exposome chimique. Cela repose sur l'hypothèse que les mesures réalisées reflètent bien l'exposition de chaque individu et que, dans le contexte de l'exposome périnatal, les transmissions *in utero* sont systématiques, ce qui n'est pas toujours le cas. En effet, les comportements individuels (sédentarité[52], alimentation[53]) ainsi que les processus biologiques de la mère et les transferts mère-fœtus peuvent largement influer sur l'exposition des individus et l'exposition *in utero*. Ainsi, les mesures d'exposition dans les matrices environnementales ne permettent pas de caractériser avec précision l'exposome chimique individuel et encore moins l'exposome chimique périnatal puisque les paramètres pharmacodynamiques (transfert, biodisponibilité, métabolisation) des différentes molécules de l'exposome chimique ne sont pas caractérisés chez la mère durant la grossesse.

Pourtant, bien que les analyses de ces matrices environnementales présentent des limites en termes d'estimation du niveau réel d'exposition d'un individu, elles restent essentielles pour évaluer le risque d'exposition chimique à l'échelle d'une population et jouent donc un rôle clé dans la surveillance sanitaire et la prévention des risques environnementaux.

L'analyse des matrices biologiques (sang, urine, etc.) permet de fournir directement des preuves d'une exposition chimique, mais cela nécessite des procédures de préparation des échantillons spécifiques à chaque type de matrice et à chaque molécule ou famille de molécules, ainsi qu'une connaissance plus précise des phénomènes de biotransformation et de biodistribution afin de ne pas passer à côté des xénobiotiques présents en faibles concentrations et subissant une métabolisation et, de trouver des marqueurs d'exposition dans les matrices sélectionnées. Dans un contexte de l'exposome chimique périnatal, le choix de la matrice biologique à étudier est crucial. En effet, au-delà de l'aspect invasif qui reste intrinsèque au prélèvement des échantillons, chaque matrice présente des caractéristiques spécifiques qui peuvent influencer les résultats obtenus après analyse, telles que la variabilité de l'ensemble des différentes molécules chimiques et les fenêtres d'exposition au cours desquelles les xénobiotiques ou leurs métabolites peuvent être détectés après exposition.

Le Tableau 1-3 présente une liste non exhaustive de matrices biologiques pouvant être étudiées dans le cadre de l'exposome chimique périnatal (fenêtre d'exposition, nature du caractère invasif, composés ou familles chimiques détectés).

Une revue sur l'exposome a montré que les matrices les plus étudiées par LC-HRMS d'après 124 articles scientifiques sont le sang et l'urine avec un taux de 46 % (57/124) pour le sérum ou le plasma, et de 41 %, (51/124) pour l'urine. Une plus faible proportion d'études (16 sur 124) s'est concentrée sur d'autres matrices, telles que le lait maternel, l'haleine, la salive, le foie, le placenta, les cheveux etc.[54] D'après les auteurs, le choix du sang et de l'urine s'expliquerait par leur caractère homogène ou facilement homogénéisable. Cette préférence manifestée par la communauté scientifique, bien qu'elle permette de renforcer les informations obtenues sur ce type de matrices avec l'identification de métabolites ou de marqueurs d'expositions, ne permet pas d'apprécier une réelle exposition sur une longue période. De plus, les informations contenues dans la littérature sur l'exposome chimique pour des matrices plus rarement étudiées sont globalement limitées. Dans un contexte d'étude de l'exposome chimique, il est nécessaire de développer des méthodes analytiques appliquées à ces matrices afin d'explorer davantage l'exposome chimique chez l'humain.

Les matrices les plus proches du fœtus tout au long de la gestation, telles que le méconium, le placenta ou le cordon ombilical, semblent être les plus intéressantes pour étudier l'exposition aux xénobiotiques et d'obtenir un reflet de l'exposome chimique périnatal. Le cordon ombilical, du fait de son interaction entre le fœtus et le placenta, permet d'observer les molécules qui le traversent ou non et donc d'en apprendre davantage sur les transferts de xénobiotiques de la mère à l'enfant. À l'inverse, le placenta accumule les xénobiotiques dont il bloque l'accès au cordon ombilical. Il permet ainsi de détecter et de caractériser des xénobiotiques sur une large fenêtre de temps, mais également d'en apprendre davantage sur la dynamique des transferts. Le méconium, de par ses propriétés d'accumulation durant les six derniers mois de la gestation chez la mère, permet de caractériser les xénobiotiques ayant pu entrer en contact avec le fœtus.[55]

Tableau 1-3 : Matrices biologiques étudiées : fenêtre d'exposition et composés chimiques détectés

Matrices	Fenêtre d'exposition	Caractère invasif ou non	Composés chimiques ou familles de composés détecté(e)s
Urine	Fenêtre restreinte	Non	Insecticides[28], phtalates[3][56], bisphénol A[3][56], pesticides[4][56], médicaments[4], métal[56] Herbicides[56], chlorophénols[56], diakyl phosphates[56], pyrethenoîdes[56], mycotoxines[57]
Serum	Fenêtre restreinte	Oui	PFAS[34], bisphénol A[58][59],POPS[56][60], pesticides[61]
Cheveux	Large fenêtre	Non	Métaux[56], pesticides[62], perturbateur endocrinien[63]
Tissus adipeux	Large fenêtre	Oui	POPS[60], composés aromatiques environmentaux[64], pesticides[65],BFR[65], métaux[65]
Lait maternel	Large fenêtre	Oui	POPS[66], pesticides[67], mycotoxine[68]
Salive	Large fenêtre	Non	Pesticides[69], métaux[69]
Méconium	Large fenêtre	Non	Médicaments[70], pesticides[70][71], drogues[72]
Sang (Cordon)	Large fenêtre	Non	Métaux[71] [56]

1.3.1.2. Préparation des échantillons

D'un point de vue global, la préparation des échantillons vise à extraire et à concentrer les composés d'intérêts afin de limiter lors des acquisitions de données, les interférences éventuelles avec d'autres constituants de la matrice. Il s'agit d'une étape cruciale qui doit être cohérente avec l'objectif ciblé ou non de l'étude. De plus, chaque type de matrice nécessite une préparation spécifique en fonction de sa nature propre et des composés recherchés.

Il n'existe donc pas de méthode universelle de préparation des échantillons. Néanmoins, on peut souligner l'existence d'étapes communes utilisées dans de nombreuses études, telles que les extractions en phase solide (SPE), l'extraction liquide-liquide et la méthode QuEChERS.[73]

D'autres étapes plus spécifiques peuvent être employées afin d'améliorer les rendements d'extraction pour certaines familles de composés spécifiques. C'est notamment le cas pour la recherche de pesticides pour caractériser l'exposome avec la méthode QuPPe[74][75] et une variante proposant une combinaison de la méthode QuEChERS et QuPPe (QuEChUp).[76]

Il est important de préciser que cette étape de préparation des échantillons va à la fois concentrer des composés d'intérêt, mais également écarter d'autres familles de composés, altérant ainsi la composition initiale de la matrice. Ces altérations peuvent affecter la qualité des données produites, compromettant sérieusement leur exploration, et donc les résultats obtenus ainsi que leurs interprétations.[77]

Idéalement, une préparation minimale des échantillons est souhaitable dans une approche non ciblée qui vise à maximiser la couverture de l'ensemble des molécules chimiques. Le protocole doit donc être simple, non sélectif et reproductible avec un nombre limité d'étapes afin de limiter la perte ou l'altération de certaines molécules et de garantir la qualité propre de l'échantillon et, par conséquent, celle du résultat analytique.

1.3.2. Outils analytiques pour produire des empreintes exposomiques exploitables

Aucune technologie ne permet à elle seule de couvrir l'intégralité des molécules chimiques en raison de sa grande diversité physico-chimique et de sa variabilité de concentration. C'est pour cette raison que la communauté scientifique préconise l'utilisation de plusieurs plateformes analytiques en métabolomique, telles que la GC-MS, la LC-MS ou la RMN, ceci afin d'accroître la couverture du métabolome.[78] Il est bien évident que l'objectif de couvrir un maximum de l'ensemble des molécules dans l'exploration de l'exposome chimique suggère l'utilisation des mêmes approches analytiques que dans les études métabolomiques.

Grâce à sa sensibilité, à sa sélectivité et à sa spécificité élevées, la spectrométrie de masse présente un réel avantage par rapport aux autres techniques analytiques.[79] Le choix des différentes techniques, c'est-à-dire le système de séparation (chromatographie liquide ou gazeuse), la source d'ionisation et l'analyseur de masse, dépend principalement des propriétés physico-chimiques de la classe de composés à analyser. L'utilisation de méthodes complémentaires, telles que des conditions chromatographiques différentes et des modes d'ionisation négative et positive, peut notamment augmenter la couverture de détection des molécules chimiques.

1.3.2.1. Méthodes de séparation en amont de la spectrométrie de masse

Le couplage GC-MS ou LC-MS combinant deux appareils analytiques, permet de séparer les composés chimiques en deux dimensions (2D). Cette séparation 2D permet de couvrir l'analyse d'un grand nombre de composés présents dans une matrice donnée. De ce fait, cette

technique est privilégiée dans de nombreuses études pour l'analyse de matrices complexes. Ce type de couplage permet non seulement de séparer différents composés, mais aussi de considérer le temps de rétention comme un paramètre supplémentaire pour caractériser chaque molécule.

Depuis longtemps, la GC-MS est la technique utilisée pour la détection de multiples composés chimiques.[80] Sa grande sensibilité et la reproductibilité des spectres de masse en mode d'ionisation électronique (IE) ont conduit au développement de bases de données indépendamment des instruments utilisés. Cependant, cette technique présente une grande limite car seuls les composés volatils et thermiquement stables peuvent être analysés. Par conséquent, une étape de dérivatisation est généralement nécessaire.

De nos jours, la LC-MS utilisant l'ionisation à pression atmosphérique (API) est l'approche la plus utilisée en analyse non ciblée. La chromatographie en phase inverse est souvent utilisée dans de nombreuses études, malgré sa capacité limitée en termes de séparation de composés très polaires, ce qui est le cas de la plupart des métabolites.[81][82] Pour améliorer la couverture des molécules polaires, d'autres techniques chromatographiques telles que la chromatographie par paires d'ions ou la chromatographie liquide d'interaction hydrophile (HILIC) ont été développées.[83] Il faut également souligner que l'introduction de la chromatographie liquide à ultra haute performance (UHPLC) a été très bénéfique pour renforcer l'analyse à haut débit.[84]

D'autres systèmes, comme l'électrophorèse capillaire (CE)[85] ou la spectrométrie de mobilité ionique (IM)[86], peuvent également être couplés à la spectrométrie de masse. L'EC-MS est bien adaptée à la séparation des composés polaires et chargés et nécessite un prétraitement minimal de l'échantillon. Cependant, elle offre une sensibilité et une reproductibilité plus faibles que d'autres techniques. La spectrométrie de mobilité ionique permet d'accéder à une autre dimension des ions en phase gazeuse en se basant sur leurs mobilités différentielles à travers un gaz tampon. Le couplage de la spectrométrie de mobilité ionique à la spectrométrie de masse (IM-MS) permet d'obtenir des données plus complètes, avec la possibilité de distinguer des isomères. Outre le temps de rétention et la valeur *m/z* fournis par la LC-MS, l'IM-MS apporte un nouveau critère d'identification des molécules, à savoir la section efficace de collision (CCS).[87]

Il est également possible de réaliser les analyses par spectrométrie de masse en mode d'injection directe (direct introduction mass spectrometry (DIMS) des échantillons sans séparation chromatographique préalable.[88] Bien que ce type d'approche permette un gain considérable dans le débit d'analyse, il est sensible aux effets de matrice et produit des empreintes analytiques moins riches que celles obtenues par des techniques de couplage.

La figure 1-2 issue de l'article de Manz *et al* .[45] résume les différentes techniques utilisées dans les études publiées pour analyser des contaminants chimiques dans diverses matrices. Dans les études examinées, la LC-MS est la technique la plus utilisée (50 %), la GC-MS est un peu moins utilisée (seulement 33 %), et l'analyse directe des échantillons n'apparaît que dans une seule étude. La combinaison de la GC-MS et de la LC-MS est employée dans 16 % des études.

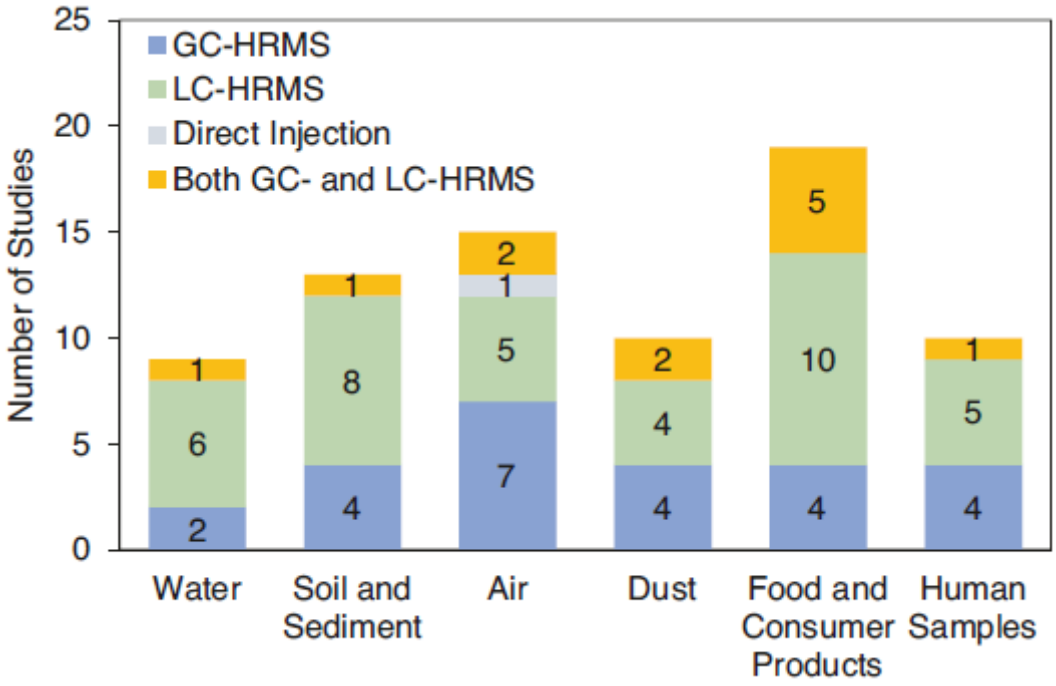


Figure 1-2 : Technique de chromatographie (GC ou LC) et/ou injection directe utilisée en couplage avec la spectrométrie de masse pour la détection de contaminants chimiques dans diverses matrices, telles que l'eau, le sol, les sédiments, l'air, la poussière, les aliments et les produits de consommation, ainsi que les échantillons humains.[45]

1.3.2.2. Conditions d'acquisition des données par spectrométrie de masse

La spectrométrie de masse permet de séparer les composés en fonction de leur rapport masse-sur-charge (*m/z*). Les espèces ionisées, chargées positivement ou négativement, sont d'abord formées par un processus d'ionisation.

Techniques d'ionisation

Divers modes d'ionisation existent et dépendent des classes chimiques des composés étudiés. Parmi les plus communes, l'ionisation en phase gazeuse, comme l'IE ou l'ionisation chimique (IC), est utilisée en GC-MS pour les composés volatils et ceux rendus volatils après dérivatisation chimique. Les sources d'ionisation à pression atmosphérique[89] comme l'APCI (ionisation chimique à pression atmosphérique), l'APPI (photoionisation à pression atmosphérique) et l'ESI (ionisation par électronébulisation) sont compatibles avec les analyses LC-MS de molécules non volatiles peu polaires à très polaires. L'ESI est particulièrement adaptée aux molécules polaires et c'est la technique d'ionisation la plus douce, ce qui permet d'accéder au poids moléculaire de la plupart des composés (Figure 1-3).

Pour toutes les techniques (excepté l'IE, qui ne fonctionne qu'en mode positif), il est possible de réaliser des analyses en mode positif ou négatif, conduisant à la formation d'espèces chargées positivement ou négativement par protonation ou déprotonation, respectivement. Il faut également noter que d'autres espèces, comme des adduits, peuvent être observées sous forme de cations (avec Na^+, K^+, NH_4^+, etc.) en mode positif ou sous forme d'anions (avec Cl^-, CH_3COO^-, etc.) en mode négatif.

Parmi les études examinées par Manz *et al* .[45], l'ESI est la technique privilégiée dans celles qui utilisent la LC-HRMS, avec 43 % en modes négatif et positif, 18 % uniquement en ESI^+, et 22 % en ESI^- (Figure 1-3a). Toutes les études réalisées par la GC-HRMS utilisaient l'IE, certaines d'entre elles (11% des études) complétaient avec la CI (Figure 1-3b).

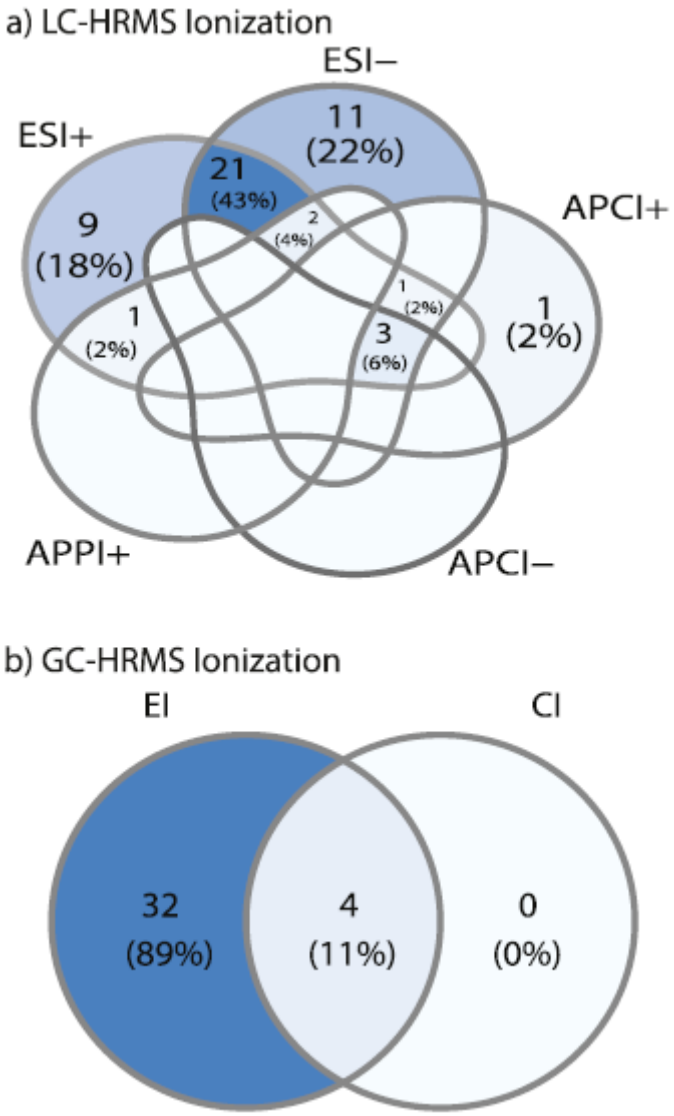


Figure 1-3 : Types d'ionisation utilisés a) en LC-HRMS et b) en GC-HRMS dans les analyses non ciblées pour les études examinées par Manz *et al* .[45] APCI+ et APCI- : ionisation chimique à pression atmosphérique en mode positif et négatif respectivement ; APPI+ : photoionisation à pression atmosphérique en mode positif et, ESI+ et ESI- : ionisation par électronébulisation en mode positif et négatif respectivement.

Performances des analyseurs de masse

Les performances des analyseurs de masse sont principalement caractérisées par leur pouvoir de résolution en masse, qui est lié à la mesure précise en masse et à la gamme dynamique (capacité d'un spectromètre de masse à détecter des signaux de très faibles intensités en présence de signaux très intenses).

Le pouvoir de résolution d'un spectromètre de masse, *Rp*, caractérise la largeur d'un pic et est défini par le rapport : $\mathrm{Rp} = \frac{\mathrm{m}}{\Delta \mathrm{m}}$ avec m = *m/z* de l'ion considéré et Δm la largeur à mi-hauteur du pic (FWHM, Full Width at Half Maximum). Ainsi plus le *Rp* est élevé, plus l'appareil est capable de séparer des espèces ioniques de rapports *m/z* très proches. C'est notamment le cas des ions isobares (mêmes masses nominales, mais masses exactes différentes). Les analyseurs de masse sont donc classés en fonction de leur *Rp*. Les instruments dont le *Rp* est supérieur à 10 000 (FWHM) sont qualifiés d'instruments à haut pouvoir de résolution en masse, tels que les instruments ToF (Time of Flight), de type Orbitrap ou à résonance cyclotronique ionique à transformée de Fourier (FTICR). Les spectromètres de masse à piège à ions et à quadripôle sont considérés comme des instruments à faible pouvoir de résolution.

Historiquement, le ToF[90] a été le premier analyseur introduit en 1946, suivi par le quadripôle[91] et le piège à ions[92] en 1953. Cependant, ces analyseurs présentaient un faible *Rp*. D'autres types d'analyseurs ont ensuite été développés, comme le FTICR introduit dans les années 1970[93] avec un *Rp* allant de 50 000 à 100 000 et l'orbitrap en 2005[94] d'après un principe observé dans les années 1920[95], avec un *Rp* pouvant atteindre 100 000. Depuis, ces instruments sont devenus plus en plus performants , notamment, l'instrument FT-ICR-MS avec un *Rp* de plus de 1 000 000[96], et les spectromètres de masse hybrides (Q-ToF, Orbitrap-ToF, etc.) pouvant atteindre des pouvoirs de résolution jusqu'à 500 000, voire plus.

Le développement de ces outils de plus en plus performants (haut pouvoir de résolution, voire très haut) a permis l'analyse de matrices complexes et la détection de composés chimiques présents à l'état de trace dans ces matrices.

Cependant, le pouvoir de résolution seul ne suffit pas pour faire le choix d'un analyseur de masse. Bien que le FT-ICR-MS possède le pouvoir de résolution le plus élevé, le temps d'acquisition de données, d'environ quelques secondes par scan, rend son couplage avec la LC peu pratique, car le nombre de points de données acquis dans le pic chromatographique peut

être insuffisant. Il est donc préférable d'utiliser des analyseurs à plus faible pouvoir de résolution, mais dont le couplage avec la LC est plus simple à réaliser. L'analyse de matrices complexes par LC-HRMS ne nécessite pas forcément un pouvoir de résolution très élevé. Cependant, la détermination structurale à partir des valeurs *m/z* des espèces ioniques d'intérêt s'avère plus difficile du fait d'un plus grand nombre de possibilités.

Qualité des données produites

La fiabilité des résultats obtenus grâce aux analyses LC-HRMS dépend de la qualité des données acquises. Des erreurs ou variations dans les mesures peuvent survenir à différentes étapes d'une analyse, affectant ainsi la qualité des données acquises, que ce soit lors de la collecte des échantillons, de leur stockage, ou de leur préparation, ou encore lors de l'acquisition des données ou de leur traitement. L'utilisation de procédures de contrôle qualité (QC) est largement recommandée pour détecter efficacement ces erreurs.[97,98]Cette détection précoce permet d'identifier les anomalies affectant les performances des instruments ou le traitement des échantillons, et éventuellement de les corriger, évitant ainsi que les analyses ultérieures soient affectées. Un échantillon QC peut être constitué de composés standards ou d'un mélange de tous les échantillons biologiques pour une étude de petite ou de moyenne taille ou, d'un sous-groupe représentatif de l'ensemble des échantillons pour une étude à grande échelle. L'échantillon QC est généralement analysé périodiquement tout au long de la séquence analytique afin de surveiller la stabilité des mesures et des dérives des signaux analytiques. Le QC est essentiel pour garantir la qualité des données produites. Il est possible de corriger ou de minimiser les écarts ou les incohérences dans l'analyse (dus à des dérives instrumentales, à des erreurs humaines ou à des fluctuations biologiques) en appliquant des algorithmes et en utilisant les signaux issus des analyses du QC.[97,98]

1.4. Analyse des jeux de données exposomiques

Les jeux de données analytiques générés par LC-HRMS nécessitent des outils bioinformatiques pour faciliter leur exploration et leur interprétation.

1.4.1. Prétraitement des données

Le prétraitement des données est indispensable pour obtenir des données exploitables à partir des données brutes contenant des valeurs erronées et des bruits parasites. Il permet en outre une analyse subséquente plus précise, une quantification fiable et une meilleure identification des composés chimiques. Sans ce prétraitement, les résultats pourraient être influencés par des signaux artefactuels, ce qui rendrait toute conclusion scientifique difficile.

1.4.1.1. Pourquoi prétraiter les données LC-MS ?

Les données brutes générées par LC-MS ou GC-MS sont composées de la totalité des scans enregistrés au cours de chaque analyse. Chaque scan comprend tous les ions détectés à des valeurs *m/z* associées à leurs intensités et leurs temps de rétention.

La figure 1-4.A représente en trois dimensions les données brutes d'un échantillon de lait maternel analysé par LC-HRMS en mode positif obtenu à partir du logiciel BatMass[99]. Chaque variable (« feature » en anglais) correspond à une espèce ionique détectée sous forme d'un pic constitué d'un ensemble de points caractérisés par une même valeur *m/z*, et un temps de rétention et une intensité maximale au sommet du pic (Figure 1-4.B).

En général, ces données brutes contiennent également des valeurs aberrantes ne correspondant à aucune espèce chimique (signaux artefactuels) : les signaux correspondants peuvent être très intenses et localisés (Figure 1-4.C) ou de faibles intensités et dispersés (Figure 1-4.D). D'autres pics de même *m/z* peuvent également être observés tout au long de l'élution chromatographique. Ils peuvent provenir de contaminants présents dans les phases mobiles, la source d'ionisation ou dans les consommables (Figure 1-4.E).

Avant de pouvoir exploiter ces données, il est nécessaire de les prétraiter afin de définir les variables correspondant à des espèces chimiques.

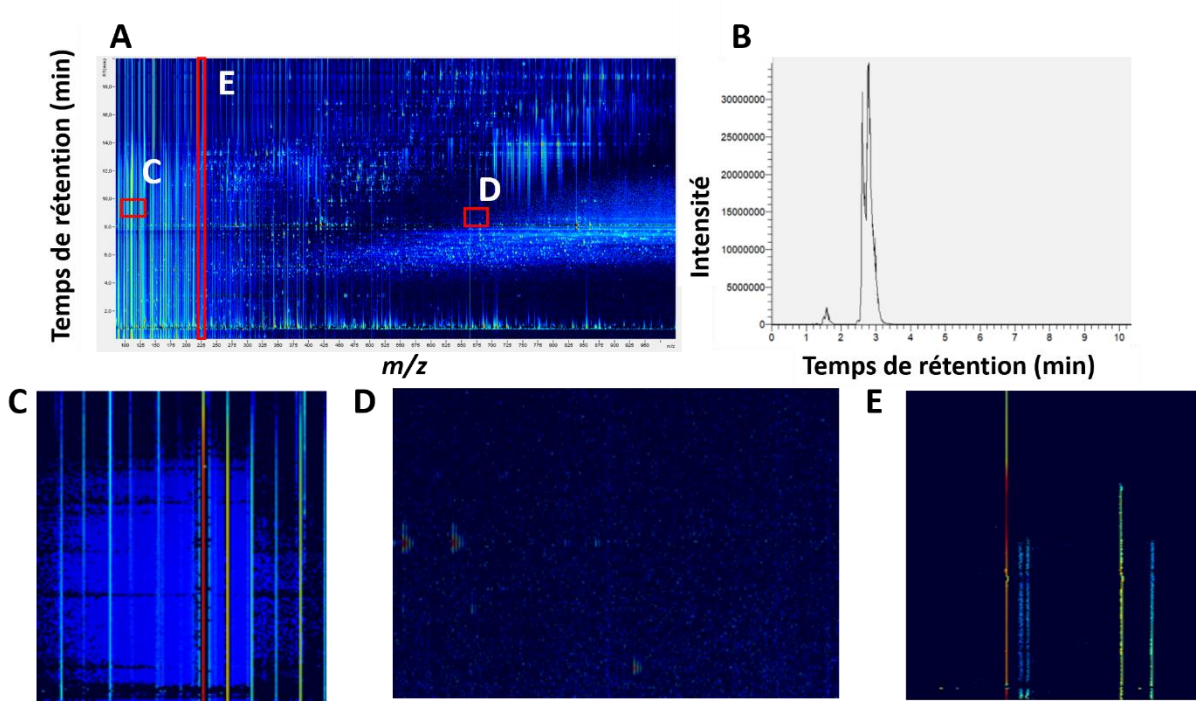


Figure 1-4 : (A) Représentation graphique des données brutes d'un échantillon de lait maternel analysé par LC-HRMS en ESI^+, avec en abscisse le rapport *m/z*, en ordonnée le temps de rétention et en graduation de couleurs (bleu à rouge) l'intensité des ions. (B) Signal d'une espèce ionique détectée sous forme d'un pic constitué d'un ensemble de points caractérisés par une même valeur *m/z* (ici *m/z* 152.07), avec en abscisse le temps de rétention et en ordonnée l'intensité. (C) Signaux artefactuels très intenses et localisés autour d'un pic très intense. (D) Signaux artefactuels de faibles intensités et dispersés (bruits). (E) Pics de même *m/z* observés tout au long de l'élution chromatographique.

1.4.1.2. Différentes étapes du prétraitement des données

De nombreux outils d'accès libre (open sources) ou proposés par des constructeurs existent pour effectuer le prétraitement des données LC-HRMS tels que XCMS, MZmine2, MetAlign, OpenMS, etc. Cependant, chacun de ces outils possède ses propres spécificités en fonction des domaines scientifiques visés (métabolomique, protéomique, etc.), incluant en conséquence des approches adaptées. Castillo *et al* . ont décrit une partie de ces outils ainsi que leurs spécificités dans une revue scientifique (Figure 1-5).[100]

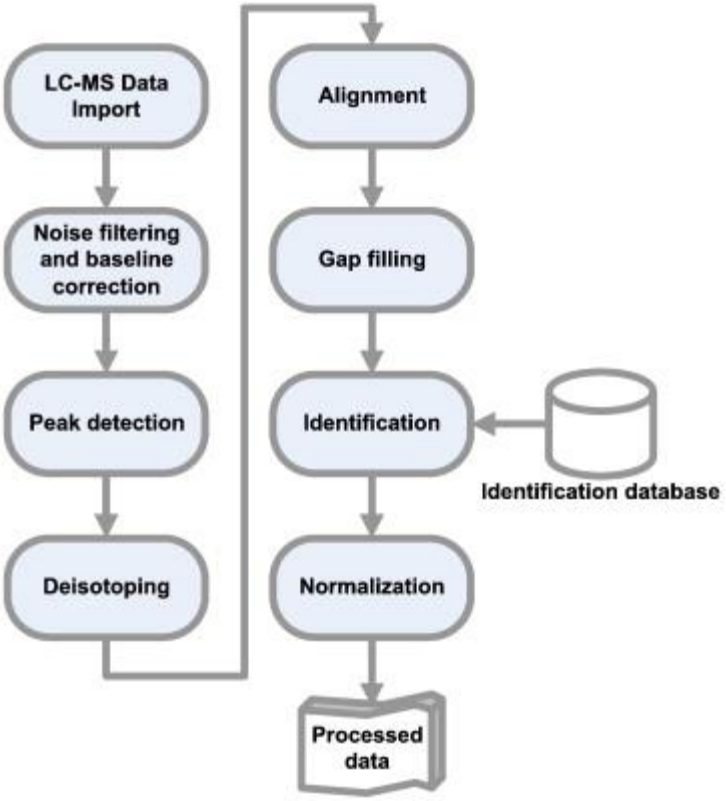


Figure 1-5 : Représentation graphique des différentes étapes de prétraitement des données LC-MS généralement utilisées.[100]

A notre connaissance il n'existe aucun outil permettant d'explorer l'exposome à partir des données brutes.

Le prétraitement des données comprend plusieurs étapes (Figure 1-5), qui sont décrites ci-dessous. Chaque étape dépend du type de données obtenues (LC-MS ou GC-MS, par exemple) et n'est pas influencée par le domaine scientifique traité.

(i) Nettoyage des données brutes (Noise filtering and baseline correction)

Le nettoyage des signaux artefactuels dans les données brutes permet de limiter les faux-positifs lors des analyses ultérieures. Les stratégies employées diffèrent en fonction des outils utilisés. XCMS[101] utilise un seuil d'intensité pour supprimer tous les signaux en dessous de celui-ci. OpenMS[102] propose plusieurs filtres, notamment un filtre gaussien, un filtre de Savitzky-Golay et une correction de la ligne de base.

Cette étape de nettoyage des données est critique. En effet, la suppression des points en dessous d'un seuil fixé peut affecter directement la forme de certains pics de faibles intensités et donc l'efficacité des étapes ultérieures pour assurer la détection des analytes. Il convient

de trouver un compromis lors de la définition de ce seuil afin de conserver un maximum de composés d'intérêt tout en éliminant le plus grand nombre d'artefacts.

(ii) Détection et intégration des pics (Peak detection)

Détecter et intégrer correctement l'ensemble des points formant chaque pic correspondant à une espèce ionique permet de garantir la qualité et l'intégrité des données.

Il n'existe pas de méthodes universelles pour détecter ces pics, chaque outil ayant sa propre manière de fonctionner avec ses avantages et inconvénients.[100] Le manque de normalisation entre algorithmes lors de cette étape est problématique pour la communauté scientifique. Une étude de Rafiei *et al* . a comparé quatre outils d'intégration de pics et a montré que les pics intégrés communs issus de ces quatre outils représentaient moins de 10% de la totalité des pics détectés.[103]

Pour des analyses qualitatives visant à identifier et caractériser des composés chimiques, une faible performance des outils de détection et d'intégration est moins impactante dès lors qu'elle permet d'identifier un pic. Notez cependant que l'application d'outils d'analyse de données ou d'étapes de prétraitement supplémentaires utilisant l'intensité des pics comme base sera fortement impactée par une mauvaise intégration.

S'il est possible de vérifier manuellement ces intégrations pour une liste restreinte de composés, cette vérification sur l'ensemble d'un grand jeu de données est impossible. Il est donc recommandé de tester et d'optimiser manuellement les paramètres en se basant sur une liste définie de pics avec une grande gamme dynamique, afin d'intégrer correctement l'intensité, le rapport *m/z* et le temps de rétention et, ainsi, de capturer au mieux les spécificités des jeux de données.

(iii) Eliminer les pics isotopiques (Deisotoping)

Cette étape vise à détecter et supprimer les pics redondants provenant des massifs isotopiques et/ou adduits de chaque espèce détectée. Si cette fonctionnalité est pertinente dans les études métabolomiques pour éviter la redondance des signaux correspondant à un même composé lors d'un traitement statistique, la détection des signaux corrélés peut également s'avérer intéressante non pas pour supprimer un massif isotopique, mais pour identifier ceux qui sont spécifiques à certains atomes comme le chlore et le brome. La détection de ces massifs isotopiques spécifiques est particulièrement intéressante dans le cas de l'exposome, car elle permet de souligner la présence de certains xénobiotiques, notamment ceux portant des atomes halogènes comme le chlore et le brome.

(iv) Alignement (Alignment)

L'alignement apporte des corrections dans les données pour les études qui nécessitent de comparer plusieurs jeux de données obtenus par plusieurs séries d'analyses. Il permet de corriger les dérives du temps de rétention pour les mêmes espèces détectées dans plusieurs échantillons et donc d'attribuer les mêmes caractéristiques (*m/z*, temps de rétention) à ces espèces.

Il est recommandé, lors des analyses, de s'assurer que les dérives du temps de rétention au cours du temps soient minimes. L'utilisation de QC (contrôle qualité, voir 1.3.3) permet, le cas échéant et si l'outil de prétraitement le permet, d'optimiser cette étape. XCMS propose d'utiliser les pics détectés dans les QC injectés périodiquement lors de la séquence analytique pour évaluer la déviation du temps de rétention et, ainsi, corriger l'ensemble des jeux de données sur cette base.

Une étude de Lange *et al*. a comparé plusieurs outils de prétraitement (OpenMS, XCMS et Mzmine) proposant une étape d'alignement.[104] En les appliquant à quatre jeux de données, deux en métabolomique, deux en protéomique, les auteurs ont montré que les performances étaient en adéquation avec le design de ces outils pour leurs domaines scientifiques respectifs (XCMS et Mzmine pour la métabolomique et OpenMS pour la protéomique). Cela implique que le choix de la stratégie d'alignement est potentiellement orienté par le domaine d'étude, ce qui rend difficile l'idée d'une méthode universelle pour corriger les jeux de données produits par LC-HRMS.

(v) Solution pour des données manquantes (Gap filling)

Dans le cas de l'analyse de séries d'échantillons, il est possible que l'étape de détection et d'intégration des données génère des données manquantes. En effet, certaines variables ne sont détectées que dans une partie des échantillons et non dans tous. Cela est probablement dû à une trop faible intensité ou à une mauvaise résolution de la forme des pics.

La non-détection d'une variable qui est présente dans les données brutes par les outils de prétraitement est problématique car elle conduit à la perte d'informations de certaines données qui peuvent être pertinentes et, donc, affecter les résultats obtenus.[105]

L'étape « gap filling » sert à retrouver l'information sur des données manquantes en corrigeant les erreurs des étapes antérieures du prétraitement en complétant les points manquants dans les données brutes sur la base des variables détectées dans les autres échantillons (temps de rétention et rapport *m/z*).

(vi) Identification

Cette étape vise à attribuer une identité aux espèces ioniques détectées dans les jeux de données LC-HRMS. À ce stade, l'identification s'effectue sur la base des données disponibles dans les bases de données en ligne ou en interne. Il existe différents niveaux d'identification. La description de ces niveaux se trouve dans la partie « 1.5. Identification des marqueurs d'exposition » à la fin de ce chapitre.

(vii) Correction des variations de la réponse analytique (Normalisation)

Une instabilité de l'intensité de certaines espèces ioniques peut être observée au cours des séquences analytiques « intra-batch » ou « inter-batch », qui est due à une dérive analytique. Les intensités des ions détectés n'étant plus représentatives de la concentration des composés présents dans les échantillons, il est essentiel de corriger ces effets. La dérive expérimentale peut être évaluée en se basant soit uniquement sur les QC, soit sur l'ensemble des données (y compris les QC). Plusieurs méthodes existent pour calculer les dérives et les corriger. On peut citer, par exemple, la moyenne et la régression linéaire appliquées aux QC, ainsi que les méthodes de régression polynomiale, de médiane glissante, de régression non-paramétrique (LOESS) et de courbes « spline » de lissage. Une étude de Rusilowicz *et al* . a testé ces différentes méthodes pour corriger sur une séquence analytique comprenant plusieurs séries d'analyses.[106] Les résultats obtenus après correction sont meilleurs que ceux obtenus sur des données non corrigées, quelle que soit la méthode utilisée. En particulier, la médiane

glissante a permis de mieux séparer deux groupes de plantes : un groupe soumis à un stress abiotique (sécheresse) et l'autre à un stress double (sécheresse et infection au Fusarium). Cependant, l'utilisation de cette même méthode a montré des résultats moins satisfaisants pour la discrimination entre les groupes témoins et « sécheresse ». Les auteurs ont également souligné que l'utilisation des QC pour la correction « inter-batch », bien que couramment pratiquée, ne permet pas toujours d'obtenir de meilleurs résultats. Ils ont par ailleurs affirmé qu'une méthode utilisant l'ensemble des variables peut s'avérer plus efficace dans certains cas. Néanmoins, une méthode de correction peut être mieux adaptée à certaines situations qu'à d'autres ; aucune méthode n'offre la correction optimale pour tous les cas.

1.4.2. Exploration des données pour caractériser l'exposome chimique

La fouille de données consiste à extraire des informations des signaux de la matrice des données obtenue après le prétraitement des données, en vue de rechercher la signature de composés chimiques spécifiques. Deux approches sont couramment utilisées : la recherche de molécules suspectées ou le « suspect screening », et la recherche sans *a priori* d'information ou recherche non ciblée.

1.4.2.1. Recherche de molécules suspectées : le « Suspect screening »

Le terme « suspect screening » a été introduit pour la première fois par Martin Krauss en 2010[107] pour désigner une approche qui consiste à rechercher les signaux parmi ceux désignant une liste de molécules suspectées. Cette liste est généralement établie sur la base d'informations préalables, en accord avec l'étude, mais elle peut également être extraite de bases de données afin de réaliser un screening plus large de composés dans un contexte exploratoire. Cette approche est couramment utilisée dans le cadre de l'exposome chimique[108,109], particulièrement lors de l'analyse de matrices complexes, qu'elles soient environnementales ou biologiques. Elle réduit artificiellement la complexité des données en se concentrant sur les signaux candidats pouvant correspondre aux composés de la liste de molécules suspectées excluant ainsi une grande partie des signaux sans intérêt.

Il faut également souligner que, selon le contexte de l'étude et les matrices analysées, cette liste de molécules suspectées peut comporter, en plus des molécules parentes, des métabolites (pouvant être formés lors du métabolisme des molécules parents) qui sont également des marqueurs d'une même exposition. Ces métabolites peuvent être ajoutés

manuellement en accord avec les informations recueillies dans la littérature mais il existe également des alternatives de prédictions *in silico* permettant d'enrichir cette liste de molécules suspectées de manière automatique.[110]

Pour comprendre comment la recherche de molécules suspectées est effectuée, il faut revenir à la nature des signaux. Pour un composé unique, plusieurs signaux sont détectés en spectrométrie de masse. Ils peuvent correspondre à différentes espèces ioniques formées (ions protonés/déprotonés et des ions adduits). De plus, pour chaque espèce, plusieurs signaux sont observés, formant un massif isotopique caractéristique des isotopes des éléments qui composent la molécule ionisée. Par exemple, le carbone possède deux isotopes naturels stables : le carbone 12 (^{12}C) et le carbone 13 (^{13}C), qui représentent respectivement 99% et 1% des isotopes du carbone ; le carbone 14 (^{14}C), isotope radioactif, n'est présent qu'à l'état de trace. La méthode couramment utilisée vise à détecter les signaux correspondant aux pics mono-isotopiques de chaque espèce, qui représentent la contribution des isotopes les plus abondants de tous les atomes constituant la molécule. Cette recherche peut parfois conduire à la détection de plusieurs signaux pour un même composé recherché. Ces signaux, également appelés faux-positifs, peuvent correspondre à des espèces isobares, des isomères ou des pics artefactuels. Il est alors difficile de confirmer ou d'infirmer la présence des composés recherchés sans aucune curation (vérification) manuelle de chaque signal. Dans le cas de l'analyse d'un grand nombre d'échantillons, cela reste irréalisable dans des délais acceptables.

Afin d'identifier le bon signal pour chaque composé recherché, plusieurs stratégies peuvent être mises en place dans le cadre d'une approche « suspect screening », notamment la prise en compte des massifs isotopiques afin de s'affranchir des artefacts et des isobares.

Les travaux de Vergeyn *et al* en 2015[111] ont démontré l'efficacité de cette stratégie dans la recherche par le « suspect screening » de 40 composés pharmaceutiques supposés, en évaluant le taux de faux-positifs et de faux-négatifs selon la prise en compte ou non du massif isotopique. Une diminution significative du nombre de faux-positifs et faux négatifs a pu être observée avec l'augmentation du nombre de pics isotopiques pris en compte (de 202 % à moins de 0,3 % et de 5,5 % à 4,2 % en considérant respectivement seulement le pic mono-isotopique et tous les isotopes abondants). Les valeurs de faux-positifs supérieures à 100 % peuvent s'expliquer par le fait qu'un composé donné est associé à plusieurs faux-positifs. De

plus, l'utilisation d'un filtre signal/bruit basé sur la forme des pics permettait également de réduire le nombre de faux-positifs et de faux-négatifs, de 202 % à 131 % et de 5,5 % à 5,3 % respectivement, en ne considérant que les pics mono-isotopiques. Cependant la combinaison des deux stratégies (recherche des isotopes abondants + filtre signal/bruit) conduisait à une diminution des faux positifs de 202 % à 0,3 % mais une augmentation des faux négatifs de 5,5 % à 6,7 %.

Bien que la détection de signaux cohérents avec la présence du massif isotopique attendu pour un composé donné permet de réduire le nombre de faux-positifs, notamment les artefacts en se basant sur la détection des carbones 13, qui est caractéristique des molécules organiques et les isobares dont les massifs isotopiques présentent des signatures différentes (Cl, Br, S, etc.), il faut souligner qu'elle ne permet pas de différencier les isobares ou isomères possédant la même signature chimique.

Afin de répondre à cette difficulté dans le cas où nous ne disposons pas de standards, des approches fondées sur la prédiction des temps de rétention sont de plus en plus utilisées. Elles visent à estimer les temps de rétention des composés de la liste de molécules suspectées sur la base de leurs structures.[112–114] Cependant, cette stratégie nécessite de remplir certaines conditions, en particulier l'annotation d'un certain nombre de composés au sein des jeux de données afin de prédire des temps de rétention en accord avec les conditions analytiques, plus précisément les colonnes chromatographiques utilisées.

1.4.2.2. Recherche sans *a priori* de marqueurs d'exposition

L'analyse non ciblée repose sur l'examen des jeux de données sans hypothèse préalable quant à la présence de composés recherchés. Cette approche vise à analyser l'intégralité du jeu de données, qui constitue une empreinte moléculaire d'une matrice analysée par LC-HRMS. Extraire des signaux pertinents dans des jeux de données complexes n'est pas facile. Diverses approches fondées sur les masses mesurées précisément ou sur les défauts de masse, c'est-à-dire la différence entre la masse précise d'un composé et sa masse nominale, ont été proposées pour explorer des données spectrales complexes.[115] Il s'agit notamment de méthodes d'analyse de données graphiques, telles que le diagramme de Kendrick[116] et le diagramme de van Krevelen[117], développées pour présenter les données complexes sous forme d'une empreinte moléculaire représentative de la matrice analysée, facilitant ainsi l'interprétation des données complexes et l'identification de certains composés ou de familles de composés spécifiques.

<u>Diagramme de Kendrick</u>

Kendrick[116] fut le premier, au début des années 60, à proposer une méthode graphique fondée sur les défauts de masses pour interpréter des données complexes produites en spectrométrie de masse à très haute résolution dans le domaine de la pétroléomique (analyse chimique détaillée du pétrole).

Dans le diagramme de Kendrick, l'échelle de masse habituelle de l'IUPAC (International Union of Pure and Applied Chemistry) est convertie en échelle Kendrick selon l'équation suivante :

$$\text{Kendrick mass} = \text{IUPAC mass} \times (M_n/M_e)$$

avec IUPAC mass : la valeur *m/z* mesurée, M_n et M_e : les masses nominales et exactes, respectivement, pour un motif de répétition donné

Pour une famille d'hydrocarbures où le motif de répétition est l'unité CH_2, la répétition en masse est de 14,01565 Da et l'échelle de Kendrick est obtenue par l'équation :

$$\text{Kendrick mass} = \text{IUPAC mass} \times (14/14{,}01565)$$

Le diagramme de Kendrick permet d'identifier des composés homologues, c'est-à-dire de même structure, mais avec un nombre différent de motifs de répétition, caractérisés par leur défaut de masse de Kendrick (KMD) identique. Il permet également de classer visuellement diverses classes de composés en fonction de leurs KMD distincts.

Kendrick mass defect = nominal Kendrick mass - exact Kendrick mass

La figure 1-6 présente le diagramme de Kendrick obtenu à partir d'un spectre de masse du pétrole brut lourd acquis en mode ESI négatif sur un instrument FTICR. Il montre une bonne résolution pour tous les points, au total 3900, avec une répartition sur l'axe vertical de différents types et classes de composés ayant des KMD différents, et sur les lignes horizontales des composés de même classe et de même type, mais avec un nombre différent d'unités CH_2. Ces derniers ont la même valeur de KMD mais sont séparés de 14 Da en masse nominale.

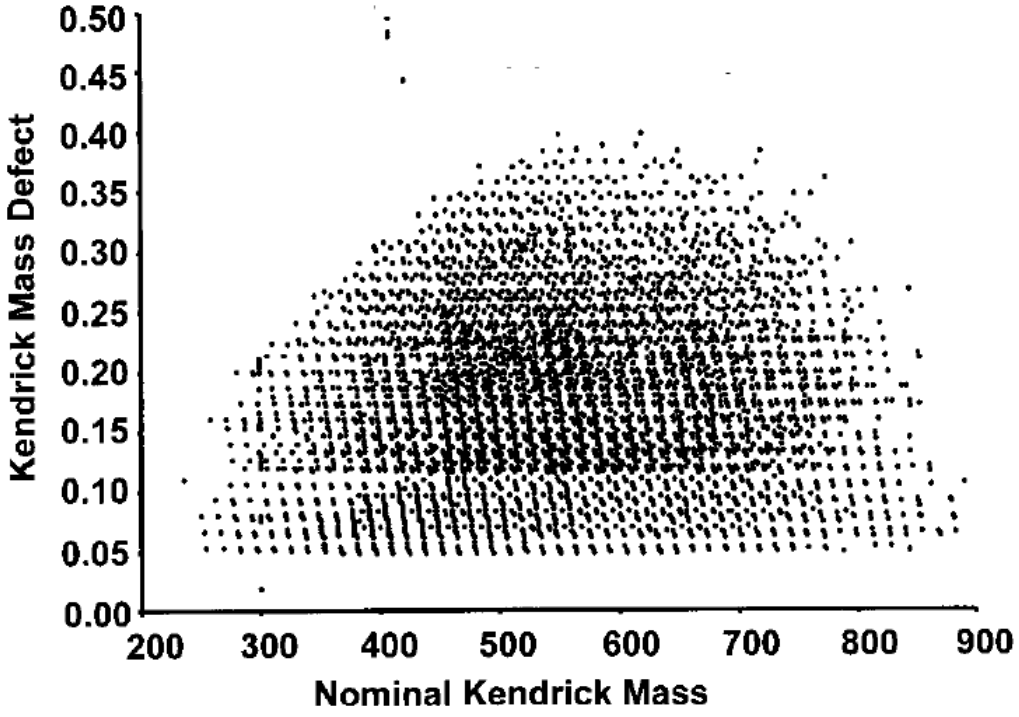


Figure 1-6 : Diagramme de Kendrick obtenu à partir de l'analyse du pétrole brut lourd par spectrométrie de masse en mode ESI négatif sur un instrument FTICR.[116]

Cette approche graphique est particulièrement intéressante pour caractériser l'exposome chimique. Elle peut révéler d'autres familles de composés en modifiant l'échelle de Kendrick d'origine, c'est-à-dire en prenant en compte des unités de base autres que le CH_2. Par exemple, l'échelle de substitution H/Br pour des composés polybromés ou l'échelle de substitution H/Cl dans le cas de composés polychlorés permet d'identifier rapidement ces composés dans des matrices biologiques. Comme le montre la figure 1-7, l'utilisation du diagramme de Kendrick dans une analyse non ciblée a révélé la présence de contaminants halogénés dans des échantillons de dauphin : les deux groupes de points (A et B), alignés et séparés de 34 Da, sont caractéristiques du massif isotopique du chlore. [118]

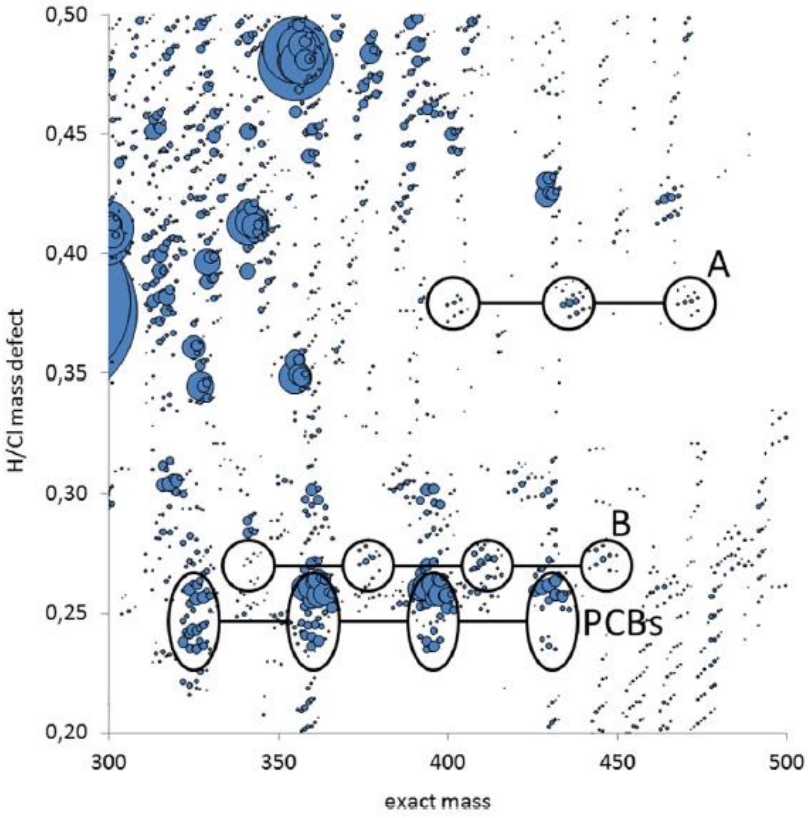


Figure 1-7 : Diagramme de Kendrick utilisant l'échelle de substitution H/Cl obtenu à partir de l'analyse non-ciblée GC×GC-HRQTOFMS d'un échantillon de muscle dorsal de dauphin. A et B sont des composés inconnus.[118]

Bien que cette méthode soit prometteuse, elle est généralement utilisée pour analyser seulement quelques échantillons, en raison de la complexité de l'interprétation manuelle. L'usage de la bioinformatique devient donc essentiel pour son application au traitement des données massives de cohortes. [119]

Diagramme de van Krevelen

Grâce à des instruments HRMS qui mesurent les masses des molécules avec une grande précision, il est possible d'attribuer une formule brute à chaque signal détecté. Exploiter les formules brutes peut aider à caractériser qualitativement les échantillons en mettant en évidence certaines familles de composés. Par exemple, les hydrocarbures aromatiques polycycliques (HAP), sont composés de nombreux atomes de carbone et d'hydrogène avec un fort taux d'insaturation, ou les lignines, qui possèdent plusieurs atomes d'oxygène. Une représentation graphique, comme celle de Van Krevelen, permet de visualiser et de séparer différentes familles de composés selon leurs rapports H/C et O/C. La figure 1-8 illustre l'utilisation du diagramme de van Krevelen dans une analyse non ciblée.[120] La comparaison entre le diagramme de van Krevelen (Figure 1-8.A) construit à partir des composés de la banque de données PhytoHub et celui issu d'une étude pilote a permis d'identifier différents types de polyphénols (Figure 1-8.B).

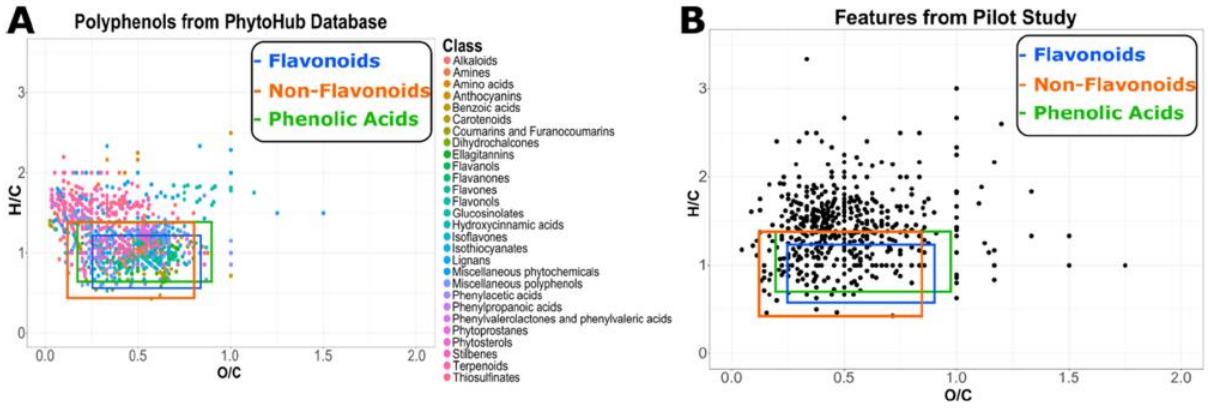


Figure 1-8 : (A) Diagramme de Van Krevelen généré à partir de plus de 2000 composés de la base de données PhytoHub. (B) Diagramme de Van Krevelen obtenu à partir des données acquises en mode DDA (Data-Dependent Acquisition) dans une étude pilote ; sont présentées dans ce graphique, les variables non annotées mais ayant une formule moléculaire générée par SIRIUS avec un score Zodiac supérieur à 0,8.[120]

Approche fondée sur des filtres de masse ou de défauts de masse

Une approche utilisant des filtres de défauts de masse (MDF, mass defect filter) a été développée dans le domaine du métabolisme des médicaments.[121] Elle permet de détecter sélectivement les signaux des métabolites présentant des défauts de masse relativement proches de celui du composé parent. Cette méthode pourrait également être très utile dans le domaine de l'exposome chimique pour rechercher des métabolites de xénobiotiques dans des jeux de données complexes en se basant sur des mesures très précises en masse et sur les modifications de ces masses par des réactions de biotransformation connues. Zhang *et al* [122] ont démontré l'efficacité de cette technique pour détecter des métabolites de médicaments dans des matrices biologiques. Des adduits au glutathion (GSH) de plusieurs médicaments ont pu être identifiés dans la bile grâce à leurs défauts de masse (Figure 1-9).

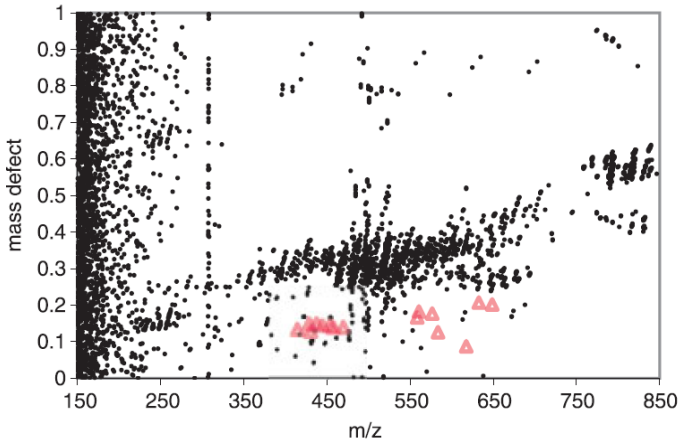


Figure 1-9 : Profil de défaut de masse de la bile humaine analysée par LC-HRMS.[122]

Les triangles rouges représentent les positions des adduits GSH de l'acétaminophène, de la carbamazépine, de la clozapine, du diclofénac, du p-crésol, du 4-éthylphénol et du 3-méthylindole. [122] L'examen de la distribution des défauts de masse peut également être appliqué à la détection d'autres métabolites conjugués (par exemple, les glucuronides), puisqu'il est évident que les positions de ces conjugués se déplaceraient vers la droite par rapport au composé parent dans la dimension *m/z*, sans que la dimension des défauts de masse n'augmente beaucoup. La limite de cette approche est que les filtres de défauts de masse ne couvrent pas toutes les réactions de biotransformation, en particulier de nombreuses réactions rares qui peuvent conduire à des changements importants de masse.

1.5. Identification des marqueurs d'exposition

L'annotation, en particulier la détermination structurale des composés d'intérêts, est primordiale afin d'identifier les marqueurs d'exposition et donc les substances chimiques auxquelles l'organisme a été exposé.

Il existe différents niveaux d'identification des composés chimiques. Cela dépend des échelles utilisées qui sont spécifiques à chaque domaine scientifique. On peut citer par exemple l'échelle MSI pour la métabolomique[123], Lipidsmap pour la lipidomique[124] et, de manière générale, le système de classification à cinq niveaux de Schymanski *et al* .[125] pour l'identification de composés par spectrométrie de masse. Dans le cas de l'exposome chimique étudié par spectrométrie de masse, il est donc préférable d'utiliser l'échelle proposée par Schymanski *et al* . (Figure 1-10), qui n'implique pas les spécificités d'autres domaines scientifiques.

Description des cinq niveaux d'identification selon Schymanski *et al* .[125] :

- Niveau 5 : Il s'agit du niveau d'identification le plus bas où l'identité d'un composé d'intérêt est attribuée en considérant la valeur *m/z* précise mesurée pour le pic mono-isotopique (prenant en compte les isotopes des éléments les plus abondants constituant la structure d'une molécule).
- Niveau 4 : L'attribution d'une formule brute (ensemble des éléments chimiques composant la structure d'une molécule) est sans équivoque en adéquation avec les observations sur les données spectrales et, en particulier, la présence du massif isotopique caractéristique (ensemble des signaux attendus considérant les isotopes des éléments constituants la formule brute et leurs abondances relatives), permettant de confirmer la présence d'éléments spécifiques et d'estimer leurs nombres en se basant sur leurs abondances relatives au sein du massif isotopique.
- Niveau 3 : La détermination de fonctions chimiques ou de sous-structures correspond à l'agencement des atomes dans les structures moléculaires, qui peuvent être caractéristiques de certaines familles de molécules. Ce niveau 3 d'identification nécessite des données issues d'expériences en spectrométrie de masse en tandem (MS/MS) afin d'observer des fragmentations caractéristiques de certaines

fonctions/sous-structures. Cela permet de donner des premières indications sur l'appartenance possible d'une espèce ionique d'intérêt à une famille de composés.

- Niveau 2 : La détermination de la structure probable d'un composé repose uniquement sur une comparaison entre les données expérimentales (MS et MS/MS) et des banques de données spectrales ou un ensemble de preuves incluant des fragmentations spécifiques.
- Niveau 1 : Le niveau d'identification le plus élevé n'est atteint que lorsqu'au moins deux données expérimentales obtenues avec des méthodes analytiques orthogonales sont similaires à celles acquises pour des composés standards analysés sur les mêmes conditions analytiques que les matrices étudiées. La disponibilité d'un composé standard est indispensable pour atteindre ce niveau 1.

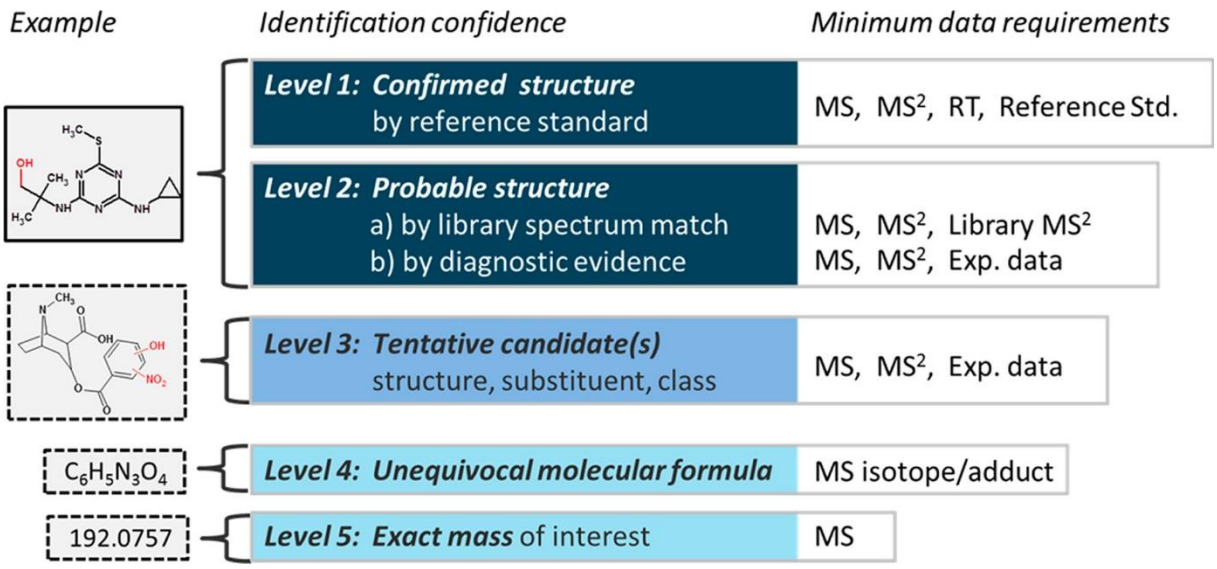


Figure 1-10 : Représentation des différents niveaux d'identification selon le système de classification à cinq niveaux de Schymanski *et al* .[125]

Les niveaux 4 et 5 d'identification sont généralement accessibles lorsqu'un instrument de haute performance est utilisé permettant des mesures précises en masse et la détection du massif isotopique caractéristique. Cependant les niveaux supérieurs nécessitent des expériences complémentaires de spectrométrie de masse en tandem pour générer des profils de fragmentation spécifiques.

Il convient de souligner que toutes les données de fragmentation ne permettent pas nécessairement de dépasser le niveau d'identification 2. En effet, l'analyse de matrices complexes contenant une quantité importante de composés et donc d'espèces ioniques, peut poser plusieurs problèmes. De manière non exhaustive on peut citer : i) des données issues de la fragmentation possible de plusieurs composés chimiques présents en mélange, rendant leur exploitation difficile ; ii) des conditions analytiques (énergie de collision) génériques non adaptées à la fragmentation de toutes les espèces ioniques (fragmentation pour certaines et défauts de fragmentation pour d'autres ne permettant pas d'apprécier un profil de fragmentation de bonne qualité) ; iii) des différences importantes au niveau des profils de fragmentation générés par plusieurs appareils, rendant difficile la comparaison avec les banques de données spectrales, réduisant ainsi les possibilités d'accéder au niveau 2 d'identification ; iv) la difficulté à obtenir des composés standards pour l'identification de niveau 1, et de mettre en place une banque de données interne, compte tenu du prix élevé des standards sur le marché.

Il est important de noter que cette échelle d'identification est reprise dans la littérature avec différentes variantes et propositions. Notamment, le temps de rétention a été proposé comme métrique supplémentaire pour améliorer les faibles niveaux d'identification (4 et 5).[126] En effet, les composés ayant un temps de rétention propre pour chaque méthode chromatographique, il est possible de limiter le nombre de candidats à ceux ayant des temps de rétention plausibles par comparaison avec des valeurs logP[113] (métrique quantifiant le caractère hydrophile des molécules en fonction de leur structure) et les temps de rétention des composés déjà identifiés.

Bien que le niveau 1 d'identification soit le plus élevé, il ne prend pas en compte les éventuels biais apportés par la méthode chromatographique. En effet, la séparation chromatographique joue un rôle important dans l'identification des composés chimiques, et notamment pour la séparation des isomères pour lesquels les spectres MS/MS ne sont pas suffisants pour les différencier. Dans ce cas, la non-séparation chromatographique du composé d'intérêt de ses isomères (même temps de rétention) peut conduire à des faux-positifs. Il serait pertinent de s'assurer de la bonne séparation chromatographique pour chaque famille de composés afin d'améliorer le niveau d'identification. Un niveau 0 supplémentaire a été proposé comme le

niveau le plus élevé d'identification incluant la détermination de la stéréochimie des métabolites.[127]

L'attribution d'une identité sans équivoque représente un défi majeur dans la caractérisation des marqueurs d'exposition. En raison de la grande diversité structurale des composés chimiques, il est souvent difficile de déterminer avec précision la structure du composé d'intérêt lorsqu'il s'agit de composés isobares (même masse nominale, mais des formules chimiques différentes) ou d'isomères. La procédure d'identification structurale des composés chimiques implique généralement l'utilisation d'instruments de haute performance pour obtenir des mesures très précises en masse. Des outils bioinformatiques sont souvent nécessaires pour extraire et annoter de façon automatique un grand nombre de signaux présents dans des jeux de données complexes. Cependant, l'expertise scientifique reste nécessaire pour une bonne interprétation des résultats en termes d'analyse structurale, afin de fournir une vraie identité aux substances d'intérêt ainsi caractérisées.

1.6. Conclusion

L'exposome est un concept global qui intègre l'ensemble des expositions environnementales, comportementales et internes propres à chaque individu. Il joue un rôle clé pour établir des liens avec le développement de certaines maladies, en particulier lorsque les expositions ont lieu durant la période de vulnérabilité de l'organisme, comme durant la phase périnatale. Étudier les différentes facettes de l'exposome exige en revanche d'avoir recours à des approches et techniques très diverses. Les approches présentées dans ce chapitre, sont dédiées à la caractérisation de l'exposome chimique, et en particulier celui périnatal. Elles impliquent l'analyse de matrices biologiques afin de détecter des marqueurs de contamination, qui peuvent ne plus être présents sous leur forme initiale mais métabolisée. Des études de cohorte mère-enfant ont été développées pour étudier l'impact de ces expositions précoces sur le développement et la santé future de l'enfant. Les matrices périnatales (placenta, méconium, etc.) collectées au sein de ces cohortes peuvent donc être analysées pour révéler une exposition très précoce, *via* l'exposition maternelle. La LC-HRMS s'est imposée comme la technique de référence pour caractériser l'exposome chimique dans les matrices biologiques. Cependant, l'ampleur des données générées, notamment compte tenu du grand nombre d'échantillons analysés, nécessite des outils bio-informatiques adaptés pour leur traitement.

C'est dans ce contexte que s'inscrit mon travail de thèse, qui vise à mettre en évidence l'exposition aux contaminants chimiques durant la période périnatale. Des matrices périnatales, telles que le méconium (n = 308) et le lait maternel (n = 320), prélevées au sein de la cohorte EDEN, ont été analysées par LC-HRMS dans le laboratoire dans le cadre des études métabolomiques, lors de deux précédentes thèses, celle de Mikaïl Berdi (soutenance le 22 décembre 2017, co-directeurs de thèse : Karine Adel-Patient et Christophe Junot) qui s'est intéressée à la composition globale du lait maternel et, celle de Nihel Bekhti (soutenance le 05 mai 2021, co-directeur de thèse : Karine Adel-Patient et François Fenaille) à la composition globale du méconium. Mon travail a donc consisté à développer des stratégies pour traiter automatiquement ce grand volume de jeux de données métabolomiques afin de rechercher des marqueurs d'exposition au sein des matrices périnatales, soit par l'approche dite de « suspect screening », soit de façon sans *a priori*. Les travaux réalisés sont présentés dans les chapitres 2 et 3, consacrés respectivement à l'approche par « suspect screening » et à l'approche sans *a priori*.

Chapitre 2. Développement d'une stratégie de recherche de marqueurs de molécules dont la présence est suspectée dans de grands jeux de données acquis par LC-HRMS

Je présente dans ce chapitre le travail que j'ai effectué sur une approche dite de « suspect screening ». Il s'agit de l'une des stratégies envisagées pour, en particulier, caractériser l'exposome chimique périnatal à partir des jeux de données métabolomiques produites par LC-HRMS sur des matrices prénatales, le méconium et le lait maternel provenant de cohortes mères-enfants françaises.

Une **introduction** sur les différentes étapes de l'approche de « suspect screening » mise en œuvre est d'abord présentée.

La **deuxième partie** est consacrée à la **méthodologie**, qui aborde la nature et le choix des matrices étudiées, les protocoles analytiques appliqués, les paramètres de prétraitement des données LC-HRMS utilisés, la sélection d'une première liste de molécules chimiques considérées comme préoccupantes pour la santé humaine ainsi que la stratégie pour explorer les données, incluant la prédiction des métabolites *in silico* de manière computationnelle ainsi que la génération et la recherche des signaux correspondant aux marqueurs potentiels d'exposition.

La **troisième partie** décrit les **résultats préliminaires** obtenus avec les premières étapes de la fouille de données automatique et **l'évaluation/validation de l'approche développée** pour, d'une part, prédire des métabolites *in silico* et, d'autre part, rechercher des signaux correspondant à la présence de xénobiotiques connus dans des matrices biologiques qui ont été expérimentalement enrichies à différentes concentrations.

La **dernière partie** présente les **résultats de l'application de notre approche** à travers deux stratégies présentées dans cette partie qui sont adaptées à la taille de la liste de composés suspectés. Elles ont été appliquées aux données d'échantillons de méconium et de lait maternel dans le cadre de l'étude présentée ici.

2.1. Introduction sur l'approche par « suspect screening »

Le travail est centré sur l'analyse des données métabolomiques afin de rechercher des marqueurs d'exposition aux contaminants chimiques durant la période périnatale, période jugée critique dans la vie d'un individu. Une des stratégies adoptées a été la recherche automatique ciblée dite de « suspect screening », sur des marqueurs de référence répertoriés dans une liste prédéfinie, ceci dans un grand nombre d'échantillons biologiques, tels que le méconium et le lait maternel provenant de différentes cohortes françaises.

Cette approche de « suspect screening » est composée de plusieurs étapes présentées dans la figure 2-1 et décrites ci-dessous :

i) Création d'une liste de marqueurs d'exposition potentiels. Cette étape comprend un enrichissement de la liste de composés ciblés, fondé sur une prédiction *in silico* de leurs métabolites possibles. Cette prédiction repose sur la structure des composés et une liste de biotransformations couramment observées dans le métabolisme des xénobiotiques. Cela permet de modéliser rapidement, *in silico,* les biotransformations potentielles des composés parents dans les organismes. Ainsi, cette méthode permet d'élargir la liste des composés à rechercher pour identifier des signaux comme marqueurs d'exposition (valeurs *m/z* de toutes les espèces pouvant être formées lors du processus d'ionisation et détectées en spectrométrie de masse).

ii) Création d'une matrice de données. Les données métabolomiques produites par LC-HRMS sur les matrices périnatales (le méconium et le lait maternel) sont prétraitées à l'aide du package XCMS[101] sous R.

iii) Recherche de marqueurs d'exposition dans la matrice de données. La recherche dans la matrice de données de signaux correspondants à la liste des composés suspectés enrichie de leurs métabolites se découpe en plusieurs étapes. Premièrement, une recherche des signaux correspondant aux pics mono-isotopiques est effectuée. Cela permet de restreindre le jeu de données aux candidats possibles. Puis, la recherche de signaux concordants correspondants aux massifs isotopiques de ces candidats est effectuée dans les spectres de masse extraits aux temps de rétention où les signaux des candidats sont détectés à la plus haute intensité ce qui permet, ainsi, de sélectionner les candidats les plus fiables.

Une vérification croisée entre scans et échantillons est ensuite réalisée sous forme de tests logiques et de corrélations. Ces tests s'assurent que les signaux des massifs isotopiques détectés suivent une tendance cohérente. Par exemple, cela permet de confirmer la présence d'un composé dans un échantillon où l'intensité du pic mono-isotopique est insuffisante pour détecter le signal du massif isotopique correspondant. Ce contrôle s'appuie alors sur un échantillon de référence où l'intensité est plus élevée, rendant le massif isotopique observable. De plus, la vérification entre scans garantit que les intensités relatives des massifs isotopiques sont maintenues sur plusieurs scans, évitant ainsi de prendre en compte des signaux artefactuels qui pourraient faussement suggérer un massif isotopique valide.

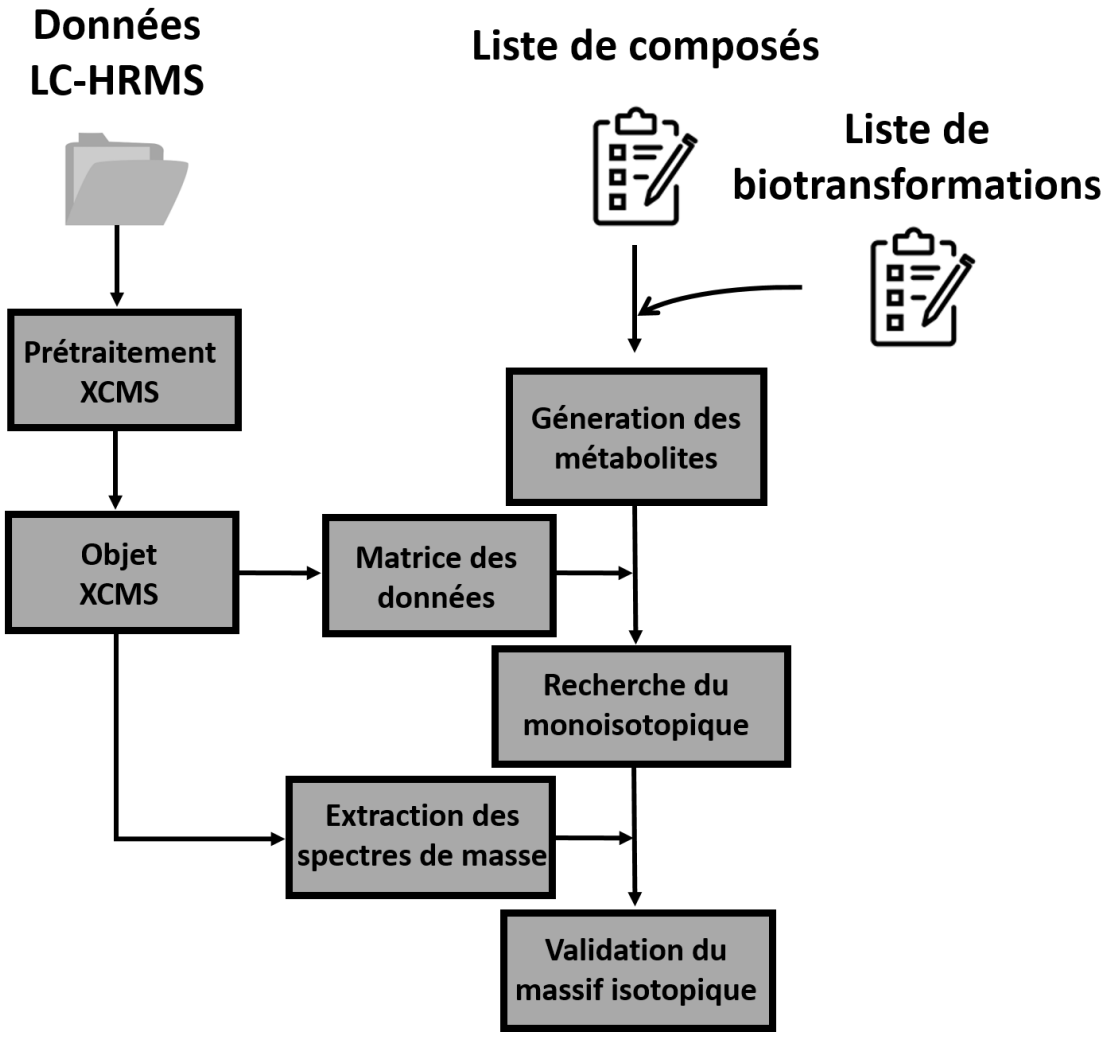


Figure 2-1 : Différentes étapes utilisées pour rechercher des marqueurs d'exposition à des xénobiotiques dans les données LC-HRMS par l'approche « suspect screening ».

2.2. Méthodologie

2.2.1. Sélection des matrices

Cohorte EDEN - La cohorte de naissance française, l'étude EDEN, vise à étudier les déterminants pré- et postnatals précoces du développement et de la santé de l'enfant.[17] Elle est constituée de 2 002 couples mère-enfant, recrutés dans deux hôpitaux universitaires français, à Nancy et Poitiers. Dans l'étude EDEN, plusieurs échantillons ont été collectés en maternité pour un sous-échantillon de la cohorte. Cela comprend des prélèvements biologiques des mères (urine, sang et lait maternel) ainsi que des échantillons issus du nouveau-né (méconium, cheveux, salive et sang de cordon). Dans le contexte de l'exposome chimique périnatal, nous nous intéressons plus particulièrement aux matrices telles que le méconium et le lait maternel dont les données métabolomiques ont déjà été acquises lors des précédents travaux de thèse de Mikaïl Berdi (soutenance le 22 décembre 2017, co-directeurs de thèse : Karine Adel-Patient & Christophe Junot) et de Nihel Bekhti (soutenance le 05 mai 2021, co-directeur de thèse : Karine Adel-Patient & François Fenaille).

Ces échantillons ont été sélectionnés en raison de leurs propriétés d'accumulation. En effet, le méconium, produit entre la 11^{e} et la 14^{e} semaine de grossesse et excrété dans les jours qui suivent la naissance, peut accumuler les xénobiotiques (ou leur métabolites) auxquels la mère a été exposée durant cette période et qui ont traversé la barrière placentaire, offrant ainsi une fenêtre d'exposition du fœtus étendue. Par ailleurs, la production du lait maternel va mobiliser les graisses maternelles, pouvant ainsi conduire à la libération de xénobiotiques liposolubles qui peuvent y être stockés et reflétant une période d'exposition de la mère plus ou moins longue (selon, par exemple, l'indice de masse corporelle (BMI) et la parité), et conduisant à une exposition néonatale du nouveau-né allaité. Il faut toutefois souligner que selon la période de prélèvement, les influences des expositions post-natales peuvent entrer en jeu, les prélèvements les plus précoces étant moins impactés.

Echantillons de méconium et de lait maternel - Au sein de la cohorte EDEN, un sous-échantillonnage de 300 couples mères-enfants (n=150 par centre) a initialement été réalisé, en sélectionnant aléatoirement les couples mère-enfant pour lesquels le méconium et le lait maternel avaient été collectés. En raison de la disponibilité des échantillons au sein de cet échantillonnage, 257 échantillons de méconiums et 274 échantillons de lait maternel ont été finalement obtenus. Des échantillons supplémentaires ont été ajoutés à cet échantillonnage

aléatoire. Ils correspondent à 51 échantillons de méconiums et 46 échantillons de lait maternel issus de couples mères-enfants supplémentaires pour lesquels les enfants ont déclaré une allergie alimentaire dans les cinq premières années de vie (design « case-cohorte »).[128] Ainsi, un total de 308 échantillons de méconium et 320 échantillons de lait maternel ont été reçus au laboratoire.

Tous les échantillons ont été conservés à -80 °C depuis leur collecte jusqu'à leur utilisation.

2.2.2. Protocole analytique

Préparation d'échantillon

Les méthodes utilisées pour traiter les échantillons de méconium et de lait maternel ont été mises au point et décrites dans des travaux antérieurs.[129,130] Les expériences de LC-HRMS ont été effectuées en suivant des protocoles optimisés, couramment utilisés au sein de la plateforme de notre laboratoire.[83] Les étalons internes ont été ajoutés lors de la préparation des échantillons (Annexe 2.1) et les étalons externes avant les analyses LC-HRMS (Annexe 2.2).

En résumé, les échantillons de méconium ont été lyophilisés à l'aide d'un lyophilisateur Triad™ Labconco (Missouri, États-Unis), avec une température de plateau fixée à 4 °C et une température de piégeage à -83 °C, sous un vide de 0,180 mbar. Dix milligrammes de méconium lyophilisé ont été dissous dans 750 µL de méthanol/eau (4:1, v/v) et homogénéisés à 4 °C à 6 500 tr/min, pendant 3 x 30 secondes dans un tube contenant des billes en céramique CK14 (Ozyme, Saint-Cyr-l'École, France) à l'aide d'un appareil Precellys 24® (Bertin Technologies, Montigny-le-Bretonneux, France). Après une incubation sur glace pendant 1 h 30 permettant une déprotéinisation complète, les échantillons ont été centrifugés à 20 000 g et à 4 °C pendant 15 minutes, puis les surnageants ont été répartis en plusieurs aliquotes de 200 µL. Enfin, les extraits ont été séchés sous un flux d'azote à 30 °C à l'aide d'un Turbovap® (Biotage, Uppsala, Suède). Les extraits séchés ont été stockés à -80 °C jusqu'à l'analyse par LC-MS.

Pour la préparation des échantillons de lait maternel, 50 µL de chaque échantillon sont prélevés, auxquels sont ajoutés 10 µL d'étalons internes et 200 µL de méthanol (MeOH). L'échantillon est ensuite incubé pendant une heure à -20 °C, puis centrifugé à 10 000 rpm pendant 15 minutes à 4 °C. 220 µL du surnageant sont prélevés et répartis en deux portions : 130 µL pour l'analyse par LC-MS sur une colonne C18 et 90 µL pour l'analyse sur une colonne HILIC. L'échantillon est ensuite évaporé à sec sous un flux d'azote et stocké à -80 °C jusqu'à l'analyse par LC-MS.

Profilage métabolique par LC-HRMS

Les expériences de LC-MS ont été réalisées sur un système chromatographique Ultimate 3000 couplé à un spectromètre de masse Q-Exactive (tous deux de Thermo Fisher Scientific, Courtaboeuf, France) équipé d'une source d'électrospray (ESI). La séparation en chromatographique a été effectuée en utilisant deux types de colonnes pour maximiser la couverture du métabolome : une colonne C18 (Hypersil GOLD C18, 1,9 µm, 2,1 × 150 mm, Thermo Fisher Scientific) associée à l'ESI en mode positif (ESI+), et une colonne ZIC-pHILIC (chromatographie liquide à interaction hydrophile ; colonne Sequant ZIC-pHILIC, 5 µm, 2,1 × 150 mm, Merck, Darmstadt, Allemagne) associée à l'ESI en mode négatif (ES*I*-).

2.2.3. Prétraitement des données

Toutes les données brutes issues des analyses LC-MS ont d'abord été converties au format mzXML à l'aide d'un outil de conversion gratuit, **msConvert** de ProteoWizard.[131] Les données brutes des fichiers mzXML ont ensuite été prétraitées à l'aide du package **XCMS**,[19] un logiciel open source dédié au prétraitement des données acquises en spectrométrie de masse. Cette bibliothèque comprend un ensemble de fonctions d'extraction de caractéristiques, notamment la détection des pics et le regroupement des données. Les paramètres suivants ont été utilisés pour la détection des pics : **peakwidth = 9-20**, **noise = 500**, **prefilter = 6**, **ppm = 3**, **mzdiff = 0.0001**, **snthresh = 10**. Ces paramètres ont été optimisés à partir d'une liste de pics correspondant à des signaux de faible intensité dans les lots étudiés, et plusieurs paramètres ont été évalués *via* l'intégration des pics issus de cette liste. Les paramètres de regroupement des données étaient les suivants : **minFrac= 0**, **bw = 5**, **binSize = 0.005**. Il est important de noter que la valeur de **minFrac** a été intentionnellement fixée à 0 afin de ne pas exclure les pics présents dans seulement quelques échantillons, car ces pics

peuvent correspondre à des marqueurs de composés exogènes présents dans seulement quelques échantillons.

2.2.4. Prédiction *in silico* des métabolites

Pour prédire la structure des métabolites, les langages spécifiques tels que le langage SMILES (Simplified Molecular Input Line Entry System), le langage SMARTS (SMiles ARbitrary Target Specification) et le langage SMIRKS (SMiles and Reaction Kinetics SMart) sont utilisés en bio-informatique pour décrire la structure des molécules chimiques, identifier des fonctions ou motifs présents dans leur structure et les réactions chimiques, respectivement. Il est ainsi possible de conditionner la prédiction des métabolites à la présence ou non de ces fonctions ou motifs en fonction des réactions de biotransformation décrites dans la littérature.

2.2.4.1. Définition des SMILES, SMARTS, SMIRKS

SMILES (Simplified Molecular Input Line Entry System)

Le SMILES est un système de notation chimique introduit par David Weininger en 1988.[132] Ce système a été conçu pour offrir une représentation simple et standard des molécules chimiques par une chaîne de caractères qui se veut intuitive au regard des structures chimiques et compréhensible par des systèmes informatiques. Depuis sa création, plusieurs modifications ont été apportées, notamment par des chercheurs de la société Daylight Chemical Information Systems et par la communauté scientifique, pour améliorer la précision du système et étendre sa capacité à couvrir un éventail plus large de structures chimiques pour répondre aux besoins de la communauté scientifique.

Dans la notation basique de SMILES, tous les atomes sont représentés par leur symbole chimique (C pour le carbone, S pour le soufre, O pour l'oxygène, etc.) à l'exception de l'hydrogène dont le nombre est implicite (sauf exception) en fonction de la valence des atomes (nombre de liaisons qu'il peut former avec d'autres atomes) et du nombre de liaisons déjà formées. La notation est linéaire et se fait en fonction de la structure. À partir de la sélection d'un atome, généralement situé en bout de chaîne, il suffit de regarder son voisin le plus proche et d'ajouter cet atome au SMILES. Par exemple, un alcool composé d'un carbone et d'un oxygène portant un hydrogène s'écrira CO.

Dans le cas d'une chaîne carbonée ramifiée, c'est-à-dire non linéaire, il est préférable de commencer par la notation de la chaîne carbonée la plus longue avant de continuer sur l'autre

embranchement. Cela se traduit par l'ajout d'une parenthèse pour chaque substituant. Par exemple, le SMILES du 2-méthyl-2-butanol est le suivant : CCC(O)(C)C. À noter qu'il existe plusieurs notations possibles pour une même structure moléculaire en fonction de l'atome choisi pour commencer l'annotation et de l'atome choisi pour chaque embranchement. Par exemple CCC(O)(C)C et C(O)(C)CCC correspondent tous deux au 2-méthyl-2-butanol.

Les liaisons sont par défaut simples mais il est possible de définir des liaisons doubles (=) ou triples (#) dans le cas des cycles. Dans ce cas, un numéro (1 à n, n étant le nombre de cycles dans la molécule) sera ajouté sur les atomes qui se trouvent au début et à la fin du cycle pour définir l'ensemble des atomes qui en font partie. Par exemple, dans le cas d'un cycle simple à 6 atomes de carbone (cyclohexane), le SMILES correspondant sera : C1CCCCC1. Dans le cas des cycles aromatiques, le procédé est le même, à l'exception du fait que les atomes compris en son sein sont en minuscule. Par exemple, un benzène sera représenté par c1ccccc1. Dans les cas les plus complexes, ces cycles peuvent s'enchevêtrer et la notation sera la même bien que plus difficile à aborder. Par exemple, le SMILES du bicyclo[2.2.0]hexane sera C1C2C3C2C1C3 (un cycle principal à 6 atomes de carbone noté 1 et deux cycles à 4 notés 2 et 3).

Les exemples ci-dessus constituent une introduction pour comprendre le système de notation SMILES mais il existe de nombreuses fonctionnalités possibles. On peut notamment citer la prise en compte de la stéréochimie avec l'ajout de « / » ou « \ » pour les configurations cis/trans, ou l'ajout d'un « @ » ou « @@ » pour les configurations respectives R ou S. Des explications plus poussées sont proposées sur le site ayant permis de déployer les SMILES :

https://www.daylight.com/cheminformatics/index.html

Les utilisations des SMILES sont nombreuses puisqu'ils permettent à des outils informatiques d'accéder à des structures de molécules chimiques aussi complexes soient-elles. On peut citer notamment le référencement et la recherche dans les banques de données, mais aussi l'utilisation des SMILES en combinaison avec une technologie d'intelligence artificielle.[133] C'est notamment le cas avec le concept de « SMILES-to-Properties-Transformer », fondé sur un type de neurone profond qui permet de prédire des propriétés biologiques ou physico-chimiques à partir des SMILES, comme les logP et logS, la toxicité, l'énergie libre, l'activité biologique, le point de fusion, la densité, etc.[134]

SMARTS (SMiles ARbitrary Target Specification)

Les SMARTS[135] sont utilisés en tant que langage informatique pour décrire de manière flexible des fonctions chimiques ou des sous-structures, et pour les rechercher au sein d'une structure donnée. Ce sont des extensions de la notation SMILES qui incorporent, des opérateurs logiques (« ou », « et », « non ») sur les atomes et les sous-structures, ainsi que des éléments de notations plus larges ou plus spécifiques selon les besoins. Par exemple, la notation X<n> permet de sélectionner, pour l'atome qui le précède, le nombre de connexions total représenté par la valeur n ; la notation R permet de spécifier que l'atome recherché au sein des structures est dans un cycle, par exemple la notation [OR] vérifie la présence d'un oxygène dans un cycle.

L'opérateur « , » peut être utilisé pour la recherche d'un atome ou d'un autre, par exemple pour rechercher un atome de brome ou de chlore, le motif moléculaire recherché sera écrit [Cl,Br] sous forme de SMARTS.

L'opérateur « **et** » peut être défini de deux manières (**;** et **&)**, il permet par exemple de rechercher un atome de carbone associé à un atome de soufre [C&S].

L'opérateur **« ! »** permet d'exclure une condition de la recherche. Par exemple, la notation [!C] permet de rechercher tous les atomes autres que le carbone et la notation [!R] tous les atomes qui ne font pas partie d'un cycle.

Il est également possible de combiner ces opérateurs logiques afin d'affiner la recherche d'un atome spécifique répondant à plusieurs conditions. Par exemple, [C,N&O] correspond à la recherche d'un carbone ou d'un azote lié à un atome d'oxygène, tandis que [C&H1&X4] correspond à un carbone lié à un hydrogène et possédant quatre connexions au total.

Plusieurs blocs, tels que [Atome1,condition1,condition2] et [Atome2,condition1,condition2], peuvent aussi être combinés pour sélectionner des motifs moléculaires plus complexes. L'exemple suivant, [C&H1&X4][O][C&H3], correspond à la recherche d'un carbone possédant quatre connexions au total, lié à au moins un hydrogène et un atome d'oxygène lui-même lié à un carbone lié à trois hydrogènes.

SMIRKS (SMiles and Reaction Kinetics SMart)

Le SMIRKS est un langage qui combine les SMILES et SMARTS. Il permet de générer des réactions chimiques de manière computationnelle. La combinaison des SMILES et SMARTS permet d'appliquer des transformations en ajoutant les SMILES/SMARTS d'un réactif (molécule avant transformation) suivis d'un produit (molécule après transformation). Par exemple, la transformation d'un alcool en cétone se notera de la manière suivante : [OH:2][C:1]>>[O]=[C:1]. Il est également possible d'annoter les éléments lors de la réaction en leur attribuant un numéro, qui peut servir de marquage pour cibler un motif moléculaire parmi plusieurs identiques au sein d'une structure.

En 2006, Greg Landrum a développé RDKit[136], un outil open-source dédié à la manipulation et à l'analyse des structures moléculaires. Son intégration dans des langages largement utilisés par la communauté scientifique, comme Python, a facilité l'accès aux systèmes de notation SMILES, SMARTS et SMIRKS pour un large public. Parmi les nombreuses fonctionnalités proposées, celle des transformations chimiques est prise en charge par RDKit, qui propose un système flexible combinant la recherche de sous-structures et leurs transformations.

La figure 2-2 présente un chemin réactionnel possible pris en compte dans notre stratégie à partir de la vinclozoline, Deux biotransformations successives sont illustrées (ouverture d'une lactone à partir d'une hydrolyse et une double oxydation sur une double liaison) ainsi que les structures des métabolites prédits

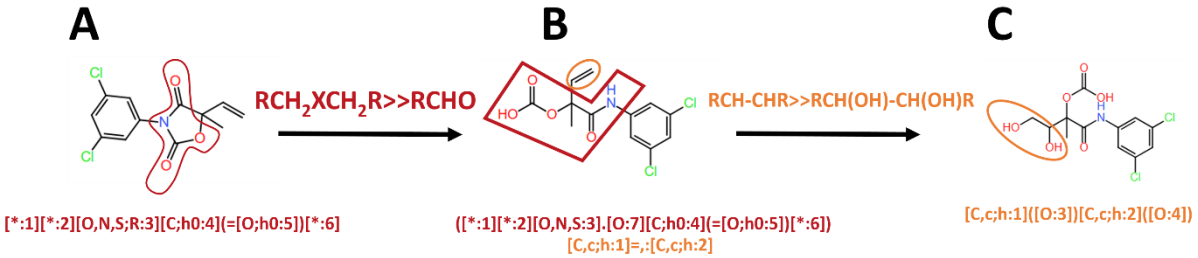


Figure 2-2 : (A) Structure de la Vinclozoline, le motif moléculaire recherché pour subir une hydrolyse avec une ouverture de cycle est entouré en rouge et encodé en SMARTS. (B) Structure du métabolite prédit à partir de la réaction en A, le motif moléculaire remplacé est entouré en rouge; le motif moléculaire recherché pour une nouvelle réaction de biotransformation telle qu'une dihydroxylation d'une double liaison est entouré en orange. (C) Métabolite prédit à partir de la réaction en B, avec le motif modifié en orange.

2.2.4.2. Procédure pour la génération des métabolites potentiels

Il est essentiel de pouvoir prédire les structures des métabolites des xénobiotiques lors d'une approche « suspect screening » appliquée à l'analyse de matrices biologiques pour mieux détecter les composés transformés dans l'organisme. En effet, de nombreux contaminants subissent des modifications métaboliques qui changent leur structure chimique de départ, rendant leur détection plus complexe. La prédiction des métabolites, notamment par des techniques *in silico*, permet d'anticiper ces transformations et d'élargir la liste des substances ciblées pour inclure les produits dérivés. Cette démarche améliore ainsi la capacité à identifier des marqueurs d'exposition, offrant une vision plus complète de la présence et de l'impact des xénobiotiques dans les matrices biologiques analysées.

La prédiction computationnelle des métabolites a déjà été décrite dans la littérature, notamment l'outil BioTransformer[110], qui permet de générer des métabolites *in silico* provenant de diverses sources (cytochrome 450, activité microbienne, etc.). Utilisant l'intelligence artificielle, cet outil propose les métabolites les plus probables en fonction de ce qui a pu être référencé dans des banques de données. Néanmoins, les transformations appliquées sont limitées à celles proposées par ces outils, ce qui peut poser problème lors d'une recherche exploratoire

Dans l'étude réalisée, nous proposons une méthode plus flexible permettant d'utiliser n'importe quelle réaction de biotransformation pour prédire les métabolites de molécules suspectées.

Le fonctionnement en est le suivant :

i) Les molécules parentes suspectées à rechercher sont encodées au format SMILES,

ii) Les réactions de transformations considérées sont ensuite appliquées sur ces molécules,

iii) Lorsqu'une ou plusieurs fonctions pouvant subir des transformations (motifs SMARTS définis) sont présentes dans les molécules suspectées (Sub), de nouvelles structures sont générées avec de nouveaux SMILES (new). À noter que si plusieurs motifs sont présents dans une molécule, le nombre de métabolites générés est égal au nombre de sous-structures détectées. Ainsi, une réaction de transformation peut être répétée autant de fois que nécessaire, et les produits

issus de la première étape seront utilisés de nouveaux pour générer d'autres métabolites.

2.3. Etude préliminaire et évaluation de l'approche développée

2.3.1. Recherche de marqueurs d'exposition

L'une des difficultés dans la recherche de marqueurs d'exposition réside sans doute dans la sensibilité de détection. Il est important de pouvoir évaluer la capacité de l'approche mise en place ici pour détecter des marqueurs d'exposition qui pourraient être présents à l'état de trace. Un premier travail a consisté à cibler des composés d'une liste réduite de marqueurs d'exposition. Cette liste contient 125 molécules parmi les 300 contaminants détectés au total dans les aliments pouvant être consommés par les mères des cohortes EDEN et ELFE, mais pas les minéraux, qui ne peuvent pas être détectés dans les conditions expérimentales utilisées pour analyser les matrices périnatales.[44] La détection de ces composés étant inhérente au projet auquel le sujet de thèse se rapporte, il était nécessaire d'obtenir les premières informations sur la possibilité de réaliser leur détection étant donné leur faible taux de contamination au sein des matrices alimentaires, qui est probablement la première source d'exposition pour les mères. Ainsi, la première étape a consisté à rechercher uniquement les signaux correspondants aux pics mono-isotopiques des formes protonées ou déprotonées de ces 125 substances chimiques (uniquement les molécules mères sans tenir compte des métabolites) dans les données métabolomiques produites par LC-HRMS en modes positif et négatif sur 308 échantillons de méconium de la cohorte EDEN, pour un nombre total de 616 fichiers de données.

51 et 37 signaux ont respectivement été détectés dans les jeux de données métabolomiques acquis en modes positif et négatif. Ces signaux pourraient correspondre à 59 des 125 contaminants de la liste de référence, pour une tolérance dans la mesure en masse de ± 3 ppm (Figure 2-3). Cependant, pour un même composé, plusieurs candidats sont possibles et, d'autre part, l'intensité des signaux observés est relativement faible, limitant les champs possibles pour des analyses supplémentaires, notamment les expériences MS/MS qui requièrent une intensité minimale pour obtenir des spectres MS/MS exploitables. Deux problèmes se sont ainsi présentés : le premier est la difficulté à discriminer des candidats fiables parmi de nombreux candidats pour un seul composé détecté et le deuxième est la non-détection de candidats pour un grand nombre de composés.

Dans le cas présent, ceci peut s'expliquer par le fait que des contaminants peuvent se cumuler dans le méconium pendant six des neuf mois de la grossesse. Cette matrice étant très dense, la détection de certains composés est donc difficile. De plus, la nature des composés qui la composent est peu connue. Il est possible que les biotransformations des xénobiotiques lors de leur passage dans l'organisme de la mère altèrent complètement leur forme d'origine, ce qui rend impossible la détection de ces derniers dans cette matrice sous la forme parentale.

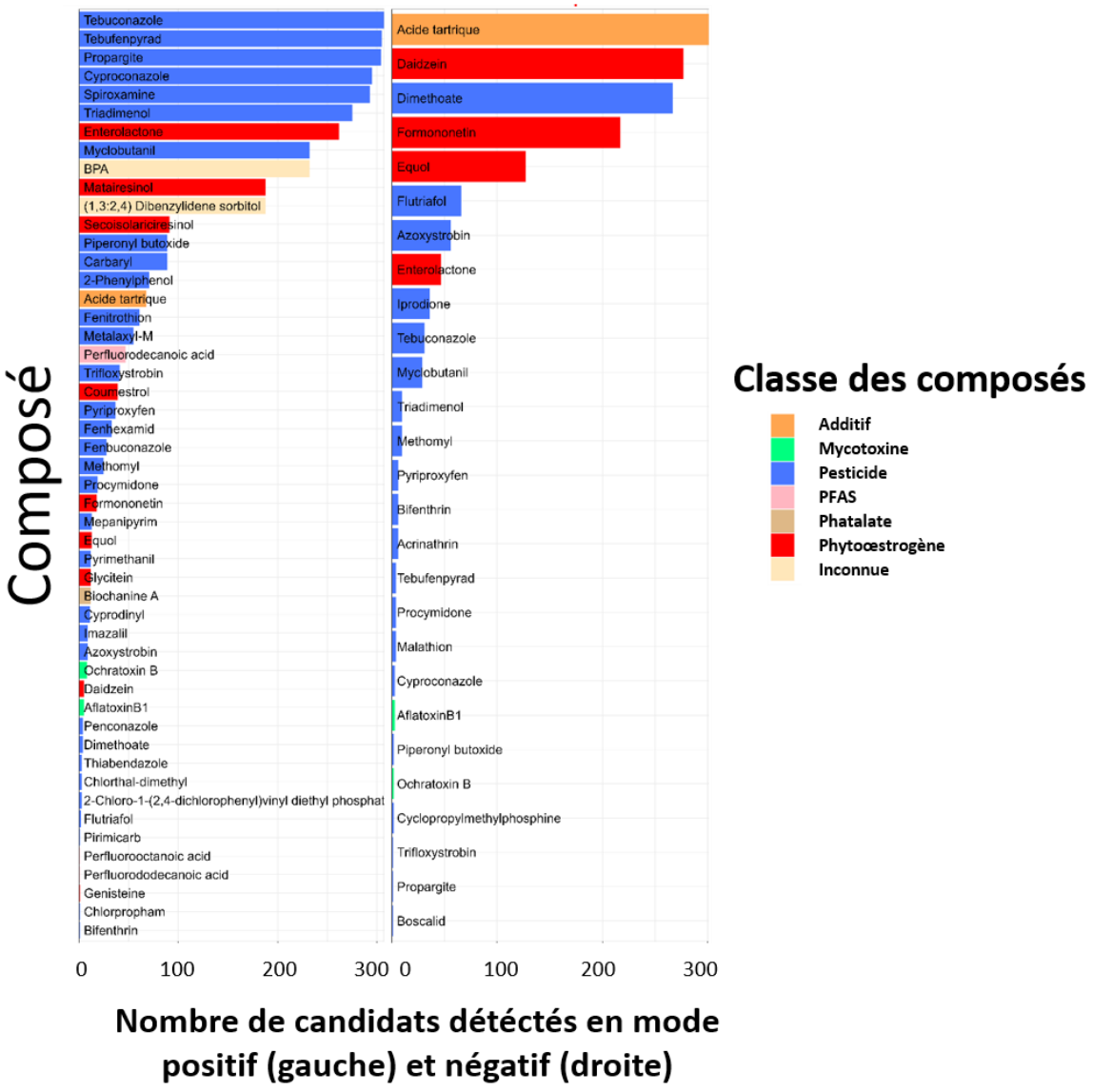


Figure 2-3 : Liste des candidats (ayant des valeurs *m/z* similaires à ± 3 ppm à ceux des contaminants recherchés) détectés dans les données produites en modes positif (gauche) et négatif (droite). L'axe des abscisses représente le nombre des échantillons dans lesquels au moins un candidat est détecté.

Ainsi, le développement de la stratégie de traitement des données présenté dans la suite inclut une étape préliminaire de prédiction des métabolites *in silico* afin d'enrichir la liste initiale des composés à rechercher et une recherche plus précise de ces composés en se fondant notamment sur leurs massifs isotopiques, sur l'extraction des spectres de masse afin d'éviter différents biais liés au prétraitement des données, ainsi que sur l'utilisation des filtres de qualité permettant de s'affranchir des signaux artefactuels.

2.3.2. Evaluation de l'approche développée

2.3.2.1. Prédiction *in silico* des métabolites connus

La base de données HMDB (Human Metabolome Database) répertorie des informations telles que les propriétés chimiques et biologiques des composés, ainsi que des données expérimentales et spectrales obtenues par RMN et LC-HRMS. Elle permet également de faire le lien avec d'autres bases de données grâce aux identifiants spécifiques (ID) des composés, et fournit des informations sur leur profil métabolique ainsi que les matrices biologiques (sang, urine, etc.) dans lesquelles ils ont été détectés, en accord avec la littérature scientifique disponible. Les données sur le profil métabolique renvoient vers la base de données SMPDB, qui relie les composés à des voies métaboliques communes, permettant de mieux comprendre leur rôle au sein de ces voies métaboliques. Bien que ces bases de données se concentrent sur le métabolome humain, certains xénobiotiques couramment observés lors de ce type d'études sont également répertoriés (médicaments, polluants environnementaux, etc.).

Afin de tester l'approche présentée ici, une première étape a consisté à sélectionner les composés xénobiotiques dont le profil métabolique est renseigné dans la base de données HMDB et qui sont également référencés dans la banque de données Drug Bank, qui est spécifique aux médicaments. Ainsi, 60 composés exogènes présentant un profil métabolique décrit ont pu être sélectionnés et sont présentés dans le tableau 2-1. Il convient de noter que ces voies métaboliques produisent des métabolites de xénobiotiques, mais qu'elles font également intervenir des molécules endogènes nécessaires à leur génération. Une liste d'exclusion, présentée en annexe 1 a donc été générée pour exclure ces molécules. Un total de 214 métabolites a ainsi pu être extrait.

Tableau 2-1 : Liste des 60 contaminants extraits de la base de données HMDB et leurs SMILES

Nom	SMILES
Imipramine	CN(C)CCCN1C2=CC=CC=C2CCC2=CC=CC=C12
Acetaminophen	CC(=O)NC1=CC=C(O)C=C1
Valproic acid	CCCC(CCC)C(O)=O
Omeprazole	COC1=CC2=C(C=C1)N=C(N2)S(=O)CC1=NC=C(C)C(OC)=C1C
Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(O)=O
Nicotine	[H][C@]1(CCCN1C)C1=CC=CN=C1
Codeine	[H][C@@]12OC3=C(OC)C=CC4=C3[C@@]11CCN(C)[C@]([H])(C4)[C@]1([H])C=C[C@@H]2O
Lansoprazole	CC1=C(OCC(F)(F)F)C=CN=C1CS(=O)C1=NC2=CC=CC=C2N1
Clopidogrel	[H][C@@](N1CCC2=C(C1)C=CS2)(C(=O)OC)C1=CC=CC=C1Cl
Celecoxib	CC1=CC=C(C=C1)C1=CC(=NN1C1=CC=C(C=C1)S(N)(=O)=O)C(F)(F)F
Venlafaxine	COC1=CC=C(C=C1)C(CN(C)C)C1(O)CCCCC1
Pantoprazole	COC1=C(OC)C(CS(=O)C2=NC3=C(N2)C=C(OC(F)F)C=C3)=NC=C1
Rabeprazole	COCCCOC1=C(C)C(CS(=O)C2=NC3=CC=CC=C3N2)=NC=C1
Rosiglitazone	CN(CCOC1=CC=C(CC2SC(=O)NC2=O)C=C1)C1=CC=CC=N1
Citalopram	CN(C)CCCC1(OCC2=C1C=CC(=C2)C#N)C1=CC=C(F)C=C1
Ramipril	[H][C@@]12CCC[C@]1([H])N([C@@H](C2)C(O)=O)C(=O)[C@H](C)N[C@@H](CCC1=CC=CC=C1)C(=O)OCC
Tramadol	COC1=CC=CC(=C1)[C@@]1(O)CCCC[C@@H]1CN(C)C
Ticlopidine	ClC1=CC=CC=C1CN1CCC2=C(C1)C=CS2
Nevirapine	CC1=C2NC(=O)C3=C(N=CC=C3)N(C3CC3)C2=NC=C1
Phenytoin	O=C1NC(=O)C(N1)(C1=CC=CC=C1)C1=CC=CC=C1
Lidocaine	CCN(CC)CC(=O)NC1=C(C)C=CC=C1C
Tenofovir	C[C@H](CN1C=NC2=C1N=CN=C2N)OCP(O)(O)=O
Methadone	CCC(=O)C(CC(C)N(C)C)(C1=CC=CC=C1)C1=CC=CC=C1
Meprobamate	CCCC(C)(COC(N)=O)COC(N)=O
Sorafenib	CNC(=O)C1=NC=CC(OC2=CC=C(NC(=O)NC3=CC(=C(Cl)C=C3)C(F)(F)F)C=C2)=C1
Gemcitabine	NC1=NC(=O)N(C=C1)[C@@H]1O[C@H](CO)[C@@H](O)C1(F)F
Fluoxetine	CNCCC(OC1=CC=C(C=C1)C(F)(F)F)C1=CC=CC=C1
Fosinopril	CCC(=O)OC(OP(=O)(CCCCC1=CC=CC=C1)CC(=O)N1C[C@@H](C[C@H]1C(O)=O)C1CCCCC1)C(C)C
Cimetidine	C\N=C(\NCCSCC1=C(C)NC=N1)NC#N
Trandolapril	[H][C@@]12C[C@H](N(C(=O)[C@H](C)N[C@@H](CCC3=CC=CC=C3)C(=O)OCC)[C@@]1([H])CCCC2)C(O)=O
Cyclophosphamide	ClCCN(CCCl)P1(=O)NCCCO1
Benazepril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@H]1CCC2=CC=CC=C2N(CC(O)=O)C1=O
Carbamazepine	NC(=O)N1C2=CC=CC=C2C=CC2=CC=CC=C12
Enalapril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@@H](C)C(=O)N1CCC[C@H]1C(O)=O
Prednisone	[H][C@@]12CC[C@](O)(C(=O)CO)[C@@]1(C)CC(=O)[C@@]1([H])[C@@]2([H])CCC2=CC(=O)C=C[C@]12C
Tamoxifen	CC\C(=C(/C1=CC=CC=C1)C1=CC=C(OCCN(C)C)C=C1)C1=CC=CC=C1
Mycophenolate mofetil	COC1=C(C\C=C(/C)CCC(=O)OCCN2CCOCC2)C(O)=C2C(=O)OCC2=C1C
Moexipril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@@H](C)C(=O)N1CC2=CC(OC)=C(OC)C=C2C[C@H]1C(O)=O
Lamivudine	NC1=NC(=O)N(C=C1)[C@@H]1CS[C@H](CO)O1
Adefovir Dipivoxil	CC(C)(C)C(=O)OCOP(=O)(COCCN1C=NC2=C(N)N=CN=C12)OCOC(=O)C(C)(C)C
Irinotecan	CCC1=C2CN3C(=CC4=C(COC(=O)[C@]4(O)CC)C3=O)C2=NC2=C1C=C(OC(=O)N1CCC(CC1)N1CCCCC1)C=C2
Etoposide	[H][C@]12COC(=O)[C@]1([H])[C@H](C1=CC(OC)=C(O)C(OC)=C1)C1=CC3=C(OCO3)C=C1[C@H]2O[C@@H]1O[C@]2([H])CO[C@@H](C)O[C@@]2([H])[C@H](O)[C@H]1O
Oxcarbazepine	NC(=O)N1C2=CC=CC=C2CC(=O)C2=CC=CC=C12

Nom	SMILES
Quinapril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@@H](C)C(=O)N1CC2=CC=CC=C2C[C@H]1C(O)=O
Felbamate	NC(=O)OCC(COC(N)=O)C1=CC=CC=C1
Azathioprine	CN1C=NC(=C1SC1=NC=NC2=C1NC=N2)[N+]([O-])=O
Doxorubicin	COC1=CC=CC2=C1C(=O)C1=C(O)C3=C(C[C@](O)(C[C@@H]3 O[C@H]3C[C@H](N)[C@H](O)[C@H](C)O3)C(=O)CO)C(O)=C1C2=O
Felodipine	CCOC(=O)C1=C(C)NC(C)=C(C1C1=C(Cl)C(Cl)=CC=C1)C(=O)OC
Mycophenolic acid	COC1=C(C\C=C(/C)CCC(O)=O)C(O)=C2C(=O)OCC2=C1C
Capecitabine	CCCCCOC(=O)NC1=NC(=O)N(C=C1F)[C@@H]1O[C@H](C)[C@@H](O)[C@H]1O
Doxepin	[H]C(CCN(C)C)=C1C2=CC=CC=C2COC2=CC=CC=C12
Ifosfamide	ClCCNP1(=O)OCCCN1CCCl
Levomethadyl Acetate	CC[C@H](OC(C)=O)C(C[C@H](C)N(C)C)(C1=CC=CC=C1)C1=CC=CC=C1
Clomipramine	CN(C)CCCN1C2=CC=CC=C2CCC2=C1C=C(Cl)C=C2
Fosphenytoin	OP(O)(=O)OCN1C(=O)NC(C1=O)(C1=CC=CC=C1)C1=CC=CC=C1
Cilazapril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@H]1CCCN2CCC[C@H](N2C1=O)C(O)=O
Spirapril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@@H](C)C(=O)N1CC2(C[C@H]1C(O)=O)SCCS2
Artemether	[H][C@@]12CC[C@@H](C)[C@]3([H])CC[C@@]4(C)OO[C@@]13[C@]([H])(O[C@H](OC)[C@@H]2C)O4
7-Ethyl-10-hydroxycamptothecin	CCC1=C2C=C(O)C=CC2=NC2=C1CN1C2=CC2=C(COC(=O)[C@]2(O)CC)C1=O
Temocapril	CCOC(=O)[C@H](CCC1=CC=CC=C1)N[C@H]1CS[C@@H](CN(CC(O)=O)C1=O)C1=CC=CS1

Ici, la liste des réactions correspond aux biotransformations dites de phase 1 ou 2, qui sont communément observées lors d'études sur le métabolisme des xénobiotiques.[122] Cette liste de biotransformations est présentée dans le tableau 2-2.

L'utilisation des SMARTS n'est pas simple pour bien prédire la structure des métabolites. Des erreurs d'encodage peuvent conduire à la génération de métabolites de structures aberrantes et donc inexistantes. De même, un encodage insuffisamment spécifique peut induire des transformations sur de mauvais motifs. Afin de s'assurer du bon fonctionnement de ces transformations, chacune d'entre elles a été testée et affinée sur des SMILES de test comportant le motif recherché (sub) ainsi que d'autres motifs similaires supposés ne pas réagir.

Tableau 2-2 : Liste des biotransformations sous forme de SMARTS

Motif à rechercher	Motif de remplacement	Réaction
[S;h1:1]	[S;h0:1](=[O;h0:2])(=[O;h0:3])[O;h1:4]	RSH>>RSO_3H
[SX2;h0:1]	[SX4;h0:1](=[O;h0:2])=[O;h0:3]	RSR>>RSO_2R
[O;h1:1]	[O;h0:1][C;h3:2]	ROH>>$ROCH_3$
[C;h0:1](=[O;h0:2])[O;h1:3]	[C:1][O;h1:3]	RCOOH>>RCH_2OH
[C;h0:1]#[N;h0:2]	[C;h0:1](=[O;h0:3])[N;h2:2]	RCN>>$RCONH_2$
[S;h1:1]	[S:1]([O;h1:2])	SH>>SOH
[N;h1:1]	[N;h0:1]([O;h1:2])	NH>>NOH
[*:1][C;h2:2][*:3]	[*:1][C;h0:2](=[O;h0:4])[*:3]	RCH_2R>>RCOR
[C;h0:1](=[O;h0:2])	[C;h0:1]([O;h1:2])	RCOR>>RCH(OH)R
[C;h2:1][C;h3:2]	[C;h2:1](=[O;h0:2])[O;h1:3]	RCH_2CH_3>>RCOOH
[N;h0:1](=[O;h0:2])(=[O;h0:3])	[N;h1:1]([O;h1:2])	RNO_2>>RNHOH
[O;h1:1][N:2]	[N:2]	RNHOH>>RNH_2
[C:1](=[O:2])([O;h1:3])	[C;h2:1]([O;h1:3])	RCOOH>>RCH_2OH
[C;h2:1][C;h2:2][C;h2:3][C;h2:4]	[C;h1:1]=[C;h1:2][C;h1:3]=[C;h1:4]	$RCH_2CH_2CH_2CH_2R$>>RCH=CH–CH=CHR
[C;h2:1][N;h2:2]	[C:1]#[N;h2:2]	RCH_2NH_2>> RCN
[N;h:1][N;h2:2]	[N;h0:1]=[N;h1:2]	$RNH–NH_2$>>RN=NH
[C;h2:1][C;h2:2]	[C;h1:1]=[C;h0:2]	RCH(OH)R>> RCOR
[C;h1:1][N;h2:2]	[C;h0:1](=[O;h0:2])	$RCHNH_2R$>>RCOR
[C,c;h:1]=,:[C,c;h:2]	[C,c;h:1]([O:3])[C,c;h:2]([O:4])	RCH-CHR>>RCH(OH) – CH(OH)R
[S;h1:1]	[S;h0:1][O;h1:2]	SH>>SOH
[C;h2:1][N;h2:2]	[C;h0:1](=[O:3])[O;h1:4]	RCH_2NH_2>>RCOOH
[C;h2:1][O;h1:2]	[C;h0:1](=[O:2])[O;h1:3]	RCH_2OH>>RCOOH
[C,c;h1:1]=,:[C,c;h1:2]	[C,c;h2:1][C,c;h2:2]	RCH=CHR>>RCH_2CH_2R

Motif à rechercher	Motif de remplacement	Réaction
[N;h1:1][N;h1:2][*:3][C;h1,h0:4](=[S;h0:5])	[N;h1:1][N;h1:2][*:3][C;h1,h0:4](=[O;h0:6])	RNHNHRC=S>>RNHNHRC=O
[C;h2:1][C;h1:2]([O;h1:3])	[C;h2:1]=[C;h1:2]	RCH_2CHOHR>>RCH=CHR
[C;h1:1]=[N;h0:2][O;h1:3]	[C;h0:1]#[N;h0:2]	RCH=N–OH>>R–CN
[C,c:1][C;h0:2](=[O;h0:3])[C,c:4]	[C,c:1][C,c:4]	RCOR>>RR
[N;h1:1][C;h2:2][C;h3:3]	[N;h2:1]	$RNHCH_2CH_3$>>RNH_2
[*:1][N;h0&O0:2]=[O;h0:3]	[*:1]	RNO>>RH
[*:1][N;h0:2](=[O:3])[O:4]	[*:1][N;h2:5]	RNO_2>>RNH_2
[C,c:1][O;h0:2][C;h3:3]	[C,c:1]	$ROCH_3$>>RH
[*:1][C;h0:2](=[O;h0:3])[O;h1:4]	[*:1]	RCOOH>>RH
[O;h0:1][N;h0:2](=[O:3])[O:4]	[O;h1:1]	$RONO_2$>>ROH
[*:1][*:2][OX2,NH,SX2;!R:3][*:4][*:5]	[O:6][*:4][*:5]	RCH_2XCH_2R>>RCHO
[*:1][*:2][O,N,S;!R:3][*:4][*:5]	[*:1][*:2][O,N,S;!R:3]	RCH_2XCH_2R>>RCHO
[*:1][*:2][OX2,N,SX2;R:3][C;h0:4](=[O;h0:5])[*:6]	([*:1][*:2][O,N,S:3].[O:7][C;h0:4](=[O;h0:5])[*:6])	RCH_2XCH_2R>>RCHO
[O;h:1][c:2][c;h1:3][c:4][c:5]([O,N;h:6])	[O,N:1][c:2][c:3]([S][C;h2][C;h1]([N;h1][C;h0](=[O;h0])[C;h2][C;h2][C;h1]([N;h2])[C;h0](=[O;h0])[O;h1])[C;h0](=[O;h0])[N;h1][C;h2][C;h0](=[O;h0])[O;h1])[c:4][c:5]([O,N:6])	glutathion – Conjugaison
[c:1][c:2]([O,N;h1:3])[c:4]([O,N;h1:5])[c;h1:6]	[c:1][c:2]([O,N;h1:3])[c:4]([O,N;h1:5])[c;h1:6]([S][C;h2][C;h1]([N;h1][C;h0](=[O;h0])[C;h2][C;h2][C;h1]([N;h2])[C;h0](=[O;h0])[O;h1])[C;h0](=[O;h0])[N;h1][C;h2][C;h0](=[O;h0])[O;h1])	Glutathion – Conjugaison

Motif à rechercher	Motif de remplacement	Réaction
[N;h1:1]([O;h1:2])	[N+;h:1](=[O:2])([O-:3])	N-oxide>>N-hydroxy
[*:1][C;h2:2][N;h2:3]	[*:1][C;h2:2][O;h1:3]	RCH_2NH_2>>RCH_2OH
[*:1][C;h1:2]([O;h1:3])[C;h3:4]	[*:1][C;h1:2](=[O;h0:3])[O;h1:4]	R-CH(OH)CH_3>>RCOOH
[n;h:1]	[O,N,S;h1:1]([C;h1:2]1[O;h0:3][C;h1:4]([C;h0:5](=[O;h0:6])([O;h1:7]))[C;h1:8]([O;h1:9])[C;h1:10]([O;h1:11])[C;h1:12]([O;h1:13])1)	Glucuroconjugaison
[O,N,S;h:1]	[O,N,S;h1:1]([C;h1:2]1[O;h0:3][C;h1:4]([C;h0:5](=[O;h0:6])([O;h1:7]))[C;h1:8]([O;h1:9])[C;h1:10]([O;h1:11])[C;h1:12]([O;h1:13])1)	Glucuroconjugaison
[O;h1:1][!$(C(=O)(C))&!S:2]	[O:1]([!$(C(=O)(C))&!S:2])[S:3](=[O:4])(=[O:5])([O;h1:6])	Sulfoconjugaison
[N&h:1][c:2]	[N&h:1]([c:2])[S:3](=[O:4])(=[O:5])([O;h1:6])	Sulfoconjugaison
[C&h2:1][F,Cl,Br:2]	[C&h2:1][O&h1:2]	Déhalogénation
[C,c:1][F,Cl,Br:2]	[C,c:1]	Déhalogénation
[*&h:1][C&h:2][F,Cl,Br:3]	[*:1]=[C:2]	Dehalogenation
[c&h:1]	[c:1][O:2]	RH>>ROH

Test de l'approche

Afin de tester cette approche, je me suis intéressé, d'une part, aux structures générées par l'approche utilisée mais, également, aux masses générées qui constituent la principale caractéristique des signaux des marqueurs d'exposition à rechercher dans les données LC-HRMS.

Les structures possibles de métabolites ont été générées à partir des 60 composés exogènes en considérant deux biotransformations successives et une seule réaction de biotransformation de phase II, afin de limiter le nombre de structures et, par conséquent, de masses générées. Ces structures ont ensuite été comparées aux 214 métabolites référencés sur les bases de données HMDB et SMPDB afin d'évaluer l'approche utilisée par rapport aux profils métaboliques référencés dans la littérature. Les valeurs des masses obtenues ont été comparées à celles générées par une approche par différence de masses. Cette approche consiste à additionner ou soustraire les masses des atomes incorporés ou retirés en fonction des réactions considérées, sans prendre en compte la structure des molécules. La figure 2-4 schématise les tests effectués.

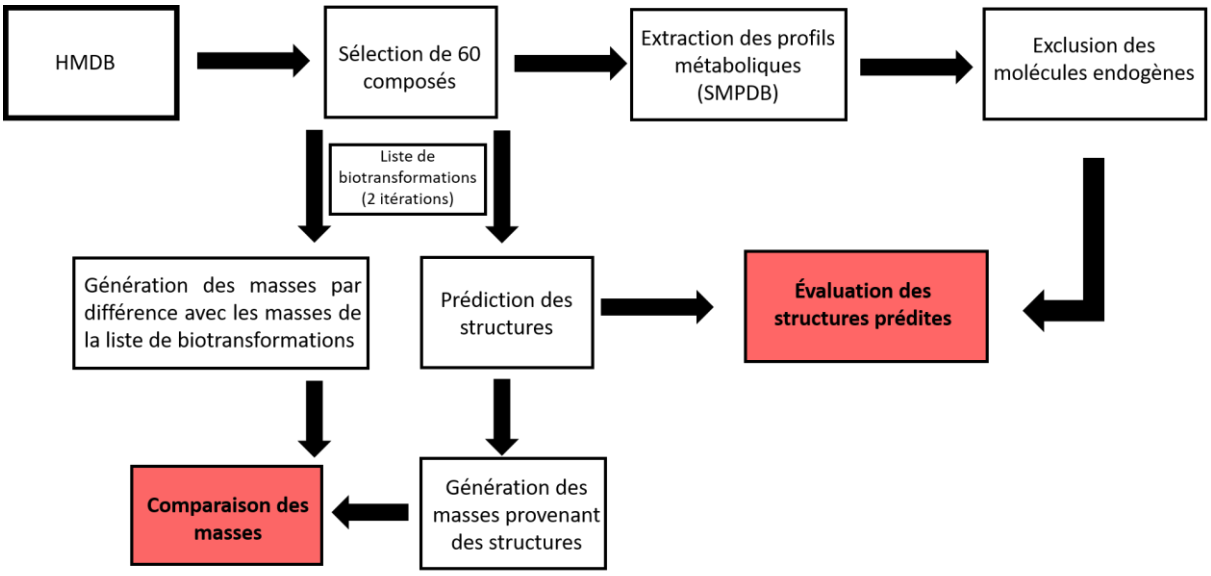


Figure 2-4 : Schéma représentant le protocole d'évaluation de la génération des structures par comparaison avec 60 composés exogènes extraits de la banque de données HMDB et ayant un profil métabolique répertorié dans la base de données SMPDB ainsi que l'évaluation du nombre de masses qui en résulte par comparaison avec une approche par différence de masses

Comparaison des structures

152 métabolites (71 %) ont été correctement générés par la stratégie développée ici (Figure 2-5). Parmi les 62 métabolites non générés (29 %), 5 correspondent à des molécules endogènes impliquées dans le processus de biotransformation et qui n'ont pas été prises en compte dans la liste d'exclusion, 20 sont issus des voies de métabolisation impliquant plus de deux réactions successives qui dépassent le nombre d'itérations sélectionnées (2), et 37 proviennent de réactions non répertoriées dans la liste de biotransformations choisie car potentiellement trop spécifiques. Même s'il est tout à fait possible d'ajouter d'autres réactions de transformation, seuls 17 parmi 37 sont a priori faciles à encoder sous forme de SMARTS contrairement aux 20 restants qui nécessitent une attention particulière. À noter que la totalité des métabolites a été générée pour 30 composés sur les 60 initiaux.

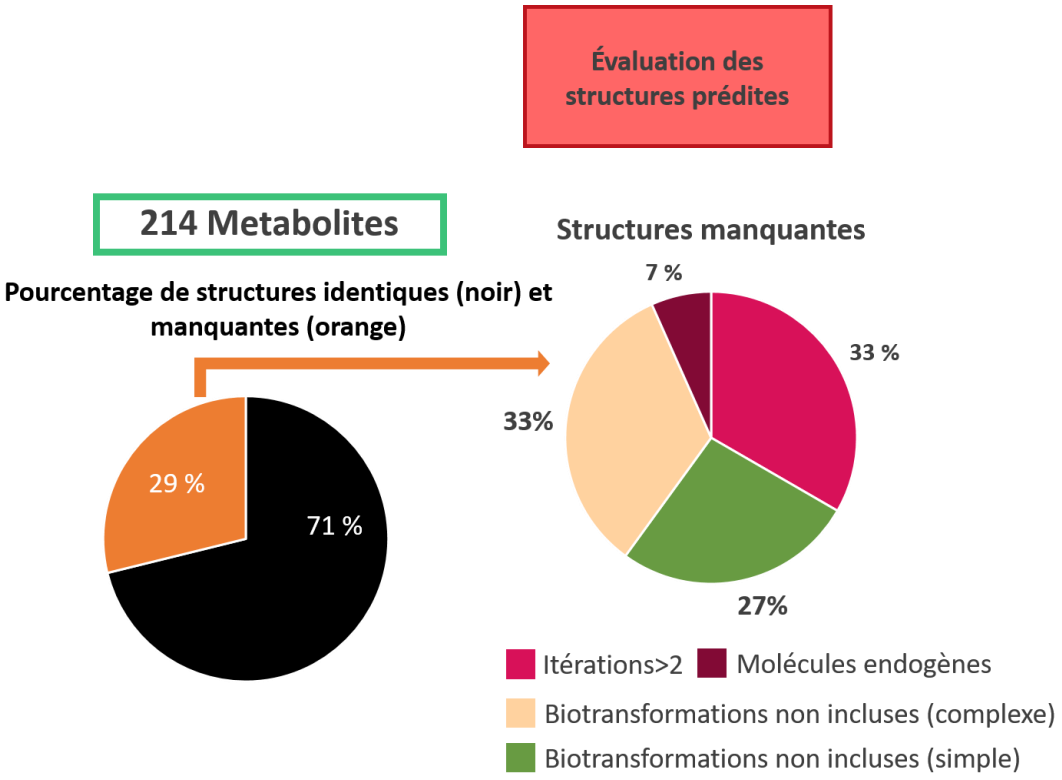


Figure 2-5 : Comparaison du nombre de structures prédites *in silico* avec les 214 métabolites décrits dans la base de données

Comparaison des masses

40 000 masses uniques ont été générées par calcul de différence de masses alors que l'approche développée, qui prend en compte la structure chimique des molécules a permis de réduire ce nombre à 2 384, soit une réduction d'un facteur 17 (Figure 2-6). Cette réduction significative permet de limiter la génération de masses incohérentes avec des structures de métabolites inexplicables et, par conséquent, de réduire le nombre de faux-positifs qui nécessiteraient une curation manuelle plus importante. De plus, certaines réactions ont été exclues, notamment les homolyses et autres réactions nécessitant une coupure au sein de la structure et dont les masses ne peuvent pas être générées avec une approche par différence de masses, car il est *a priori* impossible de définir les constituants générés par une coupure avec uniquement la masse comme paramètre.

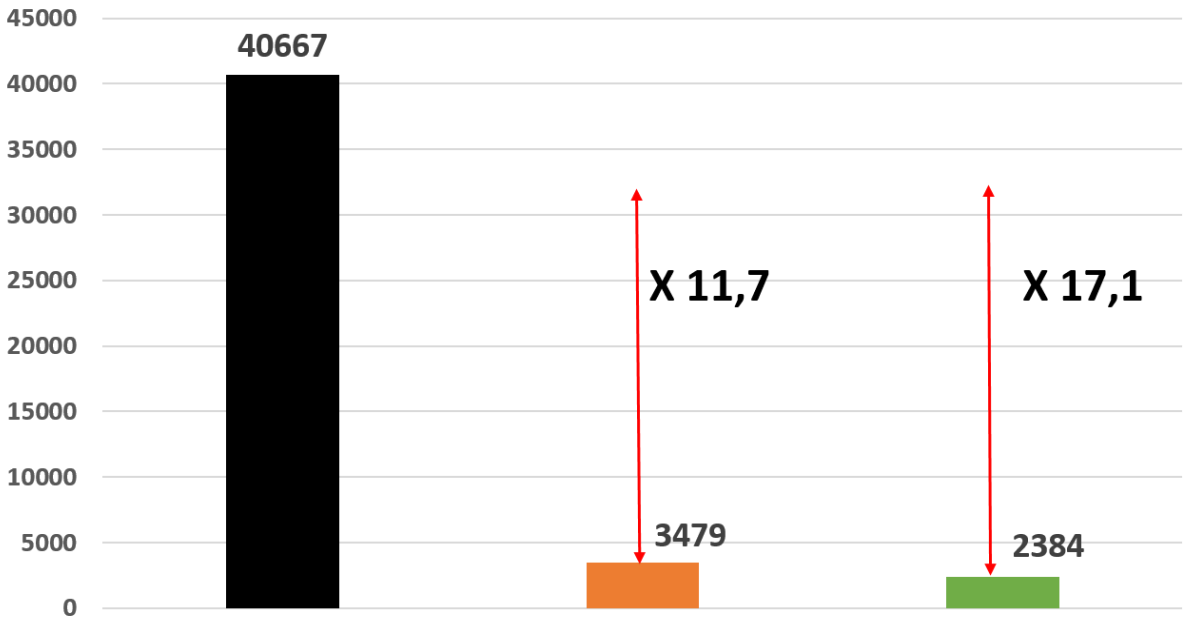


Figure 2-6 : Nombre de masses générées par calcul de différence de masses (en noir) et par l'approche développée en considérant deux itérations (en considérant les homolyses en orange et sans les considérer en vert)

Il est à noter qu'une difficulté majeure a été identifiée avec cette méthode : plus le nombre de réactions appliquées est important, plus le nombre de structures générées l'est également, ce qui conduit à une augmentation du temps de traitement lors de la prédiction des métabolites.

L'application de la liste de biotransformations à 100 composés quelconques, en fonction de plusieurs itérations, présentée dans le tableau 2-3, illustre le temps nécessaire à chaque étape. On peut noter une forte augmentation du temps de processus et du nombre de structures à partir de la troisième itération. Cette limite est problématique dans la recherche du profil métabolique complet d'une grande liste de xénobiotiques comme celle d'une banque de données, que ce soit en termes de temps nécessaire pour la génération des structures ou pour la recherche des masses résultantes au sein des jeux de données.

Tableau 2-3 : Durée nécessaire pour la prédiction des métabolites à partir de 100 composés et le nombre de structures générées en fonction du nombre d'itérations (biotransformations successives)

Nombre d'itérations	Temps	Nombre de structures
0	0	100
1	16 secondes	1 625
2	2,5 minutes	28 031
3	47,5 minutes	332 437

2.3.2.2. Examen de la procédure de génération et de recherche des signaux correspondants aux marqueurs potentiels d'exposition

Si la génération des masses correspondantes aux pics mono-isotopiques des marqueurs d'expositions pour une liste de composés d'intérêt constitue une première étape, la génération des signaux susceptibles d'être observés par LC-HRMS (formes ionisées, isotopes, etc.), leur détection et la vérification de leur structure hypothétique constituent les étapes suivantes.

Tout d'abord les données brutes ont été traitées sous R en utilisant le package XCMS permettant le prétraitement des données dans le but de définir tous les pics détectés sous la forme d'un tableau (matrice des données) contenant des caractéristiques des ions détectés (features) servant de base de travail pour la détection des signaux au sein de la matrice biologique.

Les paramètres utilisés pour définir les pics, « peak-picking » (définissant les pics) et « datagrouping » (permettant d'assigner une feature aux pics de mêmes masses et temps de rétention inter-échantillon), sont présentés dans la partie méthodologie (2.2.3. Prétraitement des données).

Comme défini dans l'introduction (voir 1.4.2.1. Recherche de molécules suspectées), plusieurs signaux peuvent correspondre à un même composé selon l'abondance relative naturelle des atomes (isotopes) le constituant (massif isotopique). Néanmoins la recherche de la totalité du massif isotopique pour chaque marqueur d'exposition augmente drastiquement le temps de recherche par l'algorithme utilisé et n'est pas réellement pertinente, notamment lorsque des signaux correspondant à des isotopes de faible abondance sont détectés sans le pic mono-isotopique ou d'autres isotopes de plus haute abondance. Par exemple, si un signal correspondant à la présence d'un carbone 13 est détecté dans le cas d'un composé organique quelconque sans la présence du pic mono-isotopique ou de l'isotope ^{12}C, il s'agira d'un faux positif.

Afin de diminuer le temps de recherche des marqueurs d'expositions et de rechercher les signaux correspondants aux massifs isotopiques des composés seulement dans le cas où le mono-isotopique est observé, la première étape consiste à générer les signaux correspondant au pic mono-isotopique pour chaque marqueur d'exposition, c'est-à-dire en ne considérant que la présence des isotopes les plus abondants pour chaque atome. Ces signaux sont ensuite

recherchés dans la matrice des données avec une fenêtre de tolérance en masse définie à 5 ppm, ceci afin d'identifier les features correspondantes.

Si un signal correspondant au pic mono-isotopique est observé, alors les signaux du massif isotopique théorique correspondant ayant une abondance relative supérieure à 0,4 % par rapport au pic le plus intense sont pris en compte

La recherche de ces signaux n'est pas effectuée directement sur la matrice des données, car l'application des algorithmes, notamment de « peak picking », pose problème. D'une part, les signaux correspondants peuvent être présents dans les données brutes mais, ne pas être détectés par l'algorithme. D'autre part, une mauvaise intégration de ces signaux peut fausser leur abondance relative, qui diffère alors de celle attendue. Tout cela réduit l'efficacité de la recherche.

Afin de minimiser cette difficulté, les signaux ont été recherchés directement dans les spectres de masse, au plus proche des données brutes, non altérées par le prétraitement.

Cette étape d'extraction considère une fenêtre de temps de rétention définie par le temps de rétention moyen répertorié dans la matrice des données, ajusté à plus ou moins d'une valeur de 3 secondes pour chaque ion mono-isotopique potentiel identifié. Pour chaque fenêtre de temps de rétention considérée pour un ion mono-isotopique donné, les spectres de masse sont extraits avec une fenêtre en masses comprise entre - 0,5 et + 4,5 Da, afin de simplifier les spectres de masse tout en conservant les informations du massif isotopique.

Par la suite, plusieurs étapes présentées ci-dessous sont effectuées afin de détecter les massifs isotopiques au sein des spectres de masse (intra-scan) et de vérifier la cohérence entre les signaux détectés pour les mêmes features entre plusieurs scans extraits (inter-scan) et entre échantillons (inter-échantillons).

Variations intra-scan en masse et dans la détection du massif isotopique

Les masses répertoriées dans la matrice des données étant la synthèse des masses observées pour tous les pics présents dans les spectres de masse, il est donc possible qu'un léger décalage soit observé entre la masse répertoriée et celle présente dans les spectres de masse extraits. Une première étape consiste à sélectionner dans les spectres extraits le signal du pic mono-isotopique qui est le plus proche de la masse répertoriée dans la matrice des données. L'étape suivante consiste à corriger les masses attendues des signaux correspondant aux massifs isotopiques en prenant en compte la différence observée entre la masse théorique et la masse expérimentale du signal attribué au pic mono-isotopique. Ensuite, la recherche d'isotopes est réalisée au sein des spectres de masse en prenant en compte les différences de masses entre l'ion mono-isotopique et ses isotopes correspondants (par exemple : +1,003355 pour le carbone 13) ainsi que leurs abondances relatives. Les fenêtres de tolérance suivantes de ± 0,001 Da et de ± 10 % respectivement pour la masse et pour l'abondance relative ont été appliquées.

À l'issue de la recherche, trois valeurs qualitatives (TRUE, PARTIAL et FALSE) sont attribuées à chaque scan en fonction des résultats obtenus. Si l'ensemble du massif isotopique recherché est détecté, la valeur devient « TRUE ». Si celui-ci est détecté partiellement, la valeur devient « PARTIAL ». Enfin, si aucun signal du massif isotopique n'est détecté ou si un seul isotope est détecté sans les isotopes d'abondance élevée, la valeur devient « FALSE ».

Variations inter-scan en masse et dans la détection du massif isotopique

Les informations relatives à la détection ou non des signaux du massif isotopique pour chaque scan sont comparées dans le but de limiter la détection de faux-positifs liés à la présence d'artefacts dont les masses et les abondances relatives correspondent de manière aléatoire aux massifs isotopiques recherchés.

Dans le cas d'un massif isotopique observé, les différences entre les scans doivent être cohérentes. Par exemple, l'augmentation de l'intensité du signal le plus intense du massif isotopique dans un scan donné doit être suivie de l'augmentation des signaux de plus faible abondance. Il se peut également qu'aucun signal ne soit détecté pour les espèces de faible abondance en raison des limites de sensibilité.

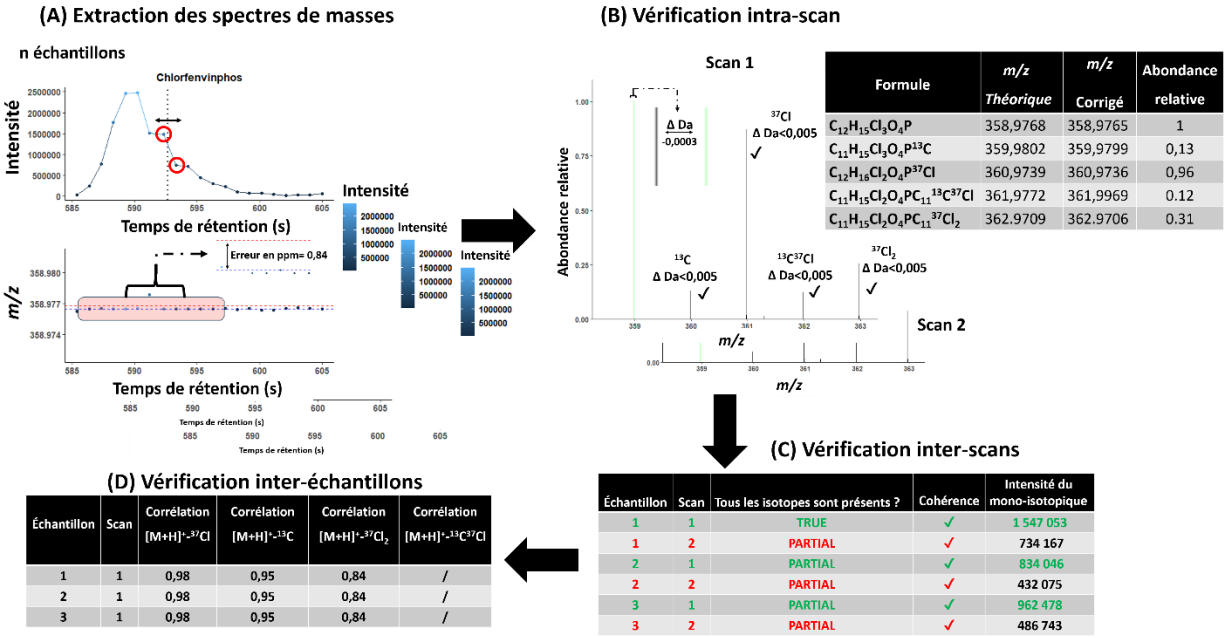
Ainsi, la cohérence entre les scans est évaluée selon la règle suivante : l'intensité des signaux correspondant aux ions mono-isotopiques doit suivre la même tendance que les valeurs

assignées, définies dans la partie intra-scan. Par exemple, l'intensité des valeurs doit respecter l'ordre suivant : TRUE > PARTIAL > FALSE.

Si cette règle est respectée, la cohérence est validée, et le scan le plus informatif est conservé pour chaque feature. Par exemple, pour un feature donné, si tous les cas de figure (TRUE, PARTIAL, FALSE) sont observés dans les n scans extraits, les informations du massif isotopique sont conservées pour le scan noté TRUE. Dans le cas où les scans ne sont pas cohérents, c'est le scan comportant l'intensité la plus élevée pour le pic mono isotopique qui sera reporté.

Variations inter-échantillons en masse et dans la détection du massif isotopique

Dans le cas d'une analyse comportant plus d'un échantillon, comme ici dans le cas d'une analyse de cohorte, les features peuvent être détectés dans plusieurs échantillons avec des intensités variables. Par conséquent, le massif isotopique attendu d'un composé peut être observé seulement pour certains échantillons et pas forcément pour tous. Suivant la même logique que lors de la recherche des massifs isotopiques pour l'inter-scan, la cohérence inter-échantillons est évaluée, en prenant en compte la moyenne des intensités reportées pour chaque valeur assignée lors de l'étape intra-scan (TRUE, FALSE, PARTIAL). De plus, lorsque cela est possible (feature détecté dans au moins trois échantillons, avec un massif isotopique observé entièrement ou partiellement dans un échantillon), un test de corrélation de Pearson est appliqué entre le signal du pic mono-isotopique et chaque signal des isotopes observés. Ce test permet d'identifier les paires de signaux dont les intensités suivent la même tendance (lorsque le signal du pic mono-isotopique augmente, les signaux des massifs isotopiques augmentent également, et inversement). Les paires de signaux présentant une corrélation inférieure à 0,60 sont exclues, car elles sont considérées comme non cohérentes. La figure 2-7 représente graphiquement les différentes étapes de l'extraction des spectres de masse à la validation inter-échantillons



Formule	*m/z* Théorique	*m/z* Corrigé	Abondance relative
$C_{12}H_{15}Cl_3O_4P$	358,9768	358,9765	1
$C_{11}H_{15}Cl_3O_4P^{13}C$	359,9802	359,9799	0,13
$C_{12}H_{16}Cl_2O_4P^{37}Cl$	360,9739	360,9736	0,96
$C_{11}H_{15}Cl_2O_4PC_{11}{}^{13}C^{37}Cl$	361,9772	361,9969	0.12
$C_{11}H_{15}Cl_2O_4PC_{11}{}^{37}Cl_2$	362.9709	362.9706	0.31

Échantillon	Scan	Tous les isotopes sont présents ?	Cohérence	Intensité du mono-isotopique
1	1	TRUE	✓	1 547 053
1	2	PARTIAL	✓	734 167
2	1	PARTIAL	✓	834 046
2	2	PARTIAL	✓	432 075
3	1	PARTIAL	✓	962 478
3	2	PARTIAL	✓	486 743

Échantillon	Scan	Corrélation $[M+H]^+$-^{37}Cl	Corrélation $[M+H]^+$-^{13}C	Corrélation $[M+H]^+$-$^{37}Cl_2$	Corrélation $[M+H]^+$-$^{13}C^{37}Cl$
1	1	0,98	0,95	0,84	/
2	1	0,98	0,95	0,84	/
3	1	0,98	0,95	0,84	/

Figure 2-7 : Représentation graphique des étapes (A) d'extractions des spectres de masses, incluant des vérifications (B) intra-scan, (C) inter-scans et, (D) inter-échantillons

2.3.2.3. Recherche des signaux de xénobiotiques connus surchargés dans des matrices biologiques

La méthode de « suspect screening » décrite ici a été testée sur des jeux de données acquis par LC-HRMS à l'aide d'un Orbitrap Fusion, en modes d'ionisation positif et négatif, sur des matrices biologiques telles que le lait maternel, le sérum et l'urine, dans le cadre du projet PARC (Partenariat européen pour l'évaluation des risques liés aux substances chimiques, collaboration A. Damont, CEA-Saclay). Des xénobiotiques ont été ajoutés dans ces trois matrices pour évaluer leur détectabilité à différentes concentrations (5, 20 et 100 $ng.mL^{-1}$), pour un total de 9 échantillons et 18 jeux de données obtenus avec les deux modes d'ionisation. La recherche de ces composés a été réalisée par un post-doctorant au sein de notre laboratoire, qui a confirmé la présence ou l'absence de ces substances dans les matrices étudiées. Le tableau 2-4 présente les composés détectés, et identifiés dans au moins une matrice, ceci quelle que soit la concentration des composés surchargés.

Tableau 2-4 : Liste des composés surchargés dans les matrices biologiques et la validation manuelle de leur détection ou non en modes positif et négatif, avec le rapport *m/z* des ions détectés

Composé	$[M+H]^+$ *m/z*	$[M-H]^-$ *m/z*	Détection mode positif	Détection mode négatif
Paracetamol-d_4	156,0957	154,0812	Oui	/
Phenanthrene- d_{10}	-	-	/	/
^{13}C-4',4-DDE (dichlorodiphenyldichloroethylene)	-	-	/	/
alpha-Hexabromocyclododecane-d_{18}	-	-	/	/
^{13}C3-Desethydesisopropyl Atrazine	149,0329	147,0183	/	/
Caffeine *d9*	204,1441	202,1296	Oui	/
2-hydroxy-4-methoxybenzophenone-*d5*	234,1173	232,1028	/	/
Chlorpyrifos-*d10*	359,9963	-	/	/
Aminomethylphosphonic acid	112,0158	110,0013	/	/
1,3-Dichloro-2-propanol	128,9868	126,9723	/	/
4-Nitrophenol	140,0342	138,0197	/	Oui
2,6-Diethylaniline	150,1277	148,1132	/	/
Acetaminophen	152,0706	150,0561	Oui	/
Chlormequat chloride	122,0731	-	Oui	/
O,O-Dimethyldithiophosphate	158,9698	156,9552	/	Oui

Composé	$[M+H]^+$ *m/z*	$[M-H]^-$ *m/z*	Détection mode positif	Détection mode négatif
2,4-Dichlorophenol	162,9712	160,9566	/	/
Glyphosate	170,0213	168,0067	/	/
2-(Diethylamino)-6-methyl-1H-pyrimidin-4-one	182,1288	180,1142	Oui	/
Benzophenone	183,0804	-	/	/
DEET	192,1383	-	Oui	/
Butylparaben	195,1016	193,0870	/	Oui
3,5,6-Trichloro-2-pyridinol	197,9275	195,9129	/	Oui
2-Hydroxyatrazine	198,1349	196,1204	Oui	Oui
Ibuprofen	207,1380	205,1234	/	/
Omethoate	214,0297	212,0152	Oui	/
2,4-Dihydroxybenzophenone	215,0703	213,0557	/	Oui
3-Phenoxybenzoic acid	215,0703	213,0557	/	Oui
Benz(a)anthracene	-	-	/	/
Bisphenol A	229,1223	227,1078	/	/
2,3,4,5-Tetrachlorophenol	230,8933	228,8787	/	Oui
Bentazone	241,0641	239,0496	/	Oui
2,4-Dibromophenol	250,8702	248,8556	/	Oui
Benzo[a]pyrene	-	-	/	/
Malathion dicarboxylic Acid	274,9807	272,9662	/	Oui

Composé	$[M+H]^+$ *m/z*	$[M-H]^-$ *m/z*	Détection mode positif	Détection mode négatif
Indeno[1,2,3-cd]pyrene	-	-	/	/
Mono-2-ethylhexyl phthalate - MEHP	279,1591	277,1445	/	Oui
1,2,3,4,5,6-Hexachlorocyclohexane	-	-	/	/
Triclosan	288,9584	286,9439	/	Oui
Dibutyl decanedioate	315,2530	-	/	/
Ipconazole	334,1681	332,1535	Oui	/
Meloxicam	352,0420	350,0275	Oui	Oui
2,2',4,4',5,5'-Hexachlorobiphenyl	-	-	/	/
Chlorfenvinphos	358,9768	356,9623	Oui	/
Bis(2-ethylhexyl) terephthalate	391,2843	-	/	/
Di(2-ethylhexyl) phthalate	391,2843	-	/	/
Perfluorohexanesulfonic acid	400,9512	398,9366	/	Oui
2,3,4,5-Tetrabromophenol	406,6912	404,6766	/	Oui
Perfluorooctanoic acid	414,9810	412,9664	/	Oui
Tris(1,3-dichloro-2-propyl) phosphate	428,8912	-	/	/
Fipronil	436,9460	434,9314	/	Oui
Fipronil sulfone	452,9409	450,9263	/	Oui
Perfluorononanoic acid	464,9778	462,9632	/	Oui
Perfluorooctanesulfonic acid	500,9448	498,9302	/	Oui
Tris(2-ethylhexyl) trimellitate	547,3993	-	/	/

Composé	$[M+H]^+$ *m/z*	$[M-H]^-$ *m/z*	Détection mode positif	Détection mode négatif
Bisphenol S Bis-b-d-Glucuronide	603,1014	601,0869	/	Oui
1,2,5,6,9,10-Hexabromocyclododecane	-	-	/	/
Tris(2,3-dibromopropyl) phosphate	692,5881	-	/	/
Bis(2-ethylhexyl) tetrabromophthalate	702,9263	-	/	/

Parallèlement, une recherche de ces mêmes composés a été effectuée sur ces données selon la stratégie utilisée ici. Cela a permis d'évaluer l'efficacité de notre approche pour retrouver ces composés tout en minimisant les faux-négatifs (composés non identifiés par la méthode utilisée alors qu'ils étaient présents).

Au total, 22 composés ont été détectés et validés manuellement au moins une fois sur les données obtenues en mode négatif grâce à l'utilisation du logiciel constructeur Xcalibur, en se basant sur les temps de rétention de composés de référence et la présence du pic mono-isotopique. La recherche de ces mêmes composés par l'approche de « suspect screening » a permis de confirmer automatiquement leur présence, grâce à la détection de leur massif isotopique concordant avec celui attendu dans au moins un des échantillons.

Le prétraitement de ces données a été effectué avec l'utilisation du package XCMS sous le logiciel R. Les paramètres de prétraitement des données sont présentés dans la partie méthodologie (1.3.3 Prétraitement des données).

Parmi les pics validés manuellement, c'est-à-dire 155 en mode négatif et 62 en mode positif, 139 et 61 ont pu être détectés respectivement en modes négatif et positif par XCMS et reportés dans la matrice des données. Une analyse rapide des intensités rapportées pour les 16 pics manquants en mode négatif indique que leurs intensités sont sensiblement plus faibles que les autres. Cela peut expliquer leur non-détection, ce qui n'est pas le cas du seul pic manquant en mode positif, provenant de l'échantillon surchargé à la plus haute concentration. Le fonctionnement de notre stratégie est flexible et peut s'adapter à une intégration approximative des signaux grâce à l'extraction des spectres de masse. Cependant,

leur non-détection rend leur validation impossible. Ainsi, l'efficacité de la méthode présentée dépend en partie de la méthode d'extraction des données et donc des paramètres du prétraitement par XCMS.

La méthode d'extraction des spectres de masse réalisée avec une tolérance de ± 3 secondes pour le temps de rétention médian de chaque pic extrait par XCMS, a permis de valider la présence de massifs isotopiques (partiellement ou en totalité) pour 114 et 57 pics respectivement en modes négatif et positif.

Ici, la vérification inter-échantillons nous a permis de valider les 139 pics et 61 pics extraits respectivement par XCMS en modes positif et négatif, en tenant compte des 114 pics et 57 pics présentant un massif isotopique concordant. En effet, le massif isotopique peut être vérifié pour les espèces ioniques suffisamment abondantes pour permettre son observation, ce qui est généralement le cas lorsque les composés sont présents à des concentrations plus élevées. En revanche, pour les espèces ioniques de faible abondance, leur massif isotopique risque de ne pas être détecté.

Ainsi, les pics validés pour les composés présents à des concentrations élevées dans les matrices analysées permettent de confirmer leur présence par la détection des pics de plus faible abondance à la même valeur *m/z* et au même temps de rétention dans les échantillons surchargés à plus faible concentration. Le tableau 2-5 répertorie les différentes étapes de la stratégie utilisée ainsi que les pics validés à chaque étape.

Tableau 2-5 : Synthèse des résultats obtenus sur les matrices biologiques surchargées, montrant le nombre d'ions validés manuellement et automatiquement à chaque étape de avec l'approche utilisée

	mode négatif	**mode positif**
Nombre de composés ajoutés dans chaque matrice	22	8
Nombre d'échantillons analysés	9	9
Nombre de pics validés manuellement	155	62
Nombre de pics mono-isotopiques extraits par XCMS	139	61
Nombre de pics validés sur la présence d'un massif isotopique concordant avec la stratégie utilisée (extraction des spectres de masses + vérification inter-scan)	114	57
Nombre de pics validés sur la présence des massifs isotopiques et validation inter-échantillons (extraction des spectres de masses + vérification inter-scan et inter-échantillons)	139	61
Nombre de composés validés avec la stratégie développée	22	8

2.4. Recherche de composés suspectés dans les données métabolomiques de la cohorte EDEN

2.4.1. Stratégies développées

Sur la base des étapes précédemment décrites, deux stratégies ont été élaborées pour s'adapter au nombre de composés présents dans la liste de molécules suspectées ainsi qu'au nombre d'échantillons à analyser

2.4.1.1. Stratégie statique

La stratégie statique (Figure 2-1) repose sur la prédiction des métabolites, qui s'effectue par le biais d'un certain nombre d'itérations (correspondant au nombre de biotransformations successives considérées), suivie de leur recherche dans les jeux de données (basée sur la détection des pics mono-isotopiques), de l'extraction des spectres de masse, et de la validation du massif isotopique dans ces spectres. Ces étapes sont réalisées en cascade, ce qui permet de détecter des signaux de métabolites indépendamment de la détection des formes parentes. Cette approche est particulièrement intéressante, car elle permet de rechercher un grand nombre de marqueurs d'exposition.

Cependant, il est impossible de garantir qu'un signal détecté à une valeur de *m/z* proche de celle d'un métabolite prédit *in silico* correspond bien à un marqueur d'exposition lorsque les jeux de données en spectrométrie de masse (MS) ne contiennent aucune information sur l'exposition des individus et en absence des données produites en MS/MS. Il est également difficile d'attester la présence d'un contaminant si seuls des métabolites potentiels sont détectés, sans la présence de l'ion correspondant à la forme parent. Dans ce contexte, la présence de plusieurs signaux, dont celui du composé parent, de préférence dans le même échantillon, peut fournir des indices supplémentaires sur sa présence.

De plus, la grande quantité de métabolites générés par la prédiction *in silico* conduit à la recherche de nombreux ions, augmentant inévitablement le nombre de signaux détectés, ce qui rend le traitement manuel des données plus complexe, notamment lors de la recherche d'un grand nombre de composés. Cette stratégie est intéressante dans le cas où des expérimentations supplémentaires en MS/MS sont possibles ou lors d'un premier screening permettant d'établir des pistes sur la présence de composés.

2.4.1.2. Stratégie dynamique

Une autre stratégie, dite dynamique (Figure 2-8), a été développée pour assurer la recherche et la détection de molécules suspectées parmi une liste plus étendue. Bien que les étapes de fouille de données soient les mêmes que celles décrites précédemment, elles sont agencées différemment. Dans cette approche, la génération des métabolites et la recherche des signaux ne constituent plus des étapes distinctes, mais deviennent des processus cycliques.

Pour une liste définie de composés suspectés, une première recherche de signaux (pics mono-isotopiques et validation du massif isotopique) est effectuée, ce qui génère deux listes : une liste des composés de la liste suspectée pour lesquels des signaux candidats ont été détectés, et une liste d'exclusion regroupant tous les composés recherchés qu'ils aient été détectés ou non. Cette dernière liste permet d'éviter la recherche du même ion à plusieurs reprises, ce qui allège le processus de recherche des signaux.

La prédiction des métabolites *in silico* n'est appliquée qu'à la liste des composés pour lesquels des signaux candidats ont été détectés, afin de repérer d'autres signaux pouvant correspondre à la présence de métabolites en plus de la molécule parent. Par exemple, une première recherche est effectuée pour les molécules parentes, puis les candidats détectés sont soumis à la prochaine itération, qui inclut la prédiction des métabolites et leur recherche. Ce processus se répète ainsi de suite, en fonction du nombre d'itérations sélectionnées (correspondant au nombre de biotransformations successives).

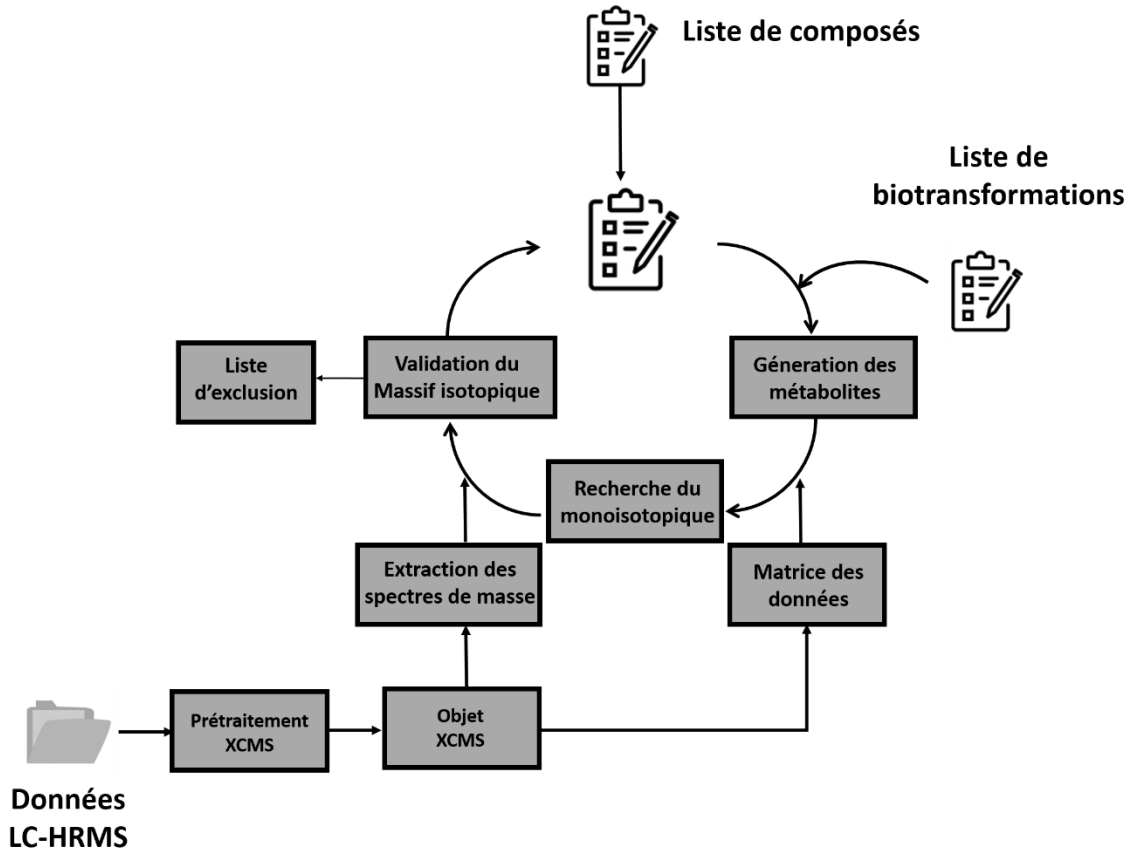


Figure 2-8 : Représentation graphique de la stratégie dynamique

2.4.2. Création des listes de molécules suspectées

Une première liste de contaminants (Liste 1) a été constituée. Elle comprend 125 contaminants issus de la liste de l'Anses[44,137], regroupant les contaminants présents dans des aliments similaires à ceux consommés par les mères des cohortes EDEN et Elfe, ainsi que 60 composés exogènes tirés de la base de données HMDB et 63 composés identifiés dans la littérature comme présentant un risque potentiel pendant la grossesse, pour un total de 248 composés.

La stratégie statique a été adoptée pour identifier la présence des molécules de la liste 1 dans nos jeux de données. Celle-ci permet de rechercher des métabolites prédits par l'approche présentée ici, que la forme parente soit présente ou non dans l'ensemble des données.

Une deuxième liste (Liste 2) a été élaborée pour tester la stratégie dynamique. Elle est extraite de la banque de données OpenFoodTox[138] établie par l'Autorité Européenne de Sécurité des Aliments (EFSA), qui recense plus de 6 300 substances chimiques présentes dans les

aliments et l'environnement. Une sélection a été effectuée en excluant les substances sous forme de sels ou hydratées, ce qui conduit à un total de 4 719 molécules à rechercher.

2.4.3. Résultats de l'analyse

La présentation des résultats de l'analyse se déroule comme suit. Dans un premier temps, les paramètres utilisés pour la génération des marqueurs d'exposition et leur recherche sont présentés.

Dans un second temps, les temps d'application des deux approches sont mesurés et comparés pour chacune des matrices auxquelles ces stratégies ont été appliquées.

Enfin, une synthèse des molécules parentes et des marqueurs d'exposition détectés à partir de la liste 1 dans nos échantillons de méconium analysés en ESI+ est présentée. Les résultats des analyses réalisées sur les autres matrices sont, quant à eux, organisés sous forme de tableaux.

Génération des marqueurs d'expositions

Pour la prédiction des marqueurs d'exposition potentiels à partir les listes 1 et 2, le nombre de métabolisations considérées est limitée à 1, ce qui signifie qu'au maximum, chaque molécule des deux listes initiales ne peut subir qu'une seule biotransformation possible. La liste des biotransformations sélectionnées pour la prédiction des marqueurs d'exposition est présentée dans le tableau 2-2

5 962 structures différentes ont été générées à partir de la liste 1 pour un total de 2 483 masses uniques ou même formule chimique (2 235 masses de métabolites et 248 masses de composés parents). Comme évoqué précédemment dans la partie « Prédiction des métabolites », l'approche utilisant des biotransformations sur des motifs moléculaires non spécifiques génère beaucoup d'isomères.

Recherche des signaux provenant des marqueurs d'expositions

La recherche des signaux des ions mono-isotopiques a été effectuée en prenant en compte une erreur en masse de (±) 5 ppm. Les spectres de masse ont été extraits pour chaque feature dont le temps de rétention médian et la valeur *m/z* médiane sont indiqués dans la matrice des données, avec une fenêtre de (±) 1 seconde en temps de rétention et celle en masse comprise entre -0,5 Da et + 4,5 Da.

L’abondance relative maximale des isotopes à rechercher pour chaque espèce dans les spectres de masse extraits est fixée à 0,4 %, et la valeur de corrélation minimale entre les intensités des signaux mono-isotopiques et celles des isotopiques détectés est de 0,60.

Evaluation de la durée du traitement des données selon les deux stratégies sur les données de méconium et lait maternel en modes ESI positif et négatif

Le temps nécessaire pour analyser chaque jeu de données issu de chacune des matrices en mode positif et négatif est présenté dans le tableau 2-6.

Tableau 2-6 : Temps d'analyse des données issues des échantillons de méconium et de lait maternel en mode positif et négatif

Matrice	**Méconium**		**Lait maternel**	
Mode	Positif	négatif	positif	négatif
Temps Liste 1 (statique)	105 minutes	25 minutes	57 minutes	12 minutes
Temps Liste 2 (dynamique)	600 minutes	222 minutes	453 minutes	102 minutes

Il convient de souligner que le temps indiqué ici correspond uniquement à l’extraction des spectres de masse et aux validations intra et inter-scans. Les autres étapes de la stratégie requièrent très peu de temps (non pris en compte dans le tableau 2-6) : moins d’une 1 minute pour la recherche des mono-isotopiques dans la matrice des données et la génération des marqueurs d’exposition, et moins de trois minutes pour la validation inter-échantillons du méconium en mode positif. Les temps maximum et minimum identifiés, rapportés par les stratégies statique et dynamique utilisées, ont été respectivement de 105 minutes et 600 minutes pour l’analyse des méconiums en mode positif ainsi que 12 minutes et 102 minutes pour le lait maternel en mode négatif

Cette différence dans le temps de traitement peut s'expliquer par la densité des signaux détectés par XCMS dans le jeu de données à analyser. En effet, en moyenne, plus de 17 000 signaux ont été détectés par échantillon de méconium en mode positif, contre 4 906 en mode négatif. Cela a pour conséquence d'augmenter drastiquement le temps d’application car le nombre de signaux des pics mono-isotopiques détectés augmente, ce qui entraîne une hausse

du nombre de spectres de masse à extraire et à vérifier. Dans tous les cas, les temps d'application relevés pour l'analyse des 308 échantillons nous semblent acceptables.

Détection des marqueurs d'exposition.

Lors de l'analyse des échantillons de méconium en mode positif, des candidats de rapport *m/z* correspondant aux espèces protonées $[M+H]^+$ de 46 composés avec validation de leurs massifs isotopiques ont été détectés. Le nombre de candidats trouvés pour chacun de ces composés lors de l'étape de détection des ions mono-isotopiques sur la matrice des données, a été comparé à celui obtenu après extraction des spectres de masse et validation des massifs isotopiques. Cette comparaison est présentée dans la figure 2-9. Une liste de ces composés et leurs structures est présentée dans l'annexe 3.

On peut noter une diminution du nombre de candidats potentiels pour les composés ayant plusieurs candidats possibles, fondée sur la détection des massifs isotopiques attendus dans les spectres de masse. Il convient toutefois de souligner qu'il ne s'agit pas, à proprement parler, d'une réduction des faux-positifs. Dans notre cas, l'impact sur les faux-positifs peut néanmoins être évalué pour trois composés annotés, par comparaison avec la banque de données interne au laboratoire : la nicotine, l'acétaminophène et la caféine.

Après extraction et vérification des spectres de masse, il ne reste qu'un seul candidat possible pour la nicotine et la caféine, détecté au même temps de rétention que les molécules de référence présentes dans la banque de données interne du laboratoire, soient respectivement 55 et 283 secondes (voir table S2 de la partie « Supplementary Information » du chapitre 3). Pour les candidats pouvant correspondre à l'acétaminophène, bien que la diminution du nombre de candidats soit plus faible, parmi les 30 candidats possibles, deux d'entre eux ont été observés au temps de rétention attendu selon les données disponibles dans la banque de données interne au laboratoire : 160 secondes pour la forme parent et 95 secondes pour celle issue de la fragmentation du métabolite glucuro-conjugué dans la source d'ionisation.

Bien sûr, il est possible que parmi les signaux candidats non retenus se trouvent ceux des composés attendus. Il pourrait toutefois s'agir d'artefacts, en l'absence d'une intensité suffisante pour accéder au profil isotopique.

La figure 2-10 représente la répartition des intensités log-transformées des ions mono-isotopiques pour les signaux candidats de la caféine, avec en rouge les signaux du candidat

validé pour lequel le profil isotopique a été vérifié (n = 304), et en bleu ceux des candidats non validés (n = 272).

Une intensité médiane plus importante est effectivement observée dans le cas du groupe validé par rapport au groupe non validé. Toutefois, envisager des analyses complémentaires, telles que de la spectrométrie de masse en tandem (MS/MS) sur ce type de signaux, peut être problématique en raison d'une intensité trop faible pour obtenir des spectres de fragmentation exploitables. De plus, seul l'accès au composé standard permettrait d'identifier de façon non ambiguë la présence des composés d'intérêt parmi les candidats possibles sur la base du temps de rétention.

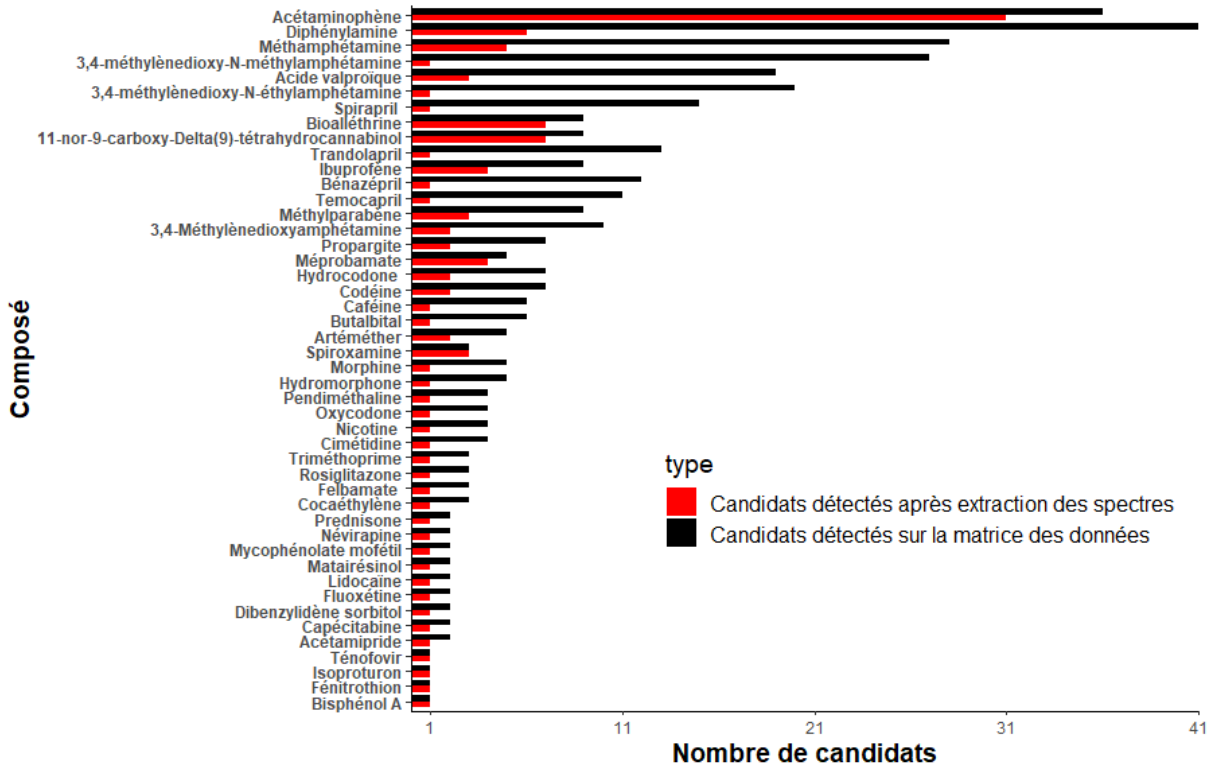


Figure 2-9 : Comparaison entre le nombre de candidats détectés sur la base de l'ion mono-isotopique dans la matrice des données (noir) et le nombre de candidats obtenu après validation du massif isotopique dans les spectres de masse extraits (rouge)

En comparaison avec l'étude préliminaire, qui ne prenait en compte que la liste de 125 composés établie à partir de l'étude de l'ANSES[44,137] (Figure 2-3), seuls quelques signaux, pouvant correspondre à la propargite, à la spiroxamine, au matairisinol, au sorbitol et au fenitrothion sous leurs formes parentales, passent les filtres de qualité appliqués ici.

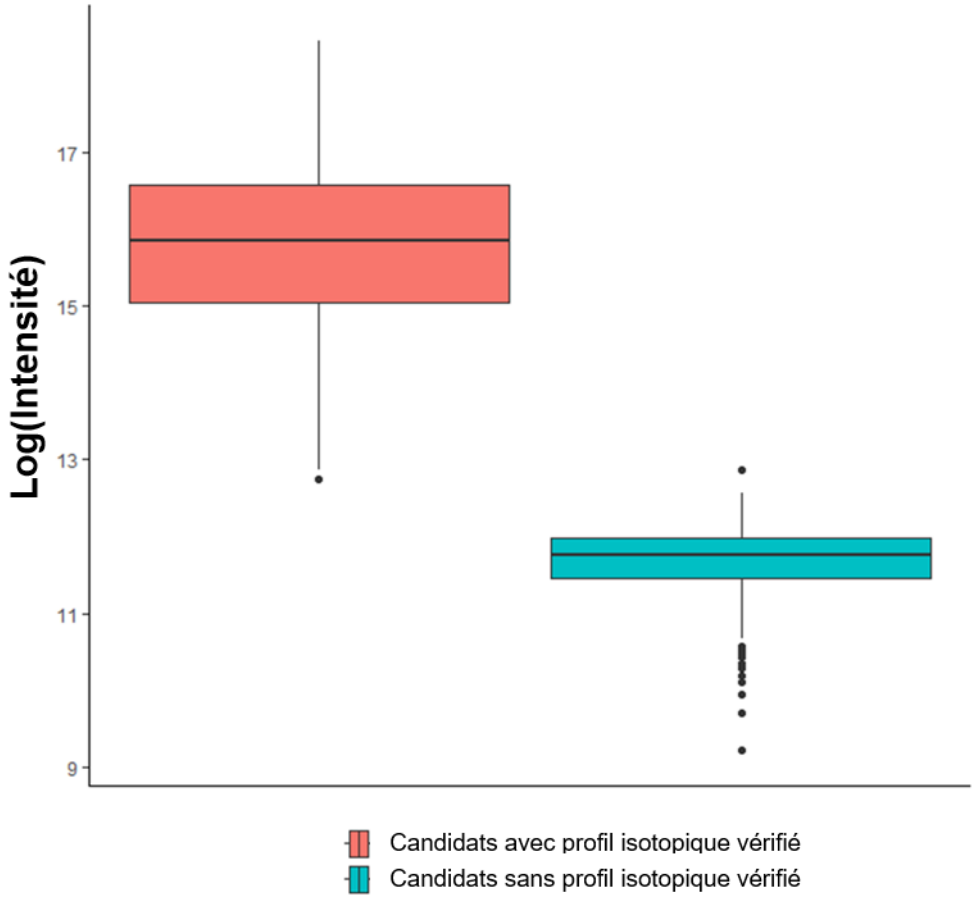


Figure 2-10 : Répartition des intensités en log des candidats dont le profil isotopique a été validé (rouge) et non validé (bleu) pour la caféine

Pour ces 46 candidats parents détectés, plusieurs signaux pouvant correspondre à des métabolites générés par l'approche utilisée ici ont été observés. Le nombre de ces métabolites candidats détectés pour chaque composé est présenté dans la figure 2-11. Une analyse plus poussée a été réalisée sur une sous-sélection de composés, tels que l'acétaminophène, la nicotine, le spiroxamine et la propargite, dont respectivement 9, 6, 5 et 7 métabolites possibles ont été détectés par l'approche utilisée. Les formules chimiques associées à ces signaux par notre approche sont présentées tableau 2-7.

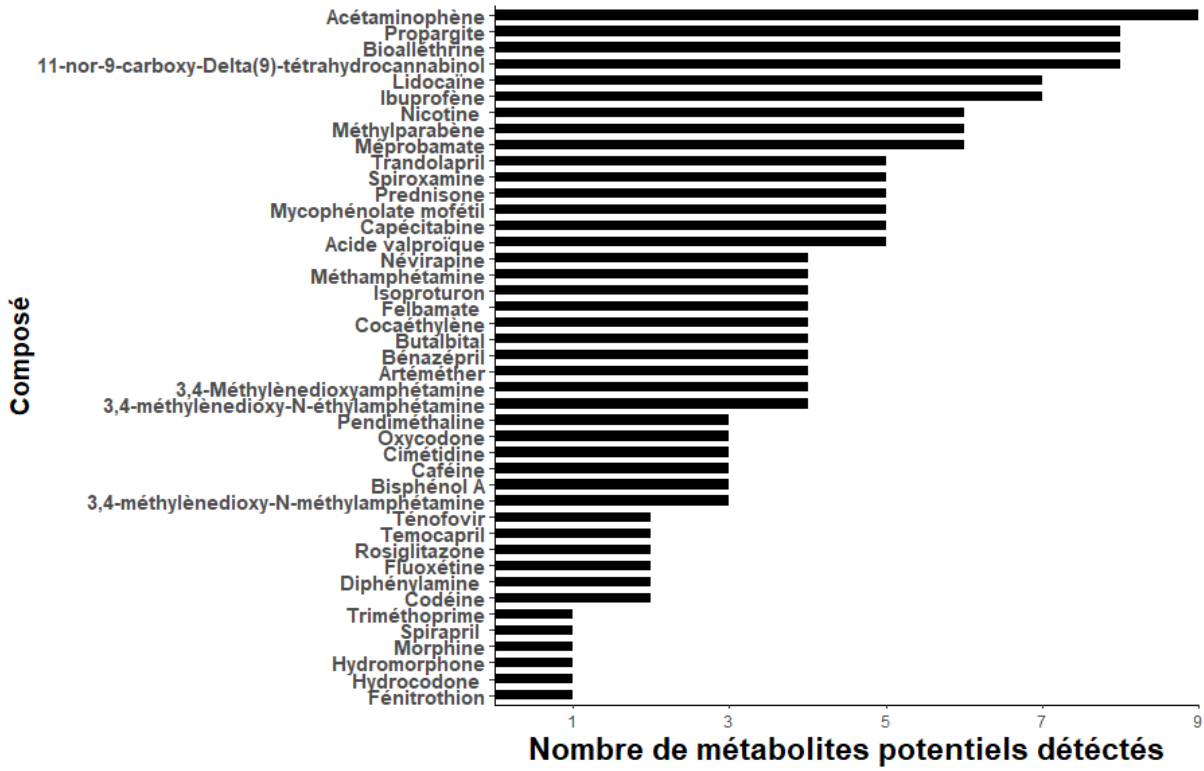


Figure 2-11 : Nombre de métabolites potentiels détectés pour les composés dont un ion pouvant correspondre au composé parent a été observé

Tableau 2-7 : Liste des métabolites prédits et détectés dans nos échantillons pour l'Acétaminophène, la Nicotine, la Spiroxamine et le Propargite, avec leur formule chimique, ainsi que la réaction qui a généré leur formation et l'écart (ΔMD) en défaut de masse entre le parent et chaque métabolite formé.

Composé parent	Métabolites	Réaction	ΔMD (mDa)
Acétaminophène ($C_8H_9NO_2$)	$C_{14}H_{17}NO_8$	Glucuroconjugaison	+32,1
	$C_8H_7NO_2$	Réduction (-H_2)	-0.0157
	$C_6H_6O_2$	Hydrolyse	-26, 7
	C_6H_7NO	Hydrolyse	-7,2
	$C_8H_{11}NO_2$	Hydrogénation (+H_2)	+0,0157
	$C_8H_{11}NO_4$	Double oxydation (+2OH)	+0.0054
	$C_8H_9NO_3$	Oxydation (+O)	-0.0051
	$C_8H_9NO_5S$	Sulfoconjugaison	-43,2
	$C_9H_{11}NO_2$	Methylation (+CH_2)	+0.0157

Composé parent	Métabolites	Réaction	ΔMD (mDa)
Nicotine ($C_{10}H_{14}N_2$)	$C_{10}H_{12}N_2$	Réduction ($-H_2$)	-0.0157
	$C_{16}H_{22}N_2O_6$	N-glucuroconjugaison	+32,1
	$C_{10}H_{12}N_2$	Réduction ($-H_2$)	-0.0157
	$C_{10}H_{14}N_2O$	Oxydation (+O)	-0.0051
	$C_{10}H_{16}N_2$	Hydrogénation ($+H_2$)	+0.0157
	$C_{10}H_{16}N_2O_2$	Double oxydation (+2OH)	+0.0054
Spiroxamine ($C_{18}H_{35}NO_2$)	$C_{17}H_{31}NO_4$	Oxydation ($+O_2$,$-CH_4$)	-0.0415
	$C_{18}H_{33}NO_2$	Réduction ($-H_2$)	-0.0157
	$C_{18}H_{35}NO_3$	Oxydation (+O)	-0.0051
	$C_{18}H_{33}NO_3$	Oxydation (+O,$-H_2$)	-0.0208
	$C_{24}H_{43}NO_8$	N-glucuroconjugaison	+32,1
Propargite ($C_{19}H_{26}O_4S$)	$C_{10}H_{14}O$	Hydrolyse	-50, 8
	$C_{16}H_{24}O_2$	Hydrolyse	22, 4
	$C_{19}H_{22}O_4S$	Réduction ($-H_4$)	-0.0313
	$C_{19}H_{24}O_5S$	Oxydation (+O,$-H_2$)	-0.0208
	$C_{19}H_{26}O_5S$	Oxydation (+O)	-0.0051
	$C_{19}H_{28}O_4S$	Hydrogénation ($+H_2$)	+0.0157
	$C_3H_4O_2S$	Hydrolyse	-162
	$C_{19}H_{28}O_6S$	Double oxydation (+2OH)	+0.0054

Parmi les métabolites prédits et détectés, une majorité provient de biotransformations couramment observées, telles que la glucuroconjugaison ou la N-glucuroconjugaison, la sulfoconjugaison, l'oxydation, la méthylation ou l'hydrolyse, qui peuvent être cohérentes avec la structure des molécules proposées. Cependant, des erreurs peuvent apparaître. Dans les cas cités ci-dessus, certaines réactions, notamment d'hydrogénations, ont été appliquées à des cycles aromatiques. Bien que la réaction encodée sous forme de SMARTS nécessite la

présence d'une double liaison entre deux carbones aliphatiques (tableau 2-2), cette réaction ne peut *a priori* pas se produire en raison de la grande stabilité des cycles aromatiques. De plus, certaines hydrolyses semblent incohérentes. Dans le cas de l'acétaminophène, une hydrolyse a été proposée, impliquant le départ d'un groupement amide. Après vérification, la réaction encodée en SMARTS manquait de précision quant au motif moléculaire recherché. Il faut souligner que les biotransformations encodées sous forme de SMARTS, lorsqu'elles sont trop simplifiées, peuvent mener à la prédiction et, par conséquent, à la détection de faux-positifs, car elles ne prennent pas en compte la complexité des réactions. Néanmoins, il est possible d'intégrer cette complexité dans les encodages afin d'améliorer la précision des prédictions.

L'examen des défauts de masse des métabolites putatifs dans un profil de défauts de masse démontre que la majorité des métabolites détectés se trouvent dans une fenêtre restreinte de défauts de masse, et en particulier autour de ceux des molécules parentes. Cette particularité a déjà été exploitée par les approches de filtre de défauts de masse permettant de détecter les métabolites de phase 1 des médicaments au sein de matrices biologiques.[122]

En effet, la majorité des biotransformations de phase 1 conduisent à des métabolites ayant des masses et des défauts de masse relativement proches de ceux des composés parents. Les approches par filtres de défauts de masse sélectionnent les signaux en appliquant une fenêtre autour de la valeur *m/z* correspondant aux précurseurs (usuellement +50 mDa en défaut de masse et 50 Da pour la masse nominale).

Bien que la détection des parents et de leurs métabolites prédits ne puisse être confirmée en l'absence d'expériences supplémentaires, il est intéressant de souligner que la majorité des métabolites prédits par l'approche utilisée ont des défauts de masse localisés dans une zone entour entourant le défaut de masse du composé parent. Par exemple, les défauts de masse des marqueurs potentiels d'exposition à l'acétaminophène, à la nicotine, à la spiroxamine et à la propargite sont observés dans les carrés rouges de la figure 2-12, en présence de l'ensemble des signaux détectés dans les 308 échantillons de méconium en mode positif. Les métabolites situés à l'extérieur de ces zones proviennent de métabolites prédits issues de biotransformations qui modifient de manière conséquente les structures chimiques, comme des réactions de phase 2 (glucuroconjugaison, sulfoconjugaison, etc.) ou des clivages issus

d’hydrolyses, pour lesquels des filtres de défauts de masse supplémentaires sont usuellement appliqués afin de les détecter.

L’approche par filtre de défauts de masse, dans notre cas, aurait conduit à l’extraction d’un grand nombre de signaux pour les quatre composés considérés. Il convient de souligner que l’utilisation de filtres de défauts de masse sur des signaux obtenus à partir de plusieurs échantillons différents peut drastiquement diminuer son efficacité en raison de la présence d'artefacts. Néanmoins, dans le contexte d’exposome chimique et en raison du profil génétique différent des individus, la présence des molécules sous forme parentale et des métabolites peut varier (voir 1.2.2. Réactions de biotransformation des xénobiotiques du chapitre 1). Il est donc nécessaire de prendre en compte l’ensemble des échantillons dans le cadre de notre étude. Il est également possible que, des métabolites de ces composés, qui n’ont pas été détectés par notre approche, se trouvent potentiellement parmi les signaux au sein de la zone définie (carré rouge). Néanmoins, l’extraction d’autant de signaux sans informations relatives à la structure des composés parents aurait requis un travail considérable de curation des données.

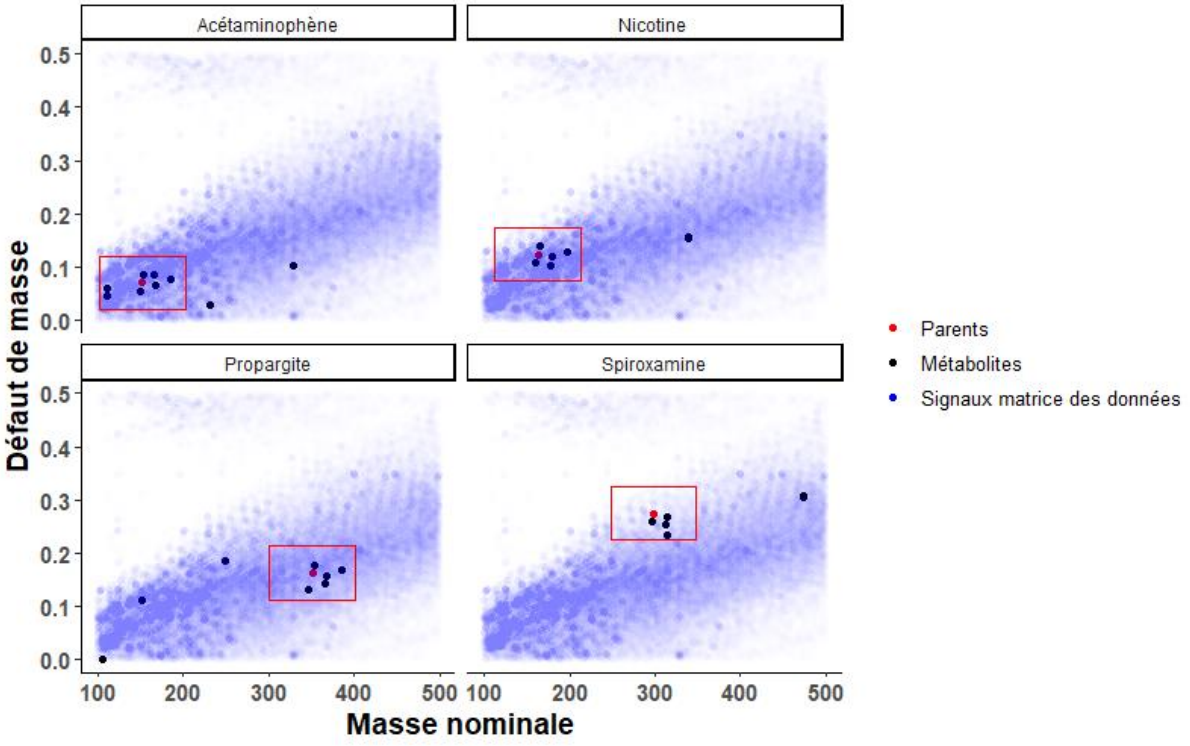


Figure 2-12 : Profils de défaut de masse obtenus à partir des données de l'ensemble 308 échantillons de méconium en mode positif, avec la présence de marqueurs d'exposition potentiels détectés pour l'acétaminophène, la nicotine, la spiroxamine et la propargite. Les points rouges correspondent aux signaux des composés parents et les points noirs aux signaux des métabolites supposés. Les points bleus correspondent aux signaux des jeux de données (n = 308) pour la fenêtre de défauts de masse de 0 à 0,5 Da et la masse nominale de *m/z* 100 à 500. Le carré rouge correspond à la fenêtre usuellement utilisée pour extraire les métabolites de phase 1 par filtres de défauts de masse (±50 mDa pour le défaut de masse et ± 50 Da pour la masse nominale appliqués sur le défaut de masse et la masse du parent).

Synthèse des résultats des analyses des jeux de données acquis en modes positif et négatif sur le lait maternel et le méconium utilisant les listes 1 et 2 de molécules

La synthèse des résultats obtenus pour la recherche des molécules des listes 1 et 2 au sein des échantillons de méconium et de lait maternel est présentée dans les tableaux 2-8 et 2-9.

Le tableau 2-8 montre que la majorité des composés parents et des métabolites sont détectés dans le jeu de données issu des analyses des échantillons de méconium en mode positif. Bien que la détection des marqueurs d'exposition soit moindre pour les autres jeux de données, la plupart d'entre eux n'ont pas été observés dans le premier jeu de données.

Parmi les composés potentiellement détectés, on observe que la plupart sont des substances xénobiotiques que les individus ingèrent généralement volontairement, telles que des médicaments (par exemple, l'acétaminophène, la codéine, le bénazépril, etc.) ou des

substances à usage récréatif (la nicotine, la 3,4-méthylènedioxy-N-méthylamphétamine, la caféine, etc.). Seuls quelques candidats, généralement présents à l'état de traces, comme certains pesticides (propargite, fénitrothion, diazinon) ont été détectés. Bien que l'identité de ces signaux reste à confirmer, il est probable que les matrices analysées par profilage métabolomique puissent révéler la présence de xénobiotiques. Cependant, la détection de ces xénobiotiques ou de leurs métabolites présents à des niveaux très faibles de concentration peut être affectée par la sensibilité de l'approche analytique utilisée.

Tableau 2-8 : Composés parents (P) et nombre de métabolites (M) détectés dans les données LC-HRMS (positif/négatif) du méconium et du lait maternel de la cohorte EDEN pour la liste 1

	Méconium		**Lait maternel**	
Composé	**Mode positif**	**Mode négatif**	**Mode positif**	**Mode négatif**
Acétaminophène	P+9M	P+7M	P+7M	P+4M
Acide tartrique	/	P+4M	/	/
Capécitabine	/	P+2M	/	/
Daidzéine	/	P	/	/
Diazinon	/	P	/	P
Méprobamate	/	P+5M	/	/
Métalaxyl	/	/	P+5M	/
Azoxystrobine	/	/	P+3M	/
Acide mycophénolique	/	/	P+3M	/
Norpropoxyphène	/	/	P+3M	/
Propoxyphène	/	/	P+3M	/
Imipramine	/	/	P+2M	/
Méprobamate	/	/	P+2M	/
Acide perfluorodécanoïque	/	/	P+1M	/
3,4-méthylènedioxy-N-éthylamphétamine	P+5M	/	P+1M	/
3,4-méthylènedioxy-N-méthylamphétamine	P+4M	/	P	/
Célécoxib	/	/	P	/
Citalopram	/	/	P	/
Coumestrol	/	/	P	/
La fosphénytoïne	/	/	P	/
Patuline	/	/	/	P
11-nor-9-carboxy-Delta(9)-tétrahydrocannabinol	P+9M	/	/	/

	Méconium		Lait maternel	
Composé	**Mode positif**	**Mode négatif**	**Mode positif**	**Mode négatif**
Propargite	P+8M	/	/	/
Lidocaïne	P+8M	/	P+5M	/
Bioalléthrine	P+8M	/	/	/
Ibuprofène	P+7M	/	/	/
Trandolapril	P+6M	/	/	/
Nicotine	P+6M	/	P+3M	/
Méthylparabène	P+6M	P+4M	P+7M	P+2M
Méprobamate	P+6M	/	/	/
Acide valproïque	P+5M	P+2M	P+2M	P+1M
Spiroxamine	P+5M	/	P+3M	/
Prednisone	P+5M	/	P+2M	/
Oxycodone	P+5M	/	/	/
Mycophénolate mofétil	P+5M	/	/	/
Cocaéthylène	P+5M	/	/	/
Capécitabine	P+5M	/	/	/
Butalbital	P+5M	P+2M	P+2M	/
Bénazépril	P+5M	/	/	/
Névirapine	P+4M	/	/	/
3,4-Méthylènedioxyamphétamine	P+4M	/	/	/
Méthamphétamine	P+4M	/	/	/
Isoproturon	P+4M	/	P+2M	/
Felbamate	P+4M	/	P+1M	/
Codéine	P+4M	/	/	/
Artéméther	P+4M	/	/	/

	Méconium		Lait maternel	
Composé	**Mode positif**	**Mode négatif**	**Mode positif**	**Mode négatif**
Temocapril	P+3M	/	P+4M	/
Pendiméthaline	P+3M	/	P+1M	/
Hydrocodone	P+3M	/	/	/
Cimétidine	P+3M	/	P+1M	/
Caféine	P+3M	P+3M	P+3M	/
Bisphénol A	P+3M	/	/	/
Ténofovir	P+2M	/	/	/
Spirapril	P+2M	/	/	/
Rosiglitazone	P+2M	/	/	/
Morphine	P+2M	/	P	/
Hydromorphone	P+2M	/	P	/
Fluoxétine	P+2M	/	/	/
Diphénylamine	P+2M	/	/	/
Triméthoprime	P+1M	/	/	/
Fénitrothion	P+1M	/	/	/
Matairésinol	P	/	P+1M	/
Acétamipride	P	/	/	/
Dibenzylidène sorbitol	P	/	P+2M	/

Tableau 2-9 : Nombre de composés parents et de métabolites prédits détectés dans les données LC-HRMS (positif/négatif) du méconium et du lait maternel de la cohorte EDEN pour la liste 2

	Méconium mode positif	Méconium mode négatif	Lait maternel mode positif	Lait maternel mode négatif
Parents	1927	823	1481	459
Métabolites	861	329	469	106

2.5. Conclusion

L'approche développée ici pour caractériser l'exposome chimique périnatal implique une étape préalable de prédiction des métabolites potentiels à partir d'une liste prédéfinie de contaminants, puis la recherche de leurs signaux dans des jeux de données LC-HRMS. Testée sur 60 composés exogènes issus de la base de données HMDB, cette approche a permis de prédire correctement plus de 70 % des métabolites connus. Les métabolites manquants étaient principalement liés à des biotransformations complexes qui n'étaient pas pris en compte dans l'approche utilisée ou qui nécessitaient plus de deux réactions successives. Une difficulté notable est l'augmentation exponentielle du nombre de structures générées en cas présence de multiples motifs ou lors de la prise en compte de réactions de biotransformation successives, ce qui allonge considérablement le temps de traitement.

L'efficacité de l'approche utilisée a pu être démontrée grâce à la détection de plus de 90 % des signaux attendus dans des matrices biologiques enrichies avec des xénobiotiques connus, en comparaison avec une analyse manuelle des données LC-HRMS produites.

L'application de cette approche aux données métabolomiques produites à partir des échantillons de méconium (n = 308) et de lait maternel (n = 320) de la cohorte EDEN a permis de détecter des signaux d'un certain nombre de marqueurs d'exposition candidats provenant des deux listes de contaminants suspectés : i) la liste contenant plus de 200 composés (incluant les contaminants jugés préoccupants par l'ANSES et ceux décrits dans la littérature) et, ii) celle plus élargie constituée de plus de 4 700 composés (base de données OpenFoodTox). Le temps nécessaire pour réaliser ces analyses augmentait avec la taille de la liste des molécules utilisée (105 minutes pour la première liste et de 600 minutes pour la deuxième). La majorité des métabolites détectés semble correspondre à la structure attendue ou prédite dont l'identité est renforcée par les valeurs de leurs défauts de masse, un paramètre utilisé dans des méthodes de filtration des données LC-HRMS pour cibler les métabolites de xénobiotiques.[122]. Une réduction des faux-positifs a également été obtenue pour des contaminants dont le temps de rétention est connu grâce à la base de données interne au laboratoire.

Ces résultats démontrent la faisabilité de notre approche pour détecter l'exposition périnatale aux xénobiotiques. Cependant, le manque de prise en compte des temps de rétention des métabolites pour lesquels nous ne disposons pas de composés standards représente une

limitation importante de notre approche. Enfin, l'utilisation de l'approche par « suspect screening » qui implique une recherche ciblée sur des contaminants prédéfinis, constitue un autre inconvénient. En effet, elle ne permet pas de détecter d'autres xénobiotiques non répertoriés dans la liste prédéfinie des molécules suspectées, bien qu'ils puissent être présents dans les matrices analysées. Une autre approche sans *a priori*, dite « non ciblée », est proposée dans le chapitre 3 pour analyser des jeux de données LC-HRMS et extraire les signaux potentiels de marqueurs d'exposition.

Chapitre 3. Développement d'une stratégie de traitement de données sans a priori pour de nombreux jeux de données acquis par LC-HRMS

Dans ce chapitre est présenté l'approche non-ciblée que j'ai développée pour le traitement de nombreux jeux de données acquises par LC-HRMS, ici appliquée à des jeux de données issus de méconium provenant de la cohorte EDEN (n=308).

L'objectif de ce travail est de pouvoir extraire de manière robuste les informations provenant des jeux de données afin de détecter des signaux reflétant l'exposition à des contaminants non répertoriés dans la liste de molécules suspectées (du chapitre 2). Pour cela, il faut préalablement extraire les données selon la procédure décrite dans la sous-section 2.2.2. L'étape suivante consiste à nettoyer les données en sélectionnant les signaux d'intérêt pouvant correspondre à des composés organiques, c'est-à-dire contenant des atomes de carbone. J'ai également développé une méthode de filtration permettant de sélectionner des couples de signaux possédant des différences de masses caractéristiques, comme i) des couples métabolites conjugués/non conjugués en considérant des réactions de biotransformation de phase 2 (glucuroconjuguaison, sulfoconjugaison conjuguaison au glutathion) communément observées lors du métabolisme des xénobiotiques ou ii) des composés halogénés tels que les molécules chlorées ou bromés, en se basant sur les caractéristiques du massif isotopique (valeurs *m/z* et abondances relatives). Par ailleurs, je me suis intéressé à la fréquence de détection des signaux, en partant du postulat que dans un contexte d'exposition à des xénobiotiques, une exposition à des contaminants serait caractérisée par une plus faible fréquence de détection.

Enfin, j'ai également développé une approche permettant l'annotation automatique (niveau 5) des espèces d'intérêt en attribuant les formules générées par un algorithme (Masstools) concordant avec toutes les informations précédemment recueillis.

Le travail réalisé sur le développement de l'approche ciblée appliquée à l'analyse des données métabolomiques acquises par LC-HRMS en mode positif sur les échantillons de méconium (n=308) de la cohorte EDEN est présenté dans ce qui suit sous forme d'un manuscrit d'article scientifique intitulé « *Exploration of chemical exposome in a mother-child cohort using a non-targeted data filtering strategy applied to a large and complex LC-HRMS dataset* », soumis au journal *Analytica Chemica Acta*.

3.1. Exploration of chemical exposome in a mother-child cohort using a non-targeted data filtering strategy applied to a large and complex LC-HRMS dataset

Dylan Saunier[1], Éric Venot[1], Sylvain Dechaumet[1], Blanche Guillon[1], Florence Castelli[1], Etienne Thevenot[1], François Fenaille[1], Blandine de Lauzon-Guillain[2], Karine Adel-Patient[1], Estelle Rathahao-Paris[1,3*]

[1] Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, MetaboHUB, 91191 Gif-sur-Yvette, France

[2] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Centre for Research in Epidemiology and StatisticS (CRESS), Paris, France

[3] Sorbonne Université, Faculté des Sciences et de l'Ingénierie, Institut Parisien de Chimie Moléculaire (IPCM), 75005 Paris, France

Corresponding author:

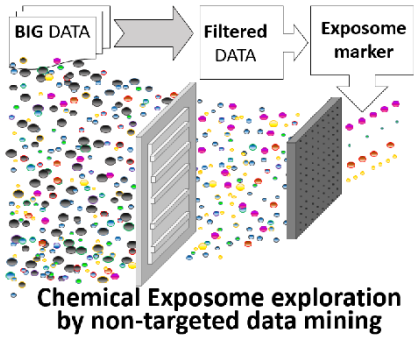
Estelle Rathahao-Paris

Département Médicaments et Technologies pour la Santé (DMTS), Université Paris-Saclay, CEA, INRAE, 91191 Gif-sur-Yvette, France;

orcid.org/0000-0002-7271-7372;

Email: estelle.paris@inrae.fr

GRAPHICAL ABSTRACT



ABSTRACT

Background: The field of exposome research has grown significantly in recent years. Although advances in analytical techniques such as high-resolution mass spectrometry (HRMS) allow large chemical coverage, especially when coupled to a separation system such as liquid chromatography (LC). However, detecting exposure markers among several hundred or even thousands of signals in untargeted data analysis remains a challenging task. A data mining strategy is needed to process such large datasets generated by LC-HRMS for the specific characterization of exogenous compounds.

Results: The developed workflow includes isotope pattern data matrix enrichment (IPDE) (e.g. specific for $^{12}C/^{13}C$ isotopes), a metabolite filter and a metric to attribute the detection frequency of every feature among all samples. The mass defect plot was used to display all data before and after the application of each data mining method. Applying this strategy to our data generated from meconium samples belonging to the EDEN mother-child study, resulted in a 6-fold reduction in the number of features. Among these cleaned datasets monohalogenated species and putative conjugated/unconjugated metabolites were highlighted. Molecular formulae of relevant features could be generated from their accurate *m/z* values using a molecular formula prediction algorithm combined with IPDE results, reducing the number of possible candidates. Common xenobiotic markers, such as paracetamol, caffeine and nicotine, were successfully detected, demonstrating the potential of meconium as a sentinel matrix to reveal *in utero* exposure to various xenobiotics.

Significance: Data cleaning is essential to provide more reliable features. Interestingly, early life exposure was demonstrated in our study through the detection of certain xenobiotic markers in meconium. The proposed data mining strategy is very promising for exploring the chemical exposome in large untargeted LC-HRMS datasets. It can also be used to process data from any matrix as well as mass spectrometry platforms.

Keywords: Chemical exposome, data filtering method, mass defect profile, early life exposure, LC-HRMS large datasets

3.1.1. Introduction

Mass spectrometry (MS) has become an essential tool in the analytical world,[79] thanks to its high sensitivity and specificity, allowing the separation of ions according to their mass-to-charge ratio (*m/z*) and the characterization of a wide range of molecules present in trace amounts in complex matrices. It is applicable to various fields, including environmental,[139] food[140] and biological analysis[141]; the approach used depends on the performance of the instrument. Targeted analysis of a specific ion or a list of known species could be performed by selected ion monitoring (SIM) or multiple reaction monitoring (MRM), respectively, using triple quadrupole or ion trap.[142,143] On the other hand, non-targeted analysis aimed at detecting a large number of ions without any information requires high-resolution mass spectrometry (HRMS) analysers such as time of flight (ToF)[144], Fourier-transform ion cyclotron resonance (FT-ICR)[145] or Orbitrap[146].

In addition, the coupling of mass spectrometry to a chromatographic system[147] provides an additional dimension for the separation of compounds based on their physicochemical properties, and the separation of certain isomers requires specialized chromatographic methods. Different techniques, such as gas chromatography (GC) or liquid chromatography (LC) can be coupled to mass spectrometry for the analysis of volatile or non-volatile molecules, respectively.[148,149] The combined use of both techniques, i.e. GC-MS and LC-MS, provides complementary results and improves the coverage of molecules, as demonstrated by Nil *et al* . who proposed its use for the comprehensive metabolic profiling in the study of aristolochic acid-induced nephrotoxicity.[150] Among the most relevant metabolites, nine metabolites were detected by GC-MS while the other six metabolites were observed by LC-MS analysis.

LC-HRMS is now widely used for complex mixture analysis, especially for omics studies such as metabolomics[151], proteomics[152] and exposomics[153]. For the latter, which is a very broad concept encompassing various exposure factors, it is perhaps more appropriate to talk about chemical exposome which refers to all chemical compounds originating from external

sources such as food and environment as well as pharmaceuticals, when using such an analytical approach.

The characterization of the chemical exposome is quite complicated, especially from complex matrices such as biological ones. From the exposure of an individual to one or more chemicals to the characterization of exposure markers, several factors can affect the results. As the initial compound to which an individual is exposed is metabolized to form metabolites[154], it may no longer be concentrated enough to be detected in its initial form. In addition, the possibility of detecting the parent molecule and/or of its metabolites depends on both the level of exposure and the type of biological sample examined, which is related to the exposure window and the biodistribution of chemicals in different organs. Another factor influencing the detection of exposure markers is the genetic supported inter-individual differences in xenobiotics metabolism, which is reinforced by the individual epigenome.[155] Nevertheless, the large number of signals detected by LC-HRMS analysis of biological samples makes the data generated complex and impossible to process manually. Computational tools are, therefore, required to facilitate their analysis, from data pre-processing[101] to data interpretation.

Several mass defect-based approaches[115] have been developed to explore such complex datasets, providing a simplified two-dimensional representation of HRMS data by reporting nominal masses (i.e. rounded to the nearest integer)[156] *versus* mass defects of all detected species: the mass defect (MD) is the difference between the accurate *m/z* value and the nominal mass of each species. Since The MD value is related to the elemental composition of the molecules, the graphical representation of the MD profile provides the molecular fingerprint of an analyzed matrix. A variant of this approach, called Kendrick mass defect (KMD), has been used to detect series of homologous compounds of the same class or type but with different numbers of repeated units in very complex HRMS data.[157] This method consists in converting all measured *m/z* values to Kendrick mass by taking into account the mass of the

repeating unit (i.e. multiplying each measured *m/z* value by the nominal mass/accurate mass ratio relative to the repeating pattern, e.g., 14.0000/14.0156 for a methylene unit (CH_2) in petroleum analysis). Another graphical data analysis, the van Krevelen diagram, exploits the property of assigned chemical formulae.[117] Although useful for revealing specific classes of compounds, this approach is reserved for data produced by very high resolution devices, such as the FT-ICR instrument, to avoid detrimental errors in the calculation of molecular formulae.[158] More interesting is the potential of MD-based approaches in the field of exposomics, as demonstrated by the work of Zhang *et al* ,[122] who selectively extracted drug metabolites from LC-HRMS data of biological matrices using a mass defect filter (MDF). Similarly, the use of data filtering methods, particularly isotope pattern filters, for the selective detection of metabolites of a chlorinated xenobiotics from the urine sample analysis has been reported.[159,160]

Although some automated processing tools exist, most of which being proprietary software designed for specific manufacturer's instruments, such as MMDF (multiple mass defect filter, Thermo Fisher Scientific) and Metabolynx (Waters), they are not easily applicable to process large amounts of data generated from cohorts. Another processing tool designed to explore multiple untargeted datasets, Haloseeker, was developed to selectively extract signals of halogenated compounds,[119] but it could not be used to detect other exogenous compounds.

Here, we propose a strategy for untargeted exploration of the chemical exposome in large LC-HRMS datasets obtained from metabolic profiling of biological matrices by extracting reliable signals suspected to be of exogenous origin without any *a priori* information. The developed workflow was designed with three main objectives: i) firstly, to clean the datasets from artefacts and keep only signals that prove the carbon isotope pattern, thereby increasing the reliability of the data matrix for subsequent analyses; ii) secondly, to automatically detect the signature of species containing specific atoms such as chlorine and bromine, which are often found in

various exogenous compounds (e.g., pesticides, flame retardants); iii) thirdly, to extend the range of chemical exposure markers by tracking pairs of conjugated and unconjugated metabolites and assigning the frequency of detection to each feature. The developed workflow was then applied to the LC-HRMS metabolomics datasets obtained from meconium collected in a French birth cohort, the EDEN mother-child study, which aims to investigate the early pre- and postnatal determinants of child development and health.[161] The MD plot was selected to display all data before and after the application of each data mining method.

3.1.2. Materials and methods

3.1.2.1. Metabolic Profiling by LC-HRMS

Meconium sample preparation - Meconium samples were collected from two French university hospitals (in Nancy and Poitiers) in the EDEN birth cohort (n = 2002 mother-child dyads)[161] and were stored at -80°C from collection until use. From samples of an EDEN sub-cohort in which we have already analyzed the levels of various cytokines and growth factors in order to assess their association with food allergy in childhood (n=317; case-cohort design), 308 samples from the same mother-child dyads were available in sufficient quantity to performed metabolic profiling. The method used to process meconium samples was developed and described in a previous methodological work.[129] Briefly, meconium samples were freeze-dried using a Triad™ Labconco freeze dryer (Missouri, USA) with temperatures set at 4 °C and – 83 °C for the tray and trap, respectively, and a vacuum of 0.180 mbar. Then, 10 mg of lyophilized meconium were suspended in 750 µL of methanol/H_2O (4:1, v/v) and homogenized at 6500 rpm, and 4 °C, for 3 x 30 s in a tube containing CK14 ceramic beads (Ozyme, Saint-Cyr-l'Ecole, France) using a Precellys 24® device (Bertin Technologies, Montigny-le-Bretonneux, France). After incubation on ice for 1.5 h to achieve complete deproteinization, the samples were centrifuged at 20,000 g for 15 min at 4 °C, then the supernatants were collected and divided into 200 µL aliquots before drying under anitrogen stream at 30 °C using

a Turbovap® (Biotage, Uppsala, Sweden). The dried extracts were stored at -80 °C until use. For LC-MS analysis, each dried extract was resuspended in 215 µL of H_2O/acetonitrile (95:5, v/v) containing 0.1% formic acid.

LC-MS data acquisition - Metabolic profiling was performed on meconium extracts from the sub-cohort (n = 308) using the optimized LC-HRMS protocol of our platform.[162] Briefly, mass spectrometry data were acquired in the positive electrospray ionization (ESI) mode on an Ultimate 3000 chromatography system coupled to a Q-Exactive mass spectrometer (both from Thermo Fisher Scientific, Courtaboeuf, France). Chromatographic separation was performed on a Hypersil GOLD C_{18} 1.9 µm, 150 × 2.1 mm i.d. column (Thermo Fisher Scientific).

3.1.2.2. Data processing workflow

The data processing workflow proposed in this study consists of several steps, ranging from data pre-processing to data mining, including data filtering methods, molecular formula calculation and data visualization *via* MD plotting (Scheme 1).

In our study, R, a free software environment for statistical computing, was used to process our LC-HRMS data from pre-processing to data mining.[163] All developed scripts can be freely available upon request.

3.1.2.2.1. Pre-processing

All raw data from the LC-MS analyses were first converted to the mzXML format using a freely available converter tool, msConvert from ProteoWizard.[131] The raw signals from the mzXML files were pre-processed into a set of features enclosed in a data matrix using the XCMS package,[101] a free open source software dedicated to pre-processing all types of mass spectrometry acquisitions. This XCMS package consists of a set of feature extraction functions, including peak picking and data grouping. The following parameters were used: peakwidth = 9-20, noise = 500, prefilter = 6, ppm = 3, mzdiff = 0.0001, sntresh = 10 for the peak picking part. These parameters were optimized based on a peak list of low intensity signals within our

batches and several parameters were evaluated by peak integration from the peak list. The data grouping parameters were as follows: minFraction = 0, bw = 5, binSize = 0.005. Note that the minFraction value was intentionally set to 0 in order not to discard features present in only a few samples, as these features may correspond to markers of rare exogenous compounds that were not widely present in the general population.

The resulting data matrix was further processed using the following data mining methods.

3.1.2.2.2. Data mining methods

Blank filter **(1)** - To remove noise and artefact signals from the data matrix, the intensity of each feature detected in the analytical blanks was compared to that observed in the biological samples. Only features which intensities were sufficiently higher than the highest one detected in the blanks were extracted. In our study, this ratio was set to three.

Isotope pattern data matrix enrichment (IPDE) **(2)** - To enrich the data matrix with features having a selected isotopic pattern, an isotope filtering method[164] was used. This involves tracking isotope ion pairs of interest in the spectral data, such as the A and (A + 1) ion pairs for the ^{12}C and ^{13}C isotopes, respectively, for organic species and, the A and (A + 2) ion pairs for species containing the $^{79}Br/^{81}Br$, $^{35}Cl/^{37}Cl$ or $^{32}S/^{34}S$ isotopes. More precisely, to perform an IPDE, a *m/z* difference value was predefined as follows: $\Delta = 1.0034 \pm 0.0005$ u (unified atomic mass unit) for $^{12}C/^{13}C$ isotopes, $\Delta = 1.9979 \pm 0.0005$ u for $^{79}Br/^{81}Br$ isotopes, $\Delta = 1.9970 \pm 0.0005$ u for $^{35}Cl/^{37}Cl$ isotopes and $\Delta = 1.9958 \pm 0.0005$ u for $^{32}S/^{34}S$ isotopes. Note that a mass tolerance window of ± 0.0005 u (i.e. 2.5 ppm at *m/z* 400) was chosen based on the reported mass accuracy of less than 3 ppm with the external mass calibration of a Q-Exactive instrument[165] this is consistent with the mass accuracy observed for the monoisotopic peak of a spiked standard reference compound (i.e. the $[M+H]^+$ species of amiloride ($C_6H_9ON_7Cl$) detected at *m/z* 230.0552 with 1 ppm error). In addition, a relative isotopic abundance (RIA) filter was applied, taking into account the theoretical RIA value of the isotope pairs considered,

using the ratio interval of ± 10 %, which is consistent with the RIA error less than 20 % for most compounds analyzed on the Orbitrap analyzer,[166] except for the $^{12}C/^{13}C$ isotopes. In this latter case, the RIA filter was based on the maximum possible carbon number, estimated from the *m/z* value of the feature considered (see Figure S1 in the Supplementary Information). Multiple signals reflecting a more complex isotope distribution may be observed for any species containing multiple halogen atoms, such as polyhalogenated compounds. These signals will not pass the RIA filter specific to monohalogenated species (e.g., species containing $^{79}Br/^{81}Br$ or $^{35}Cl/^{37}Cl$ isotopes). In this case, another module lists all putative isotopic signals in an output file for further manual investigation (different cases are illustrated in Figure S2).

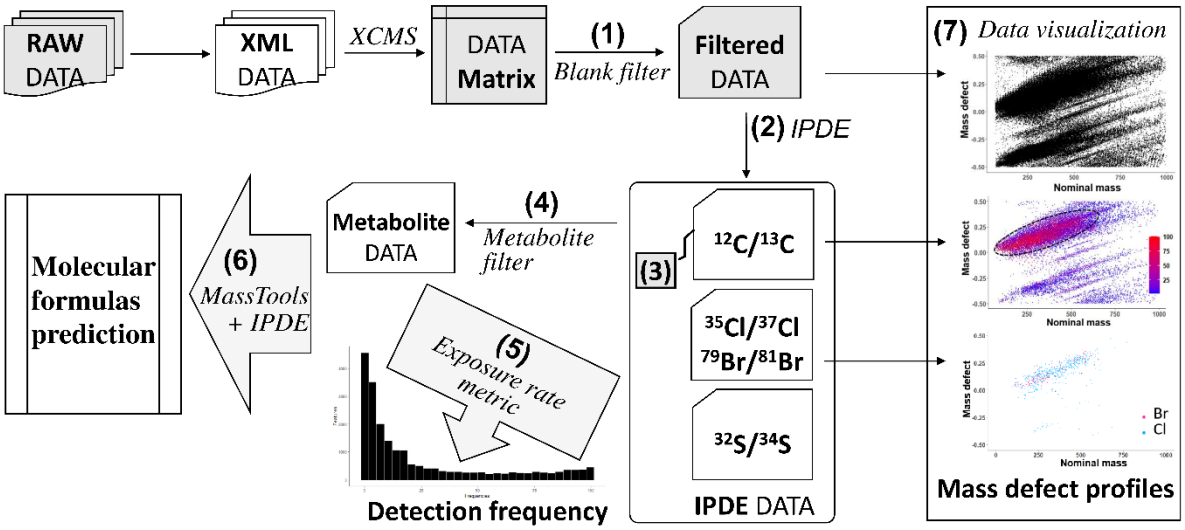
Sample-threshold filter **(3) -** For features that are not selected from all samples by IPDE specific to the $^{12}C/^{13}C$ pattern, a threshold can be set to define the minimum number of samples in which these features must be detected to ensure their reliability. Since the carbon isotope pattern is not always detected, especially for low intensity signals, a threshold of 10% was used in our study: only features which $^{12}C/^{13}C$ pattern is verified for species observed in at least 10% of the total samples were selected for further analysis.

Metabolite filter **(4) -** *(or selection of metabolites resulting from biotransformation of xenobiotics)* **-** Ion pairs corresponding to potential conjugated and unconjugated metabolites (or parent compound) were extracted based on their accurate mass differences such as 176.032 u, 305.068 u and 79.957 u for glucuro-, sulfo- and glutathione-conjugation, respectively.

Exposure rate metric **(5)** – Here, we propose to examine the number of samples in which a feature is observed in order to determine its detection frequency among all samples examined. This value, estimated in %, could be qualified as an "exposure rate metric" to highlight potential exogenous species. For example, the lower the detection frequency of a given species, the more likely the compound to which an individual or a subgroup of individuals is exposed is of xenobiotic or chemical origin.

Annotation step **(6) -** The R molecular formula calculation package from MassTools[5] was used to predict chemical formulae with the accurate *m/z* values of the features of interest. This function has many filters that take into account criteria described in the Seven Golden Rules[158] (such as restrictions on the number of elements, LEWIS and SENIOR chemical rules, hydrogen/carbon ratios, element ratios of nitrogen, oxygen, phosphorus, or sulfur *versus* carbon, and element ratio probabilities). For each candidate generated from its accurate *m/z* value by considering a mass tolerance window of 3 ppm and a predefined list of elements (e.g., containing C, Cl, N, H, O, Br, S and P), additional information obtained from IPDE results (i.e. species detected with one or more specific atoms such as Cl, Br, S) was considered, in the same way as using isotopic patterns, one of the seven golden rules to eliminate false positives. A scoring system based on the detection of the corresponding $^{12}C/^{13}C$ isotope pattern was applied: each experimental pattern was compared with those artificially simulated for different numbers of carbon atoms, from C1 to C90. The closest matches were retained. To evaluate such a method, the number of carbon atoms expected based on the simulated pattern was compared with that corresponding to its real number found in the chemical formulae of some known compounds from the internal database of our laboratory; this allows to propose a prediction metric to support further annotation of all features of interest.

Data visualization **(7) –** The mass defect profile is a way to visualize all the features of LC-HRMS data by plotting their MDs against their nominal masses. Since the nominal mass[156] is defined as the mass of a molecular ion (or molecule) obtained from the mass of the most abundant isotope of each element, rounded to the nearest integer, the nominal mass in our study is obtained by rounding up to the nearest integer value if the number to the right of the decimal point is 5 or greater and, by rounding down to the nearest integer value if the number to the right of the decimal point is less than 5. Therefore, there are positive and negative MDs.



Scheme 1 Graphical representation of data processing workflow showing different steps including conversion of raw data to XML format, data pre-processing using XCMS package, (1) noise filtration, (2) isotope pattern data matrix enrichment (IPDE) with features of species having specific isotope signals such as 12C/13C patterns validated with (3) their detection in 10 % of the total samples, and with features of species containing halogen or sulphur, (4) extraction of features of potential conjugated/unconjugated metabolite pairs (e.g., conjugates as glucuronate, sulfate and glutathione), (5) assignment of detection frequency to each feature and, (6) molecular formula determination of relevant features. (7) All features of a data set can be graphically display, as an MD profile, by plotting the MD of each feature against its nominal mass.

3.1.3. Results and discussion

In this study, a data processing workflow was developed to specifically explore the chemical exposome in large data sets generated by untargeted analysis. Here, data obtained from positive ionization LC-HRMS metabolic profiling of 308 meconium samples from the EDEN cohort were submitted to the so-developed data processing workflow.

3.1.3.1. Complexity of LC-HRMS data displayed as the mass defect (MD) plot

The MD plot was chosen to visualize the complex data set generated by LC-HRMS. Such a graphical representation has been reported to be characteristic of the analyzed matrix, as suggested by Zhang *et al* .[122] To the best of our knowledge, the MD profile of meconium has never been described before. Here, the MD profile obtained from the LC-HRMS analysis of a meconium sample is shown in Figure 1. The red lines crossing the MD profile represent the

calculated MDs of alkanes, the simplest organic compounds composed of just two elements carbon and hydrogen. These lines show an expected general trend: the MD values increase with increasing *m/z* values, which is consistent with the contribution of the positive MD of hydrogen (i.e. MD = 0.0078 u) present in the species examined. Considering that the MD is the difference between the accurate mass and the nominal mass of each species and that the nominal mass is obtained by rounding up or down to the nearest integer value, depending on the number after the decimal point, there are positive and negative MD values. Therefore, the MDs of alkanes are divided into two distinct lines: i) one red line represents positive MDs for species with *m/z* < 470 u and ii) another red line corresponds to negative MDs for species with *m/z* > 470 u (Figure 1). These two lines should represent the MD limits for the majority of organic species, since other chemical elements in addition to carbon and hydrogen contribute to the MD value and all have negative MDs[115] except nitrogen (with MD = 0.0001 u), which has a small contribution. Therefore, positive MD values smaller than those of alkane MDs are observed for species with *m/z* < 470 u and larger negative MDs for species with *m/z* > 470 u. Negative MDs close to zero surrounded by a blue box are observed for low mass species (*m/z* < 200), suggesting that they may correspond to species composed of elements which sum of their MDs gives a negative value, such as $[M+Na]^+$, $[M+K]^+$ adduct ions due to the negative MDs of the sodium and potassium.[167] The negative MDs are also characteristic of polyhalogenated species with a high degree of unsaturation, such as polychlorobiphenyl (PCB) or polybromobiphenyl (PBB) derivatives[168]. Other signals located inside the blue ellipse are probably multi-charged species which should contain an odd number of nitrogen atoms according to the nitrogen rule. For example, a molecule containing an odd number of nitrogen atoms will have an odd nominal mass, so its di-charged species (with an even number of electrons) will have a MD around 0.5.

Although the MD profile provides insight into the molecular signature of a sample, the presence of a large number of molecules in complex matrices, for example, over 19,000 features detected in one meconium sample, complicates its interpretation and exploitation. This challenge is even greater in cohort studies. More than 155,000 features were observed in the MD profile of all 308 meconium samples analyzed by LC-HRMS (Figure 2.A).

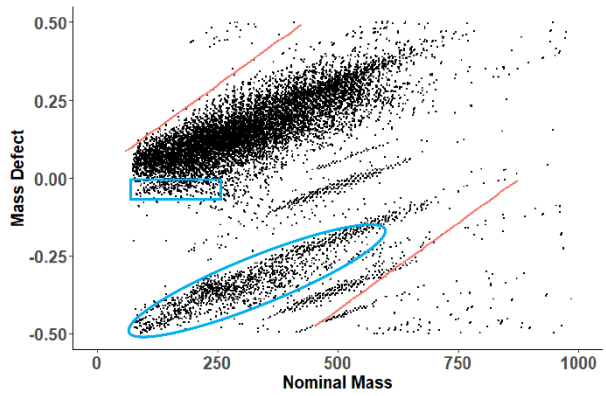


Figure 1 Mass defect profile from LC/HRMS (ESI+) data of one meconium sample containing 19,106 features, shown with two red lines representing to theoretical MDs for simulated alkanes, a blue box surrounding negative MDs species, and a blue ellipse inside which are putative multi-charged species and/or artefacts

3.2. Extraction of potential organic species and examination of their detection frequency

Data cleaning was first performed to reduce the size of the large dataset by removing non-relevant signals in order to extract more reliable features as potential organic species characterized by a specific $^{12}C/^{13}C$ isotope pattern. Our first approach was to apply the data mining methods described in the data processing workflow (Scheme 1), including 1) a blank filter to remove noise and artefact signals if they were detected in the blank injections (i.e. analysis of solvent and no sample) and their intensity was not high enough in the sample analyses, 2) an IPDE to extract only features from the characteristic $^{12}C/^{13}C$ isotope pattern and 3) a sample-threshold filter to validate the extracted features based on their detection in at least 10 % of all samples (n=308). These methods significantly reduced the number of features, from

over 155,000 to 25,276 features for the dataset from all 308 samples, which corresponds to a 6 fold reduction (Figure 2A and 2B). Examination of the detection frequency of the $^{12}C/^{13}C$ pattern observed for the 25,276 extracted features showed an average value well above the 10% threshold used in the sample-threshold filter, highlighting the relevance of all extracted variables (Figure S3C). Note that the sample-threshold filter should be less effective for signals detected in few samples, e.g., feature validation is less relevant if it is based on detecting the $^{12}C/^{13}C$ isotope pattern in 1 out of 10 samples than in 30 out of 300 samples. Although these 25,276 extracted signals have become more reliable with their detected $^{12}C/^{13}C$ pattern, there is no evidence of their origin, either endogenous or exogenous.

To highlight potential exogenous species, we examined the number of samples in which each feature was detected, which allowed us to determine its detection frequency by applying our exposure rate metric, i.e. data mining method (5). Based on the postulate that exposure of an individual or a subgroup to uncommon xenobiotics would result in features detected at low frequency, there would be a higher number of such features with low detection frequency due to the diversity of xenobiotics.

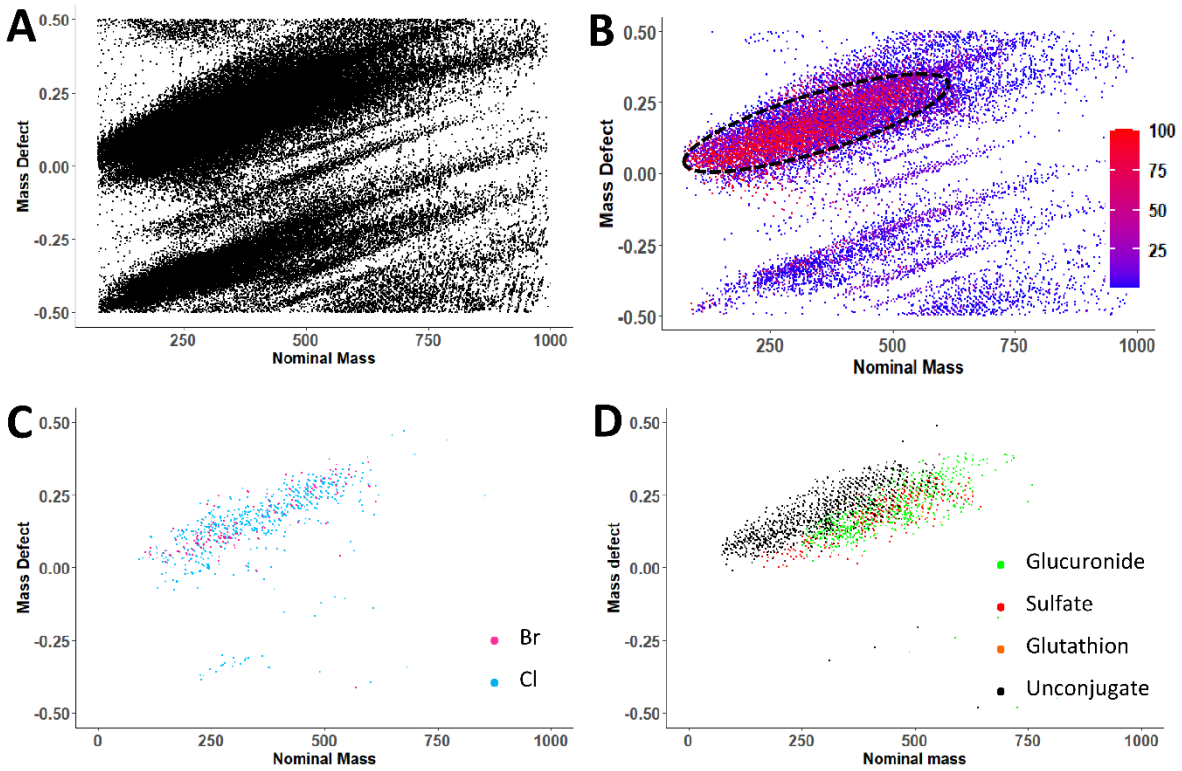


Figure 2 Mass defect profiles from LC/HRMS (ESI+) data of all 308 meconium samples showing A) 155,047 features before data mining, B) 25,276 features extracted by applying a blank filter and a $^{12}C/^{13}C$ IPDE, C) 1030 features filtered by an additional IPDE providing a monohalogen signature (i.e. brominated and chlorinated species) and, D) 5017 features selected by applying a metabolite filter to the data from B, which are potential conjugated and unconjugated metabolite pairs, including features of the parent compounds. In B, the detection frequency of each feature is represented by colors ranging from blue to red, from low to high detection frequency; a dashed black ellipse highlights a region where features are detected with high frequency

Although the low detection frequency signals may reflect a greater diversity of the chemical exposome, their relevance is limited, especially when the detection frequency is too low, i.e. between 0 and 5 %. In fact, most of them (about 60 000 features) were present in Figure 2.A, but they could not be classified as organic compounds and were then removed during the data cleaning step using the blank filter and IPDE to extract only features with the $^{12}C/^{13}C$ isotope pattern (Figures S3A and S3B).

Nevertheless, low detection frequency features observed in less than 25 % of the total number of meconium samples (n=308) remained dominant even after the data cleaning step (Figure 2B). They are more dispersed than high detection frequency features (i.e. detection frequency higher than 50 % of the total number of samples), most of which are located in a specific region within the dashed black ellipse at *m/z* below 650 u and with positive MDs. It should be emphasized that these features with high detection frequency should mainly correspond to endogenous species that are characteristic of the samples, as reported by Zhang *et al* .[169] for several biological matrices (i.e., plasma, urine, bile, and feces). However, some of them may also originate from common environmental xenobiotics as well as from drugs commonly taken by the majority of the French population.

Preliminary annotation of some of the 25,276 extracted features was performed by querying our in-house database. 164 species could be annotated based on their accurate *m/z* and retention time using tolerance windows of 2.5 ppm and of 10 seconds, respectively (Table S1). Although the majority of them were annotated as endogenous metabolites, some species could correspond to markers of everyday xenobiotics, such as acetaminophen (i.e. paracetamol), caffeine and nicotine, reflecting a very common chemical exposome. The detection frequencies of these markers were consistent with the exposure frequencies reported in the literature for pregnant women, i.e. 68 % *vs.* 67 %, 100 % *vs.* 81 %, 11 % *vs.* 16 % for exposure to acetaminophen,[170] caffeine,[171] and nicotine,[172] respectively (Figure S4 and Table S1). Note that caffeine and acetaminophen have been reported as the two most commonly detected xenobiotics in meconium.[25] Cotinine, known to be a marker of nicotine exposure in biological matrices, was also detected, providing even more accurate results with a detection frequency of 16%, similar to that already described[173] and to tobacco exposure in the EDEN cohort (i.e. 17 % from self-declaration). This demonstrated how meconium analysis could help to address women's chemical exposome during pregnancy, and more importantly, the *in utero* exposure

of the fetus. It also highlighted the challenges of interpreting data from such a complex biological matrix as meconium.

Detection frequency can help to select features for annotation if it correlates with relevant literature or cohort study data. However, without any prior knowledge, additional clues are required to select some relevant features.

3.3. Tentative detection of signature of exogenous compounds

Additional data mining methods were applied to the cleaned data of 308 meconium samples, containing 25,276 extracted features (Figure 2B) to highlight potential features of exogenous compounds. IPDE was used to selectively extract monohalogenated species based on the presence of their characteristic isotopic patterns (i.e. brominated and chlorinated species). This allowed a significant reduction in the number of features from more than 25,000 to about 1000 features (Figure 2C and Table 1). Surprisingly, the number of monohalogenated species was relatively high, representing 1/25 of the total number of features extracted after data cleaning, with a larger proportion of monochlorinated species, representing nearly 3/4 of the total number of monohalogenated species. Table 1 shows that about 50 % of the monohalogenated species were found in more than 50 % of the meconium samples (i.e. detection frequency > 50%), which probably reflects a more global exposure to these halogenated compounds whereas features with a low detection frequency, especially below 25%, should reveal specific individual exposures.

Another data mining method, a metabolite filter, was used to track specific metabolites that could result from the three known classes of phase II metabolism, i.e. glucuronidation, sulfation and glutathione conjugation. A total of 5017 features were selected by applying such a metabolite filter (Figure 2D). They could correspond 2342 unconjugated metabolites or parent compounds and 2675 putative conjugated metabolites presented in Table 1, with their detection

frequencies. A large number of glucuronide metabolites, i.e. 1700 features, were detected, and among them, 720 features had specifically low detection frequencies below 25 %. A smaller number of sulfated and glutathione-conjugated metabolites were observed, with 933 and 42 features, respectively, with different detection frequencies. 324 species could be assigned to the glucuronide-sulfate conjugates, and only 9 species to the other two double conjugates, i.e. eight glucuronide-glutathione conjugates and one sulfate-glutathione conjugate. Note that other metabolites, such as cysteine or N-acetylcysteine conjugates as well as glycosidic conjugates, could also be tracked with the same filtering method by considering specific mass differences.

Table 1 Species extracted by application of the IPDE specific to monohalogenated species or of the metabolite filter, and their distribution according to their detection frequencies within the cohort studied

Filter	Species\Frequency (%)	0-25	25-50	50-75	75-100	Total
IPDE	Chlorinated	278	132	157	220	787
	Brominated	59	43	56	85	243
Metabolite filter[a]	Glucuronide	720	353	255	372	1700
	Sulfate	241	215	182	295	933
	Glutathione	19	7	4	12	42

[a] A total of 5017 features were extracted by the metabolite filter used, of which 2342 were non-conjugate species. Among the 2675 putative conjugated metabolites, possible double conjugates were detected: 324 glucuronide-sulfate conjugates, 8 glucuronide-glutathione conjugates and one sulfate-glutathione conjugate.

Among the 1700 glucuronide metabolites extracted from the metabolite filter, one metabolite of acetaminophen, i.e. its glucuronide conjugate, was detected at 1.6 min as a protonated species at *m/z* 328.1028 ($C_{14}H_{18}NO_8$, 1 ppm error) in 134 samples, corresponding to a detection frequency of 43.5 %. The identity of this metabolite was supported by the presence of the

protonated acetaminophen species at *m/z* 152.0708 ($C_8H_{10}NO_2$, 1 ppm error) at the same retention time and therefore in the same mass spectrum, most likely due to in-source fragmentation of the glucuronide metabolite species (Figure 3). It should be noted that the detection frequency of the protonated acetaminophen species observed at 1.6 min resulting from in-source fragmentation was very close to that detected for the protonated acetaminophen annotated at 2.8 min, i.e. 64 % *vs.* 68 %. These values are very different from the detection frequency of the glucuronide metabolite (43.5 %), which could be explained by the decomposition of the conjugate metabolite leading to its non-detection in certain samples.

Although the species extracted after application of the metabolite filter cannot be assigned to exogenous sources, the presence of redundant signals for the same species (e.g., precursor and its metabolites) could help in their characterization.

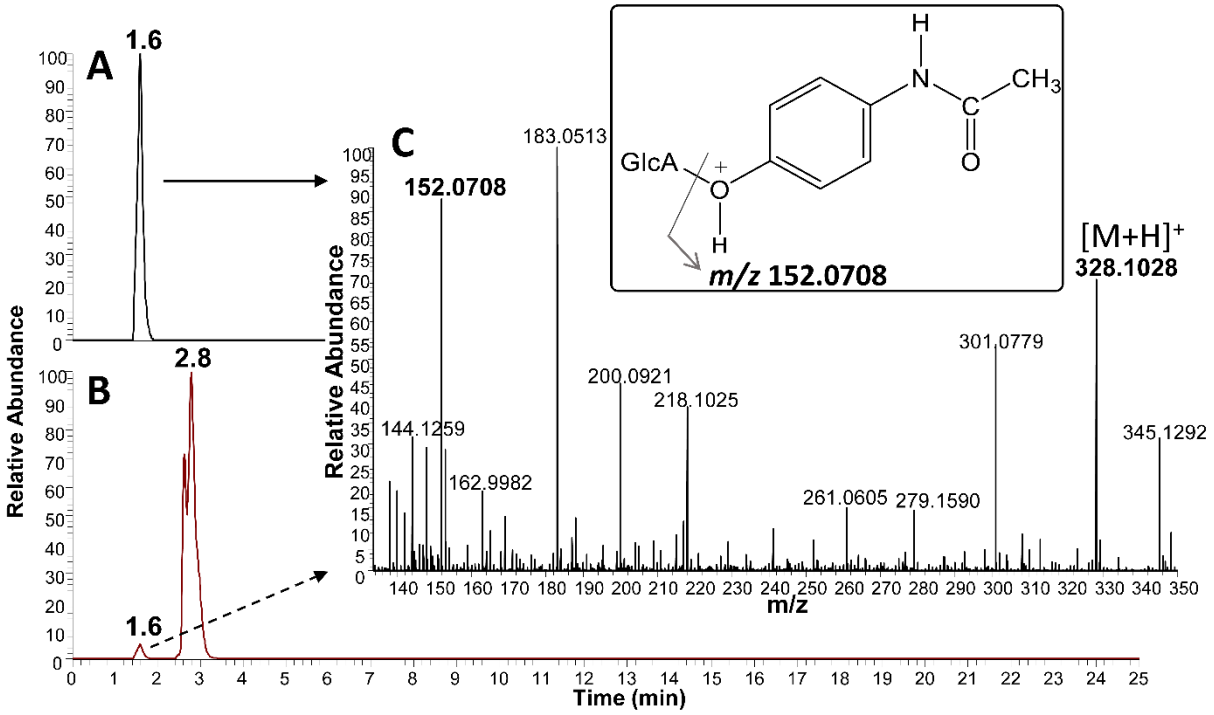


Figure 3 Results of LC-HRMS analysis of a meconium sample, showing reconstituted chromatograms A) of *m/z* 328.1028 ions annotated as the $[M+H]^+$ protonated species of acetaminophen glucuronide at 1.6 min and, B) of *m/z* 152.0708 ions annotated as the $[M+H]^+$ protonated acetaminophen at the two retention times, i.e. at 1.6 min and 2.8 min and, C) a mass spectrum recorded at 1.6 min, displaying the presence of both *m/z* 328.1030 and *m/z* 152.0708 ions, supporting the identity of the protonated species of acetaminophen glucuronide ($C_{14}H_{18}NO_8$, 1 pm error) since the *m/z* 152.0708 ions $[M+H]^+$ protonated acetaminophen ($C_8H_{10}NO_2$, 1 ppm error) probably resulted from in-source decomposition of its conjugated glucuronide species. (GlcA, glucuronic acid)

3.1.3.2. Molecular formula assignment

To go further in the assignment of a chemical formula to every species, a strategy was developed to allow automatic processing of a large number of species, such as that obtained after the data cleaning step, i.e. 25,276 features, according to the method described in the Data processing section. It should be noted that the prediction of molecular formulae from accurate *m/z* values is challenging because the number of possible candidates increases with the masses and also depends on several criteria, as demonstrated by Kind and Fiehn.[158] These authors proposed

the use of the seven golden rules to significantly reduce the number of false positives based on different rules for selecting the most probable molecular formulae.

To evaluate the efficiency of our strategy, the 25,276 features obtained after the data cleaning step (Figure 2.B) were queried against our in-house database, which contains data of more than 1,000 small molecule standards. As mentioned previously, 164 species were annotated using a mass tolerance of ± 0.0005 u and a retention time tolerance window of 10 seconds (Table S1). However, the number of molecular formulae generated by the MassTools algorithm for these 164 annotated species depended on whether or not the IPDE results were considered (Figure 4A). In fact, the number of proposed candidates decreased when the results obtained from the IPDE filter were included. In addition, 143 out of the 164 species were successfully annotated with correct molecular formulae by combining with IPDE results. The expected number of carbons estimated from the *m/z* value of the considered features was compared with that of the known molecular formulae (Figure 4B). Most molecular formulae were generated with an error of less than 4 in the number of carbon atoms and 58 % of the predicted carbon numbers matched the true formulae (i.e. zero error in the number of carbon atoms). The error in predicting carbon number seemed to increase with increasing *m/z* values (Figure S6). Our strategy was then applied to annotate the 25,276 features obtained after the data cleaning step (Figure 2B), allowing a significant reduction in the number of proposed chemical formulae, as shown in the zoomed area of the Figure 4C. From 25,276 features, at least one formula could be assigned to 20,970 and 19,285 species before and after the IPDE check, respectively.

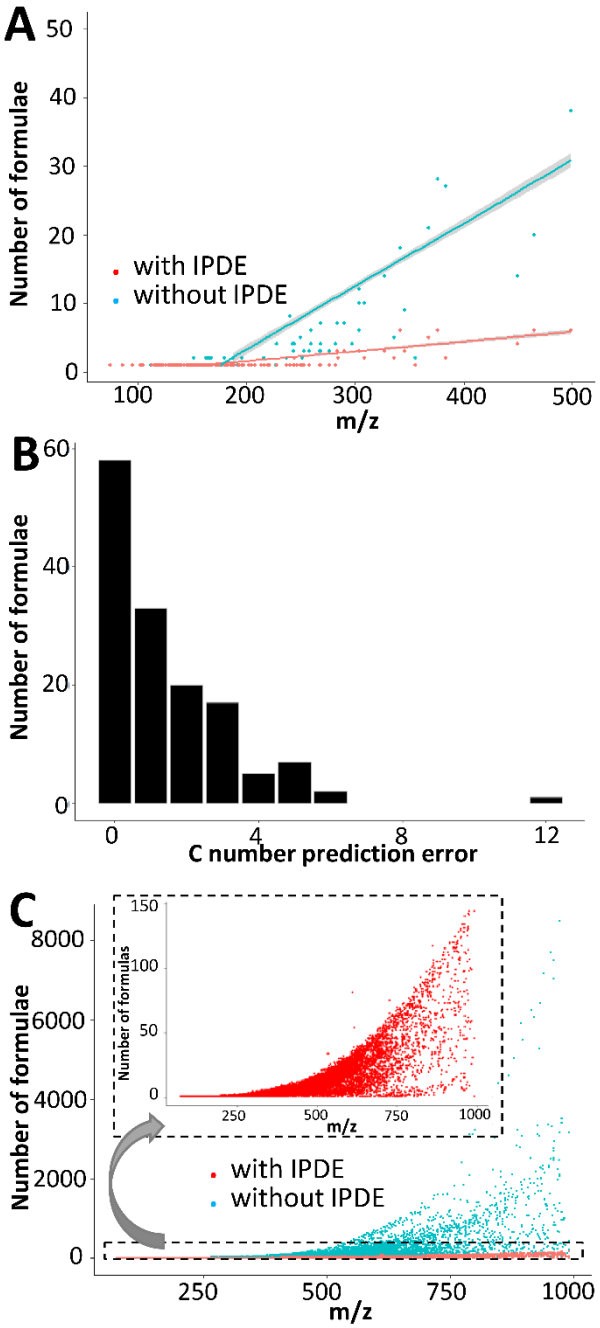


Figure 4 Graphical representation of the number of predicted molecular formulae A) for the 143 species annotated from our in-house database before and after IPDE application, B) with error in the number of carbon atoms and, C) for the 25,276 features obtained after the data cleaning step combined or not with IPDE results; the zoomed area shows the number of molecular formulae with IPDE results. (IPDE applied in red, and not applied in blue)

As an example of our approach, 10 chemical formulae (Table S2) could be proposed for species detected with a monochlorinated signature at *m/z* 331.1566 (Figure 2.C), but only four of them, i.e. $C_{19}H_{24}ClN_2O$, $C_{11}H_{28}ClN_4O_3S$, $C_{13}H_{25}ClN_6P$ and $C_{11}H_{30}ClN_4OP_2$, seemed to be more suitable due to the presence of a chlorine atom. Among these four candidates, only one contains a sulfur atom, but the specific signature of the ^{32}S /^{34}S isotope pattern was not observed for the *m/z* 331.1566 species. Therefore, the proposed chemical formula containing sulfur atom was not validated. It should be emphasized that the non-detection of low abundance isotopes could lead to the elimination of true positive hits. Nevertheless, a large tolerance on the parameters used to generate chemical formulae could lead to a too much drastic increase in false positive hits. The main objective of our approach was to focus on reliable isotope signals and signatures, in order to minimize the number of false positive results as much as possible. Furthermore, the elemental composition $C_{19}H_{24}ClN_2O$ is the closest to the expected number of carbon atoms (i.e. 22) determined from the *m/z* value of 331.1566 u, with an error of 2 ppm. This compound may correspond to hydroxyclomipramine, a metabolite of the antidepressant clomipramine (Figure S5).

3.1.4. Conclusions

In this work, we have demonstrated the efficiency of our strategy to explore the chemical exposome in large and complex LC-HRMS datasets from a cohort study. The proposed data cleaning procedure provided more reliable features, i.e. specific to organic nature based on the $^{12}C/^{13}C$ isotope pattern, for further data mining aimed at characterizing the chemical exposome, but also for more global use in classical metabolomics studies. Specific molecular signatures such as those of halogenated species and conjugated metabolites could be highlighted using more specific filtering methods. By assigning the detection frequency of the features, it is possible to distinguish species with high detection frequency, which may correspond to endogenous compounds but also to environmental xenobiotics and drugs to which most of the population is exposed, from species with low detection frequency, which are likely to be associated with individual or sub-group exposures. Rapid annotation of a large number of features with more accurate molecular formulae could be achieved through the strategy including information collected by the IPDE. We have demonstrated a positive *in utero* chemical exposome through maternal exposure during pregnancy and trans-placental passage, by detecting certain xenobiotic markers, such as paracetamol, nicotine, and caffeine in meconium, which is supported by detection frequencies consistent with those reported in the literature or with data collected in the EDEN cohort, highlighting the potential of meconium to reveal exposures during the perinatal period. Nevertheless, identifying a exposure marker with high confidence remains challenging, requiring additional analyses such as tandem mass spectrometry (MS/MS) experiments, interpretation of fragmentation pathways by an expert to achieve structural elucidation, and most importantly, but not always possible, the availability of a standard reference for unambiguous structural identification (level 1 of identification)[125]. Interestingly, as the ultra-high resolution detection has become the predominant method in untargeted analysis, the isotopic fine structure provided by such an approach is very useful for

unambiguous determination of the elemental composition of unknown compounds using the mass defect of isotopes such as ^{13}C, ^{15}N, ^{18}O, and ^{34}S.[174]

Availability

All developed scripts will be made available upon request.

Acknowledgements

The authors thank Drs Nihel Bekhti and Vincent Marie for the LC-HRMS analysis of meconium.

Funding

This work was supported by the FRM (Fondation pour la Recherche Médicale) (grant no.: ENV202109013943).

The EDEN study was supported by the FRM, French Ministry of Research: Federative Research Institutes and Cohort Program, INSERM Human Nutrition National Research Program, and Diabetes National Research Program (by a collaboration with the French Association of Diabetic Patients [AFD]), French Ministry of Health, French Agency for Environment Security (AFSSET), French National Institute for Population Health Surveillance (InVS), Paris-Sud University, French National Institute for Health Education (INPES), Nestlé, Mutuelle Générale de l'Education Nationale (MGEN), French-speaking Association for the Study of Diabetes and Metabolism (ALFEDIAM), National Agency for Research (ANR non-thematic programme), and National Institute for Research in Public Health (IRESP: TGIR 2008 cohort in health programme).

3.2. Supplementary information

SUMMARY OF CONTENTS

Figure S1. Theoretical mass spectra showing the usefulness of the relative isotope abundance (RIA) filter when applying an isotope pattern data matrix enrichment (IPDE) to extract species containing the $^{12}C/^{13}C$ isotope pattern. A) Case where two signals are detected with a mass difference of 1.0034 u ±0.0005, but their relative abundances do not match the $^{12}C/^{13}C$ RIA. B) Case where two signals are detected with an acceptable $^{12}C/^{13}C$ RIA, i.e., the abundance ratio of the ^{13}C isotope peak of about 25% is within an estimated interval.

Figure S2. Theoretical isotope patterns of mono- and multi-chlorinated compounds, such as A) clozapine, B) vinclozolin, and C) 2,3,7,8-tetrachlorodibenzo-p-dioxine (TCDD).

Figure S3. Distribution of features according to their detection frequency, A) before and after the data cleaning step using the blank filter, an IPDE to extract only features with the $^{12}C/^{13}C$ isotope pattern detected in at least 10 % of all samples (n=308), B) for features removed by the data cleaning step, C) for the $^{12}C/^{13}C$ isotope pattern actually detected (green plot) in samples which is normalized on the y-axis, the blue and red lines represent the simulation of features detected with a validated $^{12}C/^{13}C$ isotope pattern in all samples (blue line) and in 10% of the total samples (red line).

Figure S4. Distribution of the 25,276 features obtained after the data cleaning step according to their detection frequency. Some features could be annotated as common xenobiotics such as nicotine, acetaminophen and caffeine. Their detection frequency within the studied cohort is indicated.

Figure S5. A) Experimental mass spectrum from the LC-HRMS analysis of a meconium sample, showing the presence of the $[M+H]^+$ *m/z* 331.1566 species, corresponding to a chlorinated compound and, B) Theoretical mass spectrum of protonated 2-hydroxyclomipramin.

Figure S6. Error in predicting the number of carbon atoms *vs. m/z* value for the 143 species annotated using our in-house database. Regression analysis showed an increase in carbon number prediction error with the *m/z* values.

Table S1. Preliminary annotation of 164 species based on their accurate *m/z* values (with a mass tolerance window ± 0.0005 u) and retention time (RT, tolerance window of 10 seconds) using our in-house database.

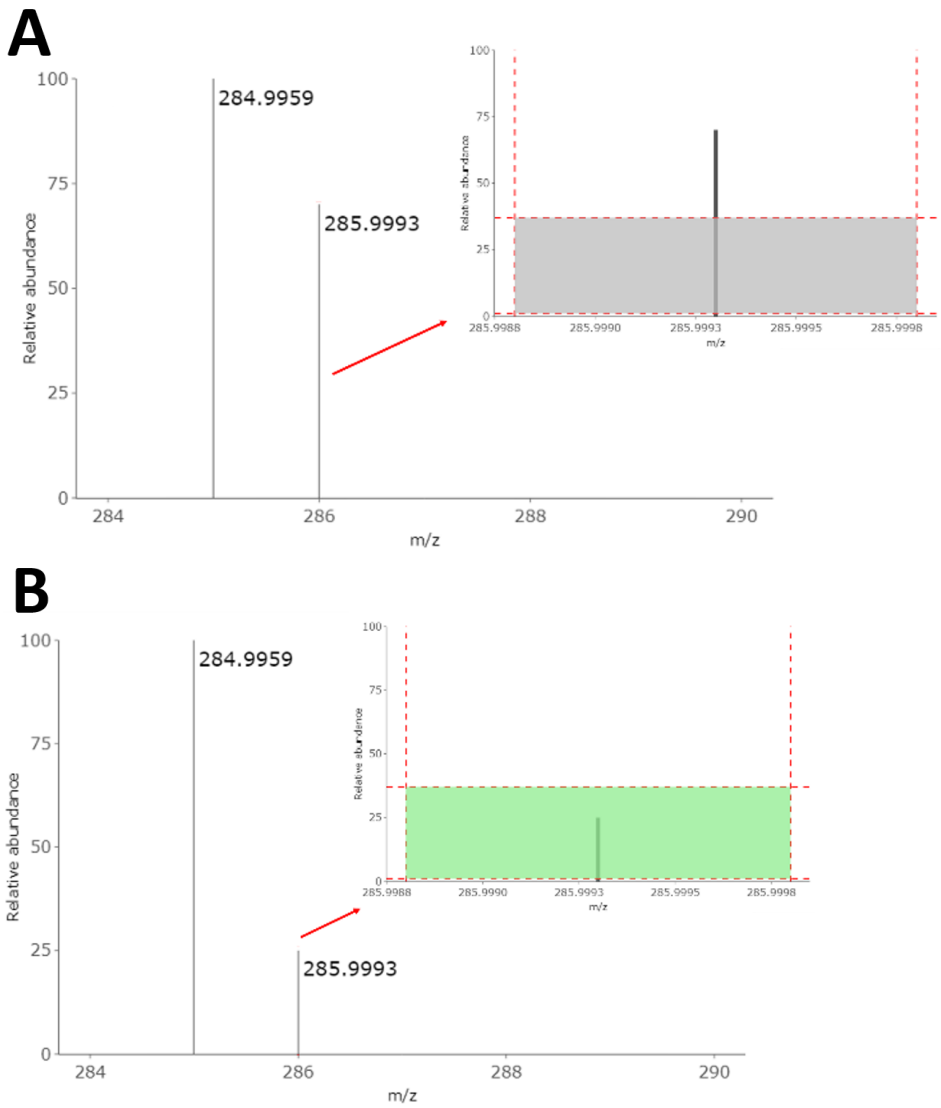


Figure S 1 Theoretical mass spectra showing the usefulness of the relative isotope abundance (RIA) filter when applying an isotope pattern data matrix enrichment (IPDE) to extract species containing the $^{12}C/^{13}C$ isotope pattern. A) Case where two signals are detected with a mass difference of 1.0034 u ±0.0005, but their relative abundances do not match the $^{12}C/^{13}C$ RIA. B) Case where two signals are detected with an acceptable $^{12}C/^{13}C$ RIA, i.e., the abundance ratio of the ^{13}C isotope peak of about 25% is within an estimated interval.

The vertical dotted lines represent the 0.0005 u tolerance window for the *m/z* value and the horizontal lines limit the RIA which maximum value is obtained from the estimated number of carbon atoms as follows.

When applying a $^{12}C/^{13}C$ IPDE, signal pairs with an accurate *m/z* difference of 1.0034 u are selected. Their RIA is then verified based on their *m/z* values. The maximum number of carbon atoms cannot be greater than the value obtained by dividing the lower *m/z* value of each signal pair by 12 (the carbon mass) + 1. This "+1" is added to minimize the error due to the possible contribution of isotopes of other elements. For example, for *m/z* 284.9959, the maximum number of carbon atoms is given by (284.9959/12) +1 = 25. Therefore, the maximum abundance of the ^{13}C isotope peak cannot be higher than the RIA of species containing 25 carbon atoms, which is approximately 25 %.

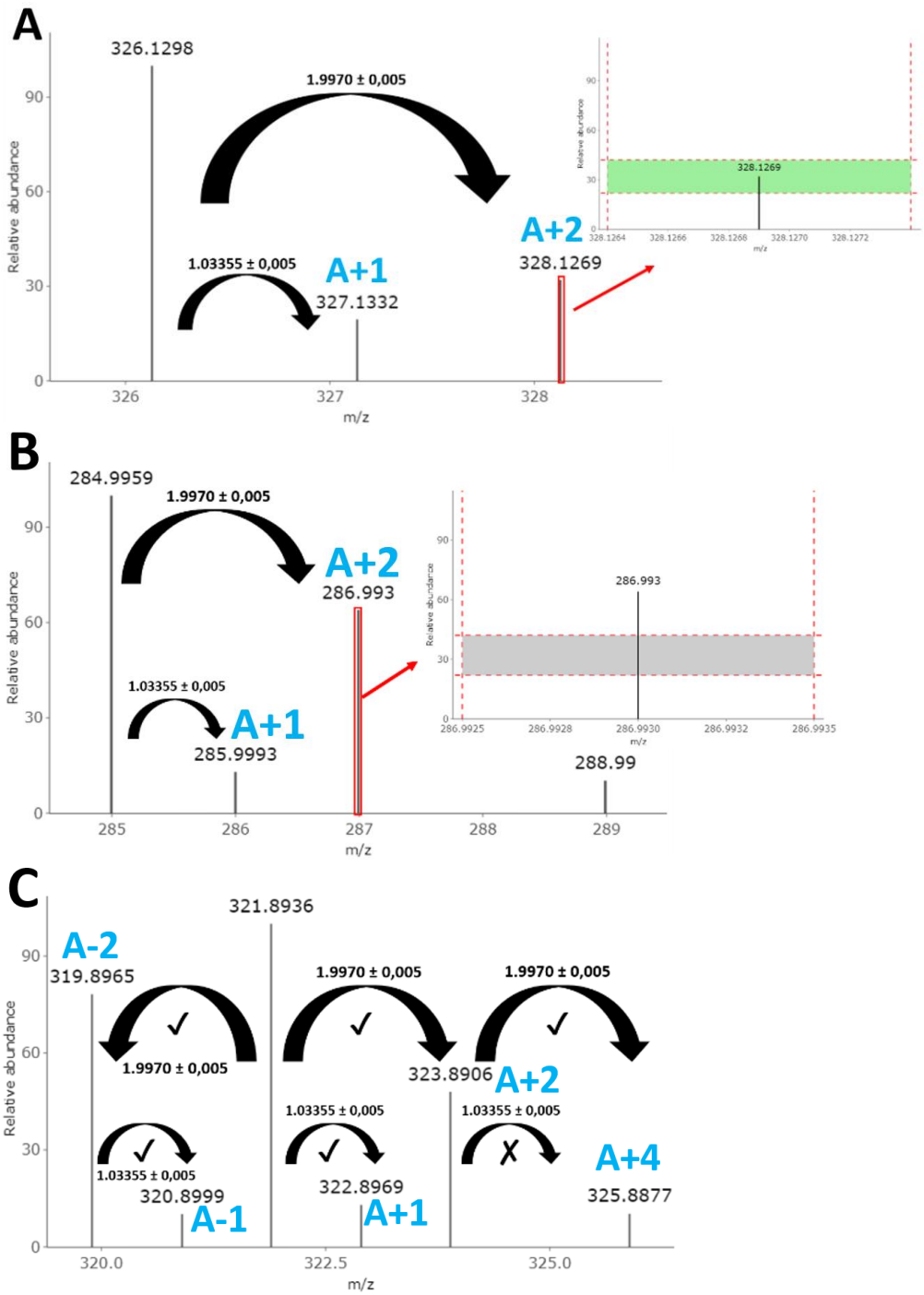


Figure S 2 Theoretical isotope patterns of mono- and multi-chlorinated compounds, such as A) clozapine, B) vinclozolin, and C) 2,3,7,8-tetrachlorodibenzo-p-dioxine (TCDD). The horizontal dotted lines represent the tolerance windows of 0.0005 u for the *m/z* value and the vertical lines correspond to 10 % for the RIA error.This figure illustrates how our developed IPDE algorithm works with three different types of molecules, each having a different number of chlorine atoms. For a molecule containing one chlorine atom (Figure S2A), if a signal is detected at A+2 with the expected RIA, the presence of one chlorine atom is confirmed and, this information is then incorporated into the initial data matrix. An A+2 signal with a mismatched RIA indicates the presence of multiple chlorines in the species considered (Figure S2B). In more complex cases,

where both A+2 and A-2 signals are detected (Figure S2C), an additional step is carried out to avoid annotation errors. This involves searching for all relevant isotopic signals (i.e. A ± 2, A ± 4, A ± 6, etc. for the chlorine isotope pattern, and A ± 1, A ± 3, A ± 5, etc. for the carbon isotope pattern) until no signal is found. The detected signals are listed in an output file for further manual analysis.

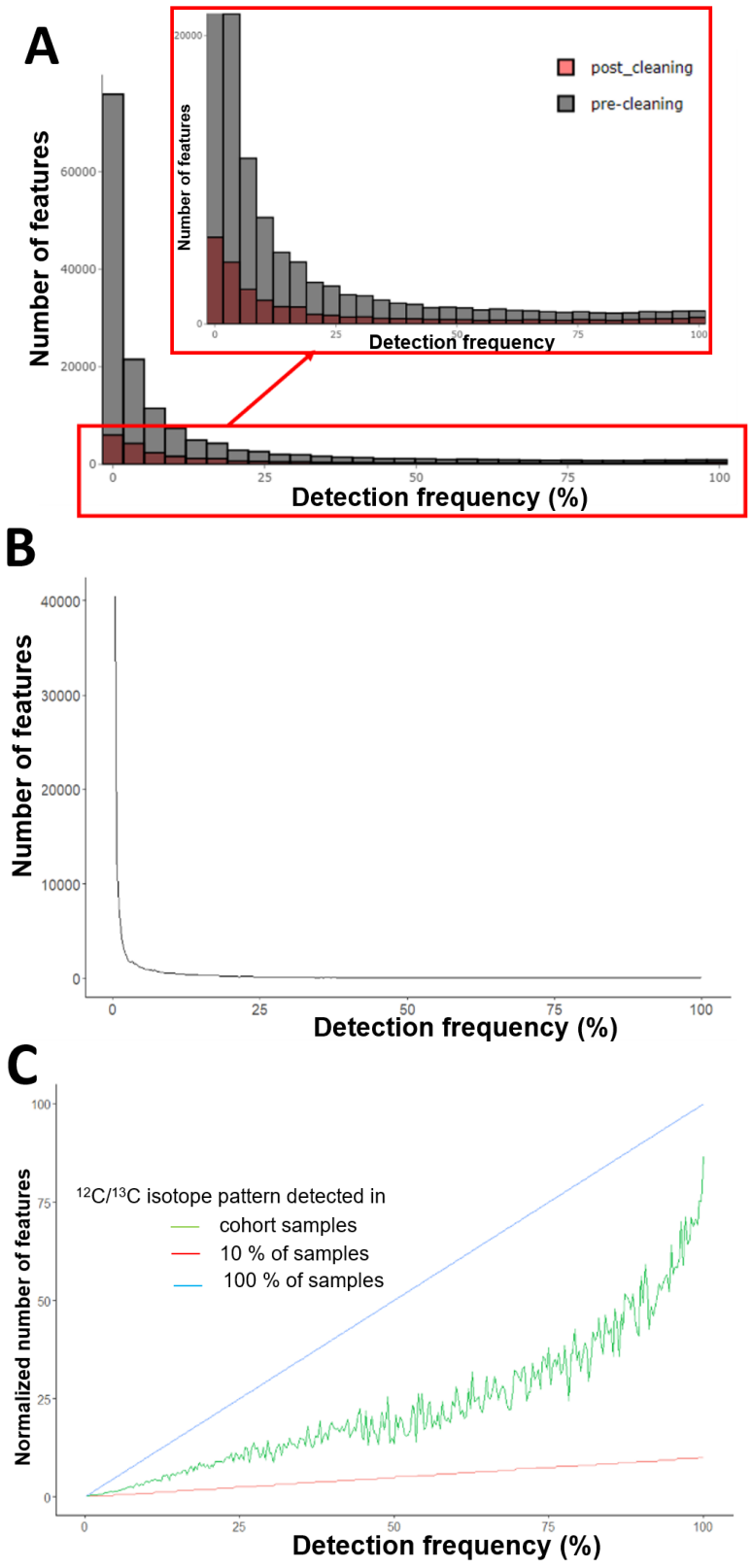


Figure S 3 Distribution of features according to their detection frequency, A) before and after the data cleaning step using the blank filter, an IPDE to extract only features with the $^{12}C/^{13}C$ isotope pattern detected in at least 10 % of all samples (n=308), B) for features removed by the data cleaning step, C) for the $^{12}C/^{13}C$ isotope pattern actually detected (green plot as the mean value of feature number) in samples which is normalized on the y-axis, the blue and red lines

represent the simulation of features detected with a validated $^{12}C/^{13}C$ isotope pattern in all samples (blue line) and in 10% of the total samples (red line).

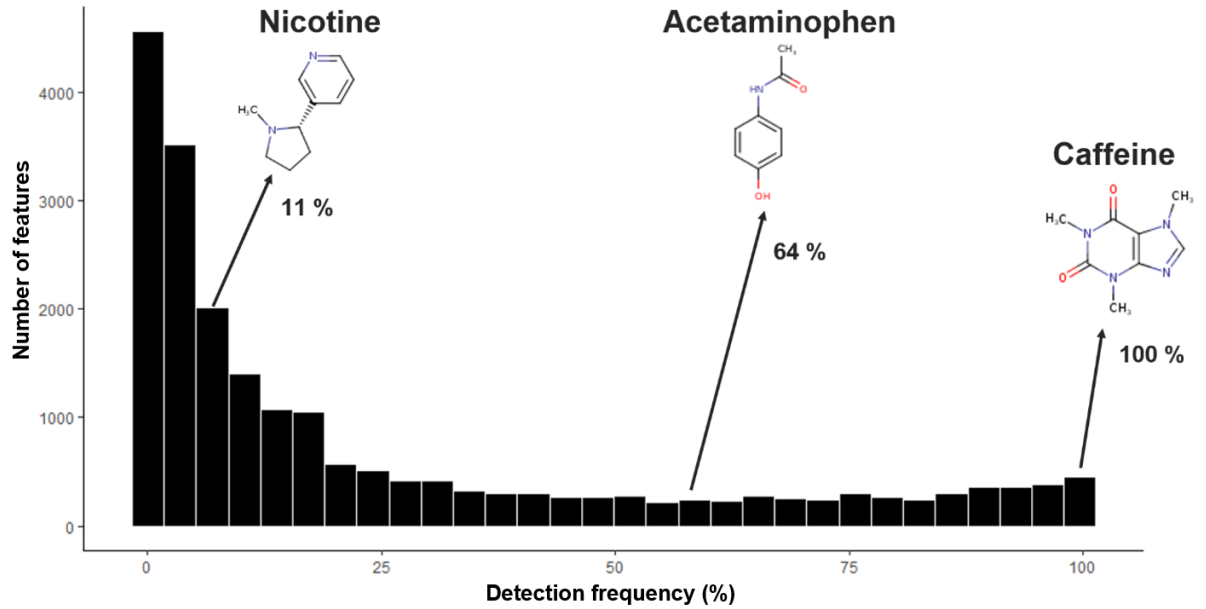


Figure S 4 Distribution of the 25,276 features obtained after the data cleaning step according to their detection frequency. Some features could be annotated as common xenobiotics such as nicotine, acetaminophen and caffeine. Their detection frequency within the studied cohort is indicated.

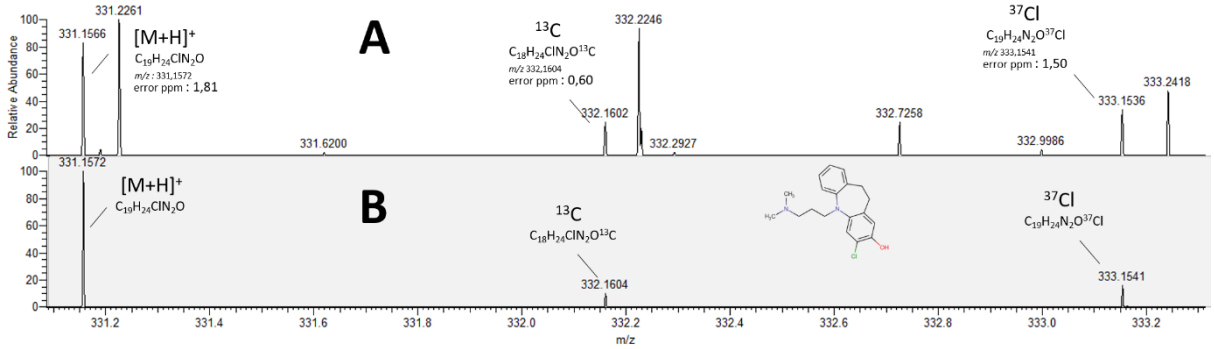


Figure S 5 A) Experimental mass spectrum from the LC-HRMS analysis of a meconium sample, showing the presence of the $[M+H]^+$ *m/z* 331.1566 species, corresponding to a chlorinated compound and, B) Theoretical mass spectrum of protonated 2-hydroxyclomipramin.

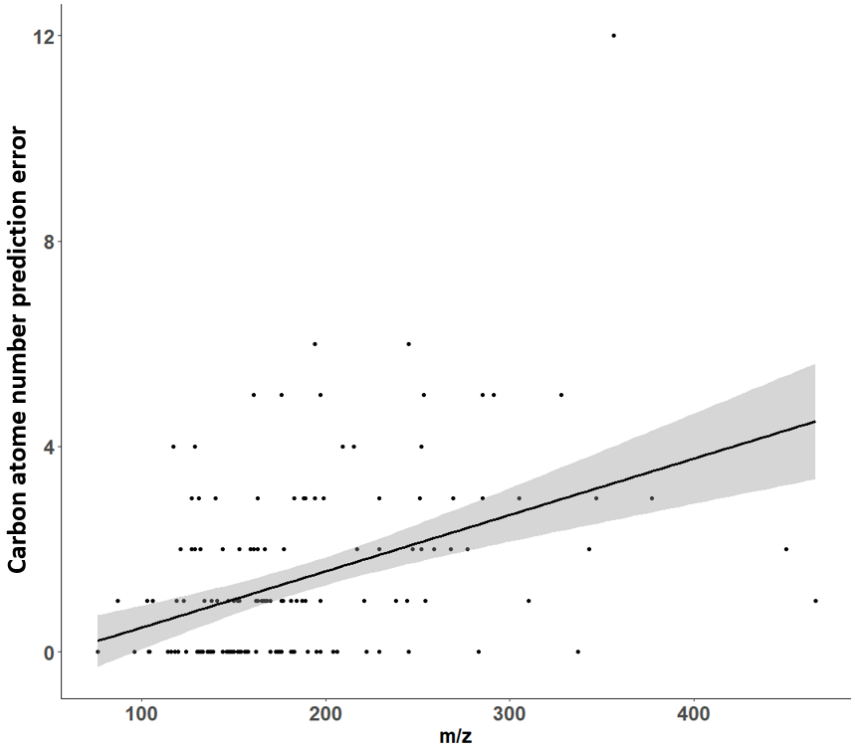


Figure S 6 Error in predicting the number of carbon atoms *vs. m/z* value for the 143 species annotated using our in-house database. Regression analysis showed an increase in carbon number prediction error with the *m/z* values.

Table S 1 Preliminary annotation of 164 species based on their accurate *m/z* values (with a mass tolerance window ± 0.0005 u) and retention time (RT, tolerance window of 10 seconds) using our in-house database.

Compounds	Species	RT (s)	*m/z*	error (ppm)
Trimethylamine-oxide	$[M+H]^+$	46,20	76,0755	2,07
Butanedione	$[M+H]^+$	109,45	87,0439	2,28
Hydroxypyridine	$[M+H]^+$	56,00	96,0442	2,19
Ethylmethylacetic-acid-(2-Methylbutyric-acid)	$[M+H]^+$	93,31	103,0752	1,55
Cadaverin	$[M+H]^+$	40,71	103,1228	1,72
gamma-aminobutyric-acid or its isomers	$[M+H]^+$	48,59	104,0705	1,17
Choline	$[M+H]^+$	45,58	104,1068	2,01
Serine	$[M+H]^+$	48,07	106,0496	2,66
Uracil	$[M+H]^+$	63,90	113,0344	1,49
Creatinine	$[M+H]^+$	47,07	114,0661	0,77
Proline	$[M+H]^+$	49,98	116,0705	0,69
Valine or its isomer	$[M+H]^+$	53,39	118,0861	1,75
Succinic-acid	$[M+H]^+$	73,91	119,0340	-0,51
6-31-Threonine or its isomers	$[M+H]^+$	48,25	120,0654	0,60
Nicotinamide	$[M+H]^+$	58,71	123,0552	0,84
Nicotinic-acid or its isomers	$[M+H]^+$	57,50	124,0393	-0,07
Thymine	$[M+H]^+$	79,97	127,0503	-0,98
Dihydrothymine	$[M+H]^+$	80,98	129,0660	-0,52
Octanal	$[M+H]^+$	5,52	129,1274	-0,08
Pyroglutamic-acid	$[M+H]^+$	65,34	130,0500	-0,65
(+-)-Mevalonolactone	$[M+H]^+$	87,09	131,0704	-1,10
N-Acetylputrescine	$[M+H]^+$	48,57	131,1179	0,06
Agmatine	$[M+H]^+$	42,17	131,1292	-0,52
cis-4-hydroxy-D-proline or its isomers	$[M+H]^+$	48,81	132,0657	-1,34
Creatine	$[M+H]^+$	48,72	132,0767	0,50
Isoleucine or its isomers	$[M+H]^+$	73,16	132,1019	-0,11
Asparagine or its isomer	$[M+H]^+$	48,27	133,0607	0,38
Ornithine	$[M+H]^+$	42,25	133,0971	0,48
L-Aspartic-acid	$[M+H]^+$	48,60	134,0447	0,51
Adenine	$[M+H]^+$	50,78	136,0618	0,17
Methylnicotinamide	$[M+H]^+$	45,71	137,0710	-0,46
Trigonelline or its isomers	$[M+H]^+$	50,52	138,0548	1,26
Aminobenzoic-acid	$[M+H]^+$	147,95	138,0552	-1,30
4-Imidazoleacrylic-acid-(urocanic-acid) or its isomers	$[M+H]^+$	52,58	139,0502	-0,23
6-Hydroxypyridine-3-carboxylic-acid	$[M+H]^+$	84,58	140,0344	-1,17
Methylimidazoleacetic-acid	$[M+H]^+$	56,99	141,0659	0,19
1-Aminocyclohexanecarboxylic-acid	$[M+H]^+$	87,78	144,1020	-0,69
Stachydrine	$[M+H]^+$	50,96	144,1019	0,12
4-Acetamidobutanoic-acid	$[M+H]^+$	81,64	146,0812	-0,12
alpha-Ketoglutaric-acid	$[M+H]^+$	64,99	147,0288	-0,16
Glutamine	$[M+H]^+$	48,47	147,0764	0,23
Lysine	$[M+H]^+$	42,56	147,1128	-0,06
Glutamic-acid or its isomers	$[M+H]^+$	48,60	148,0604	0,96
Mevalonic-acid	$[M+H]^+$	93,49	149,0810	-1,08
Methionine or its isomers	$[M+H]^+$	59,23	150,0585	-1,52
Triethanolamine	$[M+H]^+$	45,99	150,1124	0,66
Acetaminophen-(4-Acetamidophenol)	$[M+H]^+$	164,92	152,0706	0,11
Xanthine	$[M+H]^+$	63,99	153,0407	0,22
Histidine	$[M+H]^+$	43,99	156,0767	0,77
2-5-Furandicarboxylic-acid	$[M+H]^+$	110,57	157,0132	-0,02
Imidazolelactic-acid	$[M+H]^+$	49,12	157,0607	0,70
3-Methylcrotonyl-glycine	$[M+H]^+$	243,43	158,0813	-0,51
Allantoin	$[M+H]^+$	51,37	159,0511	1,29
D-Ala-D-Ala	$[M+H]^+$	47,35	161,0919	0,95
Alpha-D-Aminoadipic-acid	$[M+H]^+$	47,45	162,0760	0,77
Carnitine	$[M+H]^+$	47,00	162,1123	1,07
3-Hydroxy-3-methylglutaric-acid	$[M+H]^+$	77,43	163,0600	0,58
5-Hydroxylysine	$[M+H]^+$	41,91	163,1077	-0,10

Compounds	Species	RT (s)	*m/z*	error (ppm)
N-Acetyl-L-Cystein	$[M+H]^+$	92,44	164,0376	0,20
L-Fucose	$[M+H]^+$	50,52	165,0756	1,24
methylguanine	$[M+H]^+$	57,14	166,0723	-0,16
1-Methylxanthine	$[M+H]^+$	108,03	167,0561	1,72
Pyridoxal or its isomers	$[M+H]^+$	51,64	168,0654	0,86
Uric-acid	$[M+H]^+$	58,37	169,0354	0,95
Norepinephrine	$[M+H]^+$	80,32	170,0811	0,83
Gly-Pro or its isomers	$[M+H]^+$	52,44	173,0918	1,46
N-Acetyl-D-allo-isoleucine	$[M+H]^+$	319,53	174,1125	0,03
N-Acetyl-ornithine	$[M+H]^+$	50,74	175,1076	0,78
Arginine	$[M+H]^+$	44,56	175,1187	1,44
N-Acetyl-L-aspartic-acid	$[M+H]^+$	61,59	176,0552	0,72
Guanidinosuccinic-acid or its isomers	$[M+H]^+$	48,65	176,0664	1,39
Ascorbic-acid or its isomers	$[M+H]^+$	52,58	177,0392	0,97
D-(+)-Gluconic-acid-D-lactone	$[M+H]^+$	51,58	179,0548	1,29
Theobromine	$[M+H]^+$	155,38	181,0719	0,74
Paraxanthine or its isomers	$[M+H]^+$	231,55	181,0719	0,32
Tyrosine or its isomers	$[M+H]^+$	67,03	182,0810	0,92
DL-p-Hydroxyphenyllactic-acid	$[M+H]^+$	245,94	183,0652	-0,04
D-Mannitol or its isomers	$[M+H]^+$	50,50	183,0861	1,10
Pyridoxic-acid	$[M+H]^+$	84,00	184,0602	0,95
L-Alanyl-L-proline	$[M+H]^+$	56,83	187,1072	2,41
N8-Acetylspermidine	$[M+H]^+$	46,00	188,1755	1,33
N-acetyl-L-glutamine	$[M+H]^+$	58,36	189,0866	1,94
N6-Acetyl-L-lysine	$[M+H]^+$	52,40	189,1231	1,83
N6-N6-N6-Trimethyl-L-lysine	$[M+H]^+$	44,10	189,1595	1,67
Kynurenic-acid	$[M+H]^+$	267,94	190,0498	0,38
N-Acetyl-L-glutamic-acid or its isomers	$[M+H]^+$	67,56	190,0708	1,21
Citric-acid or its isomers	$[M+H]^+$	64,52	193,0345	-0,80
Caffeine	$[M+H]^+$	283,71	195,0879	-1,17
1-3-Dimethyluric-acid	$[M+H]^+$	136,02	197,0672	-1,30
3-7-Dimethyluric-acid	$[M+H]^+$	95,90	197,0671	-1,26
1-7-Dimethyluric-acid	$[M+H]^+$	203,39	197,0670	-0,62
Guaiacol-glyceryl-ether	$[M+H]^+$	363,74	199,0967	-1,24
Acetyl-L-carnitin	$[M+H]^+$	55,21	204,1230	-0,05
Xanthurenic-acid	$[M+H]^+$	247,94	206,0451	-1,33
L-Kynurenine	$[M+H]^+$	132,81	209,0923	-0,74
N-alpha-acetyl-L-arginine	$[M+H]^+$	51,49	217,1294	0,25
5-hydroxy-dl-tryptophan	$[M+H]^+$	84,97	221,0921	-0,20
Acetylgalactosamine or its isomers	$[M+H]^+$	51,13	222,0971	0,40
Deoxycytidine	$[M+H]^+$	49,61	228,0978	0,51
9-12-dioxo-dodecanoic-acid	$[M+H]^+$	423,20	229,1435	-0,02
Pro-Leu	$[M+H]^+$	227,86	229,1547	0,13
6-Biopterin	$[M+H]^+$	64,01	238,0933	0,76
L-Cystine	$[M+H]^+$	47,96	241,0310	0,52
Cytidine	$[M+H]^+$	56,18	244,0925	1,05
Uridine or its isomers	$[M+H]^+$	62,92	245,0766	0,73
N-acetyl-DL-tryptophan	$[M+H]^+$	363,45	247,1077	-0,19
Muramic-acid	$[M+H]^+$	49,14	252,1075	1,24
Ala-Tyr-(alanyltyrosine)	$[M+H]^+$	93,13	253,1182	0,42
D-(+)-Neopterin	$[M+H]^+$	52,80	254,0881	1,14
D-Glucosamine-6-phosphate	$[M+H]^+$	48,25	260,0527	0,99
Adenosine	$[M+H]^+$	64,44	268,1037	1,25
Inosine	$[M+H]^+$	73,33	269,0877	1,49
N-Acetyl-L-carnosine	$[M+H]^+$	51,33	269,1240	1,40
L-Saccharopine	$[M+H]^+$	48,61	277,1389	1,63
2-O-Methylinosine	$[M+H]^+$	100,95	283,1035	0,73
Xanthosine	$[M+H]^+$	87,46	285,0827	1,22
Argininosuccinic-acid	$[M+H]^+$	49,40	291,1295	1,36
2-O-Methylguanosine	$[M+H]^+$	92,15	298,1144	0,57
(-)-N-Acetylneuraminic-acid	$[M+H]^+$	52,45	310,1129	1,45
Acetaminophen-glucuronide	$[M+H]^+$	95,46	328,1024	0,77
D-Maltose or its isomer	$[M+H]^+$	50,71	343,1230	1,42

Compounds	Species	RT (s)	*m/z*	error (ppm)
Dehydroisoandrosterone-3-sulfate or its isomers	$[M+H]^+$	480,34	369,1727	0,91
N-Acetyl-D-lactosamine	$[M+H]^+$	51,41	384,1495	1,24
Glycoursodeoxycholic-acid	$[M+H]^+$	482,56	450,3209	1,03
2-1-Glycocholic-acid	$[M+H]^+$	478,80	466,3159	0,77
Caproic-acid	$[M+H]^+$	5,52	117,0910	0,27
Citraconic-acid or its isomers	$[M+H]^+$	96,38	131,0340	-0,87
3-Methylxanthine or its isomer	$[M+H]^+$	91,91	167,0563	0,74
9-9-Thymidine	$[M+H]^+$	111,04	243,0974	0,94
72-173-Stearic-acid	$[M+H]^+$	4,87	285,2785	1,19
6-beta-Hydroxytestosterone	$[M+H]^+$	425,07	305,2108	0,86
Prostaglandin-A1	$[M+H]^+$	499,68	337,2369	1,17
Deoxycortisol or its isomer	$[M+H]^+$	488,27	347,2213	1,03
Riboflavin	$[M+H]^+$	301,25	377,1458	-0,54
5-hydroxymethylfurfural	$[M+H]^+$	125,60	127,0390	-0,25
Hypoxanthine	$[M+H]^+$	60,76	137,0459	-0,56
2-Acetamidophenol	$[M+H]^+$	6,17	152,0707	-0,34
2-4-Quinolinediol	$[M+H]^+$	370,85	162,0549	0,49
Phenylacetaldehyde	$[M+H]^+$	5,52	121,0649	-0,50
Nicotine	$[M+H]^+$	55,68	163,1229	0,92
Cotinine	$[M+H]^+$	63,99	177,1021	0,63
Mandelic-acid	$[M+H]^+$	280,42	153,0548	-1,14
Acetylhistamine	$[M+H]^+$	50,14	154,0974	0,53
Indoxyl-acetate	$[M+H]^+$	324,66	176,0706	-0,27
2-Methylhippuric-acid-(ortho-Methylhippuric-acid)	$[M+H]^+$	332,87	194,0816	-1,87
gamma-Glutamylcysteine	$[M+H]^+$	58,13	251,0694	0,99
Deoxyadenosine	$[M+H]^+$	69,57	252,1088	1,01
AICAR	$[M+H]^+$	60,13	259,1037	0,00
o-Toluic-acid	$[M+H]^+$	301,63	137,0600	-2,01
1-Methylhistidine or its isomers	$[M+H]^+$	45,12	170,0921	1,64
2-Amino-1-phenylethanol	$[M+H]^+$	5,52	138,0915	-1,56
4-Hydroxy-3-methylbenzoic-acid	$[M+H]^+$	355,42	153,0547	-0,56
Tryptamine	$[M+H]^+$	282,54	161,1074	-0,32
Resveratrol	$[M+H]^+$	263,33	229,0861	-0,81
L-Homocystine	$[M+H]^+$	49,24	269,0620	1,33
Spermidine	$[M+H]^+$	44,10	146,1652	-0,27
Val-Pro	$[M+H]^+$	111,22	215,1391	-0,41
Arachidoyl-ethanolamide	$[M+H]^+$	872,27	356,3525	-0,67
D-Glucosamine-6-sulfate	$[M+H]^+$	47,07	260,0432	0,99
Itaconic-acid	$[M+H]^+$	123,31	131,0341	-1,31
4-Methylhippuric-acid-(para-Methylhippuric-acid)	$[M+H]^+$	372,34	194,0814	-1,28
Taurodeoxycholic-acid or its isomers	$[M+H]^+$	496,63	500,3035	0,93
N-Acetyl-Asp-Glu	$[M+H]^+$	68,59	305,0979	-0,13
N1-Acetylspermine	$[M+H]^+$	41,75	245,2334	0,65

Table S 2 Molecular formulae proposed by the MassTools algorithm (without considered IPDE results) for a species detected at *m/z* 331.1566

Molecular formula	**$[M+H]^+$ *m/z***	**Error (ppm)**
$C_8H_{23}N_6O_8$	331,1572	1,8
$C_{10}H_{28}N_4O_4SP$	331,1564	-0,6
$C_{11}H_{24}N_8SP$	331,1577	3,3
$C_{11}H_{28}ClN_4O_3S$	331,1565	-0,3
$C_{11}H_{30}ClN_4OP_2$	331,1578	3,6
$C_{12}H_{25}N_6OP_2$	331,1560	-1,8
$C_{13}H_{25}ClN_6P$	331,1562	-1,2
$C_{16}H_{27}O_5S$	331,1574	2,4
$C_{18}H_{24}N_2O_2P$	331,1570	1,2
$C_{19}H_{24}ClN_2O$	331,1572	1,8

3.3. Conclusion

Le travail présenté dans ce chapitre a démontré l'efficacité de l'approche non ciblée, combinant plusieurs méthodes de nettoyage, de filtration et de sélection des données, pour la recherche, de façon sans *a priori*, de la signature spécifique aux xénobiotiques dans un grand jeu de données produites par l'analyse LC-HRMS non ciblée. L'étape de nettoyage des données utilisée a permis d'éliminer un grand nombre de signaux en ne conservant qu'un sixième du nombre initial de l'ensemble des signaux grâce à la détection des couples d'isotopes ^{12}C et ^{13}C caractéristiques de composés organiques et validés dans au moins 10 % des échantillons où ces signaux sont présents. Des signaux correspondant putativement aux espèces halogénées et aux couples de métabolites conjugués et non conjugués ont pu être détectés dans des données obtenues sur les premières selles des nouveau-nés de la cohorte EDEN. Plus intéressant encore, la présence de signaux caractéristiques de de xénobiotiques courants tels que le paracétamol, la caféine et la nicotine, a pu être mise en évidence grâce à la base de données du laboratoire. De plus, la fréquence de leur détection dans les échantillons analysés est en accord avec celle décrite dans la littérature et avec des métadonnées recueillies au sein de la cohorte EDEN. Ce résultat constitue la preuve d'une exposition très précoce, en début de la vie, aux xénobiotiques.

La stratégie d'approche non ciblée présentée ici semble très prometteuse pour aborder plus largement l'exposome chimique.

Conclusion générale et perspectives

L'exposome, un domaine de recherche émergent, vise à étudier l'ensemble des expositions auxquelles un individu est confronté au cours de sa vie.[6] L'objectif est de pouvoir établir un lien possible ou probable entre ces expositions et le développement de nombreuses maladies chroniques telles que les allergies. Des expositions, en particulier chimiques, qui surviennent pendant des périodes de vulnérabilité comme la fenêtre périnatale (fœtus, petite enfance) durant laquelle l'organisme en développement est plus sensible aux perturbations environnementales, peuvent avoir des conséquences sur la santé ultérieure de l'individu.

Le travail de thèse présenté ici s'est intéressé à l'exposome chimique périnatal. L'objectif était d'identifier les marqueurs d'exposition présents dans des matrices périnatales telles que le méconium et le lait maternel issus de la cohorte mère-enfant de l'étude épidémiologique EDEN.[161] Pour cela, deux stratégies ont été adoptées pour réaliser une fouille des données métabolomiques déjà produites lors de deux précédentes thèses réalisées au laboratoire. La première, présentée dans le chapitre 2, est une approche par "suspect screening" consistant en la recherche automatique ciblant des signaux attendus tandis que la deuxième, décrite dans le chapitre 3, est une approche plus globale dédiée à une recherche plus exhaustive et sans *a priori* de marqueurs d'exposition.

Ces deux approches, ciblée ou sans *a priori*, se révèlent très complémentaires. L'approche par « suspect screening » permet de caractériser rapidement une exposition aux contaminants suspectés. Elle est idéale pour confirmer ou infirmer une hypothèse liée à l'évaluation du risque de la présence de contaminants jugés préoccupants. Toutefois, elle ne permet pas de capturer l'ensemble des expositions chimiques auxquelles un individu est confronté. En revanche, l'approche sans *a priori* offre une vision plus globale et plus exhaustive des expositions que représente l'exposome chimique et, ainsi, permet d'aborder les expositions sans se limiter à des listes de molécules suspectées.

L'application conjointe de ces deux approches, dans le cadre de ma thèse, a permis de mieux appréhender l'exposition périnatale aux contaminants chimiques dans des matrices biologiques complexes. Les résultats obtenus, notamment la détection de marqueurs d'exposition périnatale aux xénobiotiques, témoignent de l'efficacité des deux approches

utilisées pour analyser de grands jeux de données générées par LC-HRMS et ouvrent des perspectives intéressantes pour des études futures.

Cependant, des études supplémentaires seront nécessaires pour valider les résultats obtenus avec ces deux approches. En particulier, l'identification structurale des marqueurs d'exposition détectés représente un défi majeur dans la caractérisation de l'exposome chimique. D'une part, plusieurs candidats ont été détectés par l'approche « suspect screening » à des temps de rétention différents pour une même formule chimique, et il est difficile de limiter des faux-positifs par manque de données sur les temps de rétention. En effet, pour la majorité des molécules suspectées, nous ne disposons pas de composés standards et encore moins de standards de métabolites prédits *in silico*, ce qui empêche la prise en compte des temps de rétention dans la caractérisation de ces marqueurs d'exposition. Une alternative possible serait de pouvoir générer *in silico* des temps de rétention à l'aide d'outils de prédiction fondés sur la structure de composés d'intérêt.[113]Il convient de souligner que des métabolites de référence pourraient être générés *via* des expériences conduites *in vitro* pour les composés standards disponibles, bien que ceci ne soit pas toujours simple à réaliser. D'autre part, l'identification structurale des candidats potentiels nécessite une curation manuelle. Cela impliquerait des analyses structurales conduites par la spectrométrie de masse en tandem (MS/MS). Or, la spectrométrie de masse étant une technique destructive, cette analyse complémentaire ne peut être réalisée sans consommer des échantillons rares de la cohorte qui sont souvent disponibles en quantité limitée. L'évaluation du rapport « bénéfice/risque » dans la conduite expérimentale complémentaire à mettre en œuvre doit alors être prise en compte.

Un autre problème rencontré lors du traitement des données est la non-détection de signaux d'intérêt pourtant présents dans les données brutes. L'utilisation de la méthode d'extraction des signaux directement à partir des spectres de masse a permis de s'affranchir de ces limites liées à la détection et à l'intégration des pics. La poursuite du développement des méthodes d'extraction de signaux d'intérêt pourrait permettre, dans l'avenir, d'augmenter la couverture de l'exposome chimique à partir de ces mêmes données LC-HRMS, en améliorant la détection d'autres marqueurs d'expositions qui n'ont pas pu être révélés à l'heure actuelle par manque de sensibilité de la détection en spectrométrie de masse.

A court terme, l'exposome chimique périnatal de la cohorte EDEN pourra être complété par les données obtenues par la fouille de données complémentaires, notamment par l'application de l'approche sans *a priori* aux données produites en modes positif et négatif sur les échantillons de lait maternel ainsi que sur celles du méconium en mode négatif. Par ailleurs, l'exposome chimique périnatal sera également exploré à partir des données issues d'une autre cohorte mère-enfant, la cohorte ELFE.[19] Cela permettra d'approfondir l'exposome chimique périnatal, notamment en termes d'expositions environnementales qui pourraient être différentes selon la cohorte considérée et, d'évaluer à terme leurs impacts potentiels sur la santé.

Une autre partie de l'exposome chimique, correspondant à sa part réactive caractérisée par « l'adductome », pourrait également être étudiée pour mettre en évidence les produits chimiques et/ou les métabolites pertinents sur le plan toxicologique.[175] La caractérisation de l'adductome, qui résulte des métabolites réactifs susceptibles de se lier de manière covalente à des macromolécules cellulaires, peut révéler des biomarqueurs d'une exposition de long terme tels que les adduits à l'acide désoxyribonucléique (ADN) ou aux protéines, reflétant ainsi une altération de ces macromolécules.

A plus long-terme, il serait intéressant d'utiliser ces données d'exposome chimique périnatal pour documenter les métadonnées de la cohorte étudiée, puis d'évaluer la corrélation globale qui pourrait exister entre ces métadonnées d'exposition et celles liées à la santé des enfants afin d'établir un lien avec le développement ultérieur de maladies chroniques telles que les allergies. Ces travaux s'inscrivent dans une dynamique plus large de recherche sur l'exposome qui permet d'accroître les connaissances sur les interactions entre environnement chimique et santé humaine. Ils ouvrent ainsi la voie à des applications futures dans le domaine de la santé publique, avec pour objectif de prévenir les impacts négatifs des expositions précoces et d'améliorer les conditions de vie des générations futures.

Références bibliographiques

[1] Z. Wang, G.W. Walker, D.C.G. Muir, K. Nagatani-Yoshida, Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories, Environ. Sci. Technol. 54 (2020) 2575–2584. https://doi.org/10.1021/acs.est.9b06379.

[2] B. Geueke, L.V. Parkinson, K.J. Groh, C.D. Kassotis, M.V. Maffini, O.V. Martin, L. Zimmermann, M. Scheringer, J. Muncke, Evidence for widespread human exposure to food contact chemicals, J. Expo. Sci. Environ. Epidemiol. (2024). https://doi.org/10.1038/s41370-024-00718-2.

[3] M. Gascon, M. Casas, E. Morales, D. Valvi, A. Ballesteros-Gómez, N. Luque, S. Rubio, N. Monfort, R. Ventura, D. Martínez, J. Sunyer, M. Vrijheid, Prenatal exposure to bisphenol A and phthalates and childhood respiratory tract infections and allergy, J. Allergy Clin. Immunol. 135 (2015) 370–378. https://doi.org/10.1016/j.jaci.2014.09.030.

[4] Ž. Tkalec, G. Codling, J.S. Tratnik, D. Mazej, J. Klánová, M. Horvat, T. Kosjek, Suspect and non-targeted screening-based human biomonitoring identified 74 biomarkers of exposure in urine of Slovenian children, Environ. Pollut. 313 (2022) 120091. https://doi.org/10.1016/j.envpol.2022.120091.

[5] X. Shen, H. Yan, C. Wang, P. Gao, C.H. Johnson, M.P. Snyder, TidyMass an object-oriented reproducible analysis framework for LC–MS data, Nat. Commun. 13 (2022) 4365. https://doi.org/10.1038/s41467-022-32155-w.

[6] C.P. Wild, Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology, Cancer Epidemiol. Biomarkers Prev. 14 (2005) 1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456.

[7] C.P. Wild, The exposome: from concept to utility, Int. J. Epidemiol. 41 (2012) 24–32. https://doi.org/10.1093/ije/dyr236.

[8] M. Vrijheid, The exposome: a new paradigm to study the impact of environment on health, Thorax 69 (2014) 876–878. https://doi.org/10.1136/thoraxjnl-2013-204949.

[9] G.W. Miller, D.P. Jones, The Nature of Nurture: Refining the Definition of the Exposome, Toxicol. Sci. 137 (2014) 1–2. https://doi.org/10.1093/toxsci/kft251.

[10] J. Higginson, Distribution of different patterns of cancer, Isr. J. Med. Sci. 4 (1968) 457–468.

[11] K. Czene, P. Lichtenstein, K. Hemminki, Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database, Int. J. Cancer 99 (2002) 260–266. https://doi.org/10.1002/ijc.10332.

[12] D.J.P. Barker, C. Osmond, INFANT MORTALITY, CHILDHOOD NUTRITION, AND ISCHAEMIC HEART DISEASE IN ENGLAND AND WALES, The Lancet 327 (1986) 1077–1081. https://doi.org/10.1016/S0140-6736(86)91340-1.

[13] M.-A. Charles, C. Delpierre, B. Bréant, Le concept des origines développementales de la santé - Évolution sur trois décennies, médecine/sciences 32 (2016) 15–20. https://doi.org/10.1051/medsci/20163201004.

[14] U. Simeoni, J.-B. Armengaud, B. Siddeek, J.-F. Tolsa, Perinatal Origins of Adult Disease, Neonatology 113 (2018) 393–399. https://doi.org/10.1159/000487618.

[15] J.R. Wozniak, E.P. Riley, M.E. Charness, Clinical presentation, diagnosis, and management of fetal alcohol spectrum disorder, Lancet Neurol. 18 (2019) 760–770. https://doi.org/10.1016/S1474-4422(19)30150-4.

[16] J.E. Bruin, H.C. Gerstein, A.C. Holloway, Long-Term Consequences of Fetal and Neonatal Nicotine Exposure: A Critical Review, Toxicol. Sci. 116 (2010) 364–374. https://doi.org/10.1093/toxsci/kfq103.

[17] B. Heude, A. Forhan, R. Slama, L. Douhaud, S. Bedel, M.-J. Saurel-Cubizolles, R. Hankard, O. Thiebaugeorges, M. De Agostini, I. Annesi-Maesano, M. Kaminski, M.-A. Charles, EDEN mother-child cohort study group, Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development, Int. J. Epidemiol. 45 (2016) 353–363. https://doi.org/10.1093/ije/dyv151.

[18] C. Warembourg, C. Monfort, N. Costet, F. Rouget, F. Pelé, R. Garlantézec, S. Cordier, C. Chevrier, Cohort Profile: The PELAGIE mother-child cohort, Int. J. Epidemiol. 53 (2024) dyae064. https://doi.org/10.1093/ije/dyae064.

[19] A. Oleko, F. Betsou, H. Sarter, C. Gerdil, I. Desbois, M.A. Charles, H. Leridon, S. Vandentorren, A Pilot Study of the ELFE Longitudinal Cohort: Feasibility and Preliminary Evaluation of Biological Collection, Biopreservation Biobanking 9 (2011) 223. https://doi.org/10.1089/bio.2010.0032.

[20] L. Maitre, J. de Bont, M. Casas, O. Robinson, G.M. Aasvang, L. Agier, S. Andrušaitytė, F. Ballester, X. Basagaña, E. Borràs, C. Brochot, M. Bustamante, A. Carracedo, M. de Castro, A. Dedele, D. Donaire-Gonzalez, X. Estivill, J. Evandt, S. Fossati, L. Giorgis-Allemand, J. R Gonzalez, B. Granum, R. Grazuleviciene, K. Bjerve Gützkow, L. Småstuen Haug, C. Hernandez-Ferrer, B. Heude, J. Ibarluzea, J. Julvez, M. Karachaliou, H.C. Keun, N. Hjertager Krog, C.-H.E. Lau, V. Leventakou, S. Lyon-Caen, C. Manzano, D. Mason, R. McEachan, H.M. Meltzer, I. Petraviciene, J. Quentin, T. Roumeliotaki, E. Sabido, P.-J. Saulnier, A.P. Siskos, V. Siroux, J. Sunyer, I. Tamayo, J. Urquiza, M. Vafeiadi, D. van Gent, M. Vives-Usano, D. Waiblinger, C. Warembourg, L. Chatzi, M. Coen, P. van den Hazel, M.J. Nieuwenhuijsen, R. Slama, C. Thomsen, J. Wright, M. Vrijheid, Human Early Life Exposome (HELIX) study: a European population-based exposome cohort, BMJ Open 8 (2018) e021311. https://doi.org/10.1136/bmjopen-2017-021311.

[21] P. Vineis, M. Chadeau-Hyam, H. Gmuender, J. Gulliver, Z. Herceg, J. Kleinjans, M. Kogevinas, S. Kyrtopoulos, M. Nieuwenhuijsen, D.H. Phillips, N. Probst-Hensch, A. Scalbert, R. Vermeulen, C.P. Wild, T.Expo. Consortium, The exposome in practice: Design of the EXPOsOMICS project, Int. J. Hyg. Environ. Health 220 (2017) 142. https://doi.org/10.1016/j.ijheh.2016.08.001.

[22] N. Li, R. Friedrich, C.N. Maesano, E. Medda, S. Brescianini, M.A. Stazi, C.E. Sabel, D. Sarigiannis, I. Annesi-Maesano, Lifelong exposure to multiple stressors through different environmental pathways for European populations, Environ. Res. 179 (2019) 108744. https://doi.org/10.1016/j.envres.2019.108744.

[23] HMB4EU, HBM4EU (n.d.). https://www.hbm4eu.eu/ (accessed November 4, 2024).

[24] P. Marx-Stoelting, G. Rivière, M. Luijten, K. Aiello-Holden, N. Bandow, K. Baken, A. Cañas, A. Castano, S. Denys, C. Fillol, M. Herzler, I. Iavicoli, S. Karakitsios, J. Klanova, M. Kolossa-Gehring, A. Koutsodimou, J.L. Vicente, I. Lynch, S. Namorado, S. Norager, A. Pittman, S. Rotter, D. Sarigiannis, M.J. Silva, J. Theunis, T. Tralau, M. Uhl, J. van Klaveren, L. Wendt-Rasch, E. Westerholm, C. Rousselle, P. Sanders, A walk in the PARC: developing and implementing 21st century chemical risk assessment in Europe, Arch. Toxicol. 97 (2023) 893–908. https://doi.org/10.1007/s00204-022-03435-7.

[25] C. Guivarch, A.H. Cissé, M.-A. Charles, B. Heude, B. de Lauzon-Guillain, Parental feeding practices as potential moderating or mediating factors in the associations between children's early and later growth, Int. J. Obes. 2005 47 (2023) 190–196. https://doi.org/10.1038/s41366-023-01255-y.

[26] K.-A. Kallas, K. Marr, S. Moirangthem, B. Heude, M. Koehl, J. van der Waerden, N. Downes, Maternal Mental Health Care Matters: The Impact of Prenatal Depressive and Anxious Symptoms on Child Emotional and Behavioural Trajectories in the French EDEN Cohort, J. Clin. Med. 12 (2023) 1120. https://doi.org/10.3390/jcm12031120.

[27] J.Y. Bernard, M. De Agostini, A. Forhan, T. Alfaiate, M. Bonet, V. Champion, M. Kaminski, B. de Lauzon-Guillain, M.-A. Charles, B. Heude, Breastfeeding Duration and Cognitive Development at 2 and 3 Years of Age in the EDEN Mother–Child Cohort, J. Pediatr. 163 (2013) 36-42.e1. https://doi.org/10.1016/j.jpeds.2012.11.090.

[28] C. Cartier, C. Warembourg, G. Le Maner-Idrissi, A. Lacroix, F. Rouget, C. Monfort, G. Limon, G. Durand, D. Saint-Amour, S. Cordier, C. Chevrier, 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. 124 (2016) 674–680. https://doi.org/10.1289/ehp.1409472.

[29] J.-F. Viel, C. Warembourg, G. Le Maner-Idrissi, A. Lacroix, G. Limon, F. Rouget, C. Monfort, G. Durand, S. Cordier, C. Chevrier, Pyrethroid insecticide exposure and cognitive developmental disabilities in children: The PELAGIE mother-child cohort, Environ. Int. 82 (2015) 69–75. https://doi.org/10.1016/j.envint.2015.05.009.

[30] N. Costet, R. Garlantézec, C. Monfort, F. Rouget, B. Gagnière, C. Chevrier, S. Cordier, Environmental and Urinary Markers of Prenatal Exposure to Drinking Water Disinfection By-Products, Fetal Growth, and Duration of Gestation in the PELAGIE Birth Cohort (Brittany, France, 2002–2006), Am. J. Epidemiol. 175 (2012) 263–275. https://doi.org/10.1093/aje/kwr419.

[31] SPF, Imprégnation des femmes enceintes par les polluants de l'environnement en France en 2011 - Tome 3 : synthèse et conclusions, (n.d.). https://www.santepubliquefrance.fr/notices/impregnation-des-femmes-enceintes-par-les-polluants-de-l-environnement-en-france-en-2011-tome-3-synthese-et-conclusions (accessed September 27, 2024).

[32] C. Vernet, I. Pin, L. Giorgis-Allemand, C. Philippat, M. Benmerad, J. Quentin, A.M. Calafat, X. Ye, I. Annesi-Maesano, V. Siroux, R. Slama, EDEN Mother–Child Cohort Study Group, In Utero Exposure to Select Phenols and Phthalates and Respiratory Health in Five-Year-Old Boys: A Prospective Study, Environ. Health Perspect. 125 (2017) 097006. https://doi.org/10.1289/EHP1015.

[33] C. Philippat, B. Heude, J. Botton, N. Alfaidy, A.M. Calafat, R. Slama, EDEN Mother–Child Cohort Study Group, Prenatal Exposure to Select Phthalates and Phenols and Associations with Fetal and Placental Weight among Male Births in the EDEN Cohort (France), Environ. Health Perspect. 127 (2019) 17002. https://doi.org/10.1289/EHP3523.

[34] N. Stratakis, D. V. Conti, R. Jin, K. Margetaki, D. Valvi, A.P. Siskos, L. Maitre, E. Garcia, N. Varo, Y. Zhao, T. Roumeliotaki, M. Vafeiadi, J. Urquiza, S. Fernández-Barrés, B. Heude, X. Basagana, M. Casas, S. Fossati, R. Gražulevičienė, S. Andrušaitytė, K. Uppal, R.R.C. McEachan, E. Papadopoulou, O. Robinson, L.S. Haug, J. Wright, M.B. Vos, H.C. Keun, M. Vrijheid, K.T. Berhane, R. McConnell, L. Chatzi, Prenatal Exposure to Perfluoroalkyl Substances Associated With Increased Susceptibility to Liver Injury in Children, Hepatology 72 (2020) 1758. https://doi.org/10.1002/hep.31483.

[35] C.M. Villanueva, A. Espinosa, E. Gracia-Lavedan, J. Vlaanderen, R. Vermeulen, A.J. Molina, P. Amiano, I. Gómez-Acebo, G. Castaño-Vinyals, P. Vineis, M. Kogevinas, Exposure to widespread drinking water chemicals, blood inflammation markers, and colorectal cancer, Environ. Int. 157 (2021) 106873. https://doi.org/10.1016/j.envint.2021.106873.

[36] J.O. Grimalt, M. Torrent, J. Sunyer, The influence of organochlorine compound exposure on the physiological development of children, Med. Balear (2014) 25–36. https://doi.org/10.3306/MEDICINABALEAR.29.03.25.

[37] E. Croom, Chapter Three - Metabolism of Xenobiotics of Human Environments, in: E. Hodgson (Ed.), Prog. Mol. Biol. Transl. Sci., Academic Press, 2012: pp. 31–88. https://doi.org/10.1016/B978-0-12-415813-9.00003-9.

[38] M. Nakajima, T. Yokoi, Interindividual Variability in Nicotine Metabolism: C-Oxidation and Glucuronidation, Drug Metab. Pharmacokinet. 20 (2005) 227–235. https://doi.org/10.2133/dmpk.20.227.

[39] R.G. Arora, H. Frölén, Interference of Mycotoxins with Prenatal Development of the Mouse, Acta Vet. Scand. 22 (1981) 535–552. https://doi.org/10.1186/BF03548678.

[40] L.-E. Appelgren, R.G. Arora, Distribution of 14C-labelled ochratoxin a in pregnant mice, Food Chem. Toxicol. 21 (1983) 563–568. https://doi.org/10.1016/0278-6915(83)90141-2.

[41] M.R. Syme, J.W. Paxton, J.A. Keelan, Drug Transfer and Metabolism by the Human Placenta, Clin. Pharmacokinet. 43 (2004) 487–514. https://doi.org/10.2165/00003088-200443080-00001.

[42] D. Zalko, A.M. Soto, L. Dolo, C. Dorio, E. Rathahao, L. Debrauwer, R. Faure, J.-P. Cravedi, Biotransformations of bisphenol A in a mammalian model: answers and new questions raised by low-dose metabolic fate studies in pregnant CD1 mice., Environ. Health Perspect. 111 (2003) 309–319. https://doi.org/10.1289/ehp.5603.

[43] D.A. Devault, S. Karolak, Y. Lévi, N.I. Rousis, E. Zuccato, S. Castiglioni, Exposure of an urban population to pesticides assessed by wastewater-based epidemiology in a Caribbean island, Sci. Total Environ. 644 (2018) 129–136. https://doi.org/10.1016/j.scitotenv.2018.06.250.

[44] T. Traoré, A. Forhan, V. Sirot, M. Kadawathagedara, B. Heude, M. Hulin, B. de Lauzon-Guillain, J. Botton, M.A. Charles, A. Crépet, To which mixtures are French pregnant women mainly exposed? A combination of the second French total diet study with the EDEN and ELFE cohort studies, Food Chem. Toxicol. 111 (2018) 310–328. https://doi.org/10.1016/j.fct.2017.11.016.

[45] K.E. Manz, A. Feerick, J.M. Braun, Y.-L. Feng, A. Hall, J. Koelmel, C. Manzano, S.R. Newton, K.D. Pennell, B.J. Place, K.J. Godri Pollitt, C. Prasse, J.A. Young, Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome, J. Expo. Sci. Environ. Epidemiol. 33 (2023) 524–536. https://doi.org/10.1038/s41370-023-00574-6.

[46] S.J. Hird, B.P.-Y. Lau, R. Schuhmacher, R. Krska, Liquid chromatography-mass spectrometry for the determination of chemical contaminants in food, TrAC Trends Anal. Chem. 59 (2014) 59–72. https://doi.org/10.1016/j.trac.2014.04.005.

[47] J.R. Sobus, J.F. Wambaugh, K.K. Isaacs, A.J. Williams, A.D. McEachran, A.M. Richard, C.M. Grulke, E.M. Ulrich, J.E. Rager, M.J. Strynar, S.R. Newton, Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA, J. Expo. Sci. Environ. Epidemiol. 28 (2018) 411–426. https://doi.org/10.1038/s41370-017-0012-y.

[48] A. David, J. Chaker, L. Multigner, V. Bessonneau, Exposome chimique et approches « non ciblées » - Un changement de paradigme pour évaluer l'exposition des populations aux contaminants chimiques, médecine/sciences 37 (2021) 895–901. https://doi.org/10.1051/medsci/2021088.

[49] N. Núñez, J. Saurina, O. Núñez, Liquid Chromatography–High-Resolution Mass Spectrometry (LC-HRMS) Fingerprinting and Chemometrics for Coffee Classification

and Authentication, Molecules 29 (2024) 232. https://doi.org/10.3390/molecules29010232.

[50] H. Li, M. Riva, P. Rantala, L. Heikkinen, K. Daellenbach, J.E. Krechmer, P.-M. Flaud, D. Worsnop, M. Kulmala, E. Villenave, E. Perraudin, M. Ehn, F. Bianchi, Terpenes and their oxidation products in the French Landes forest: insights from Vocus PTR-TOF measurements, Atmospheric Chem. Phys. 20 (2020) 1941–1959. https://doi.org/10.5194/acp-20-1941-2020.

[51] I.A. Rather, W.Y. Koh, W.K. Paek, J. Lim, The Sources of Chemical Contaminants in Food and Their Health Implications, Front. Pharmacol. 8 (2017) 830. https://doi.org/10.3389/fphar.2017.00830.

[52] R. Laumbach, Q. Meng, H. Kipen, What can individuals do to reduce personal health risks from air pollution?, J. Thorac. Dis. 7 (2015) 96–107. https://doi.org/10.3978/j.issn.2072-1439.2014.12.21.

[53] E. Papadopoulou, L.S. Haug, A.K. Sakhi, S. Andrusaityte, X. Basagaña, A.L. Brantsaeter, M. Casas, S. Fernández-Barrés, R. Grazuleviciene, H.K. Knutsen, L. Maitre, H.M. Meltzer, R.R.C. McEachan, T. Roumeliotaki, R. Slama, M. Vafeiadi, J. Wright, M. Vrijheid, C. Thomsen, L. Chatzi, Diet as a Source of Exposure to Environmental Contaminants for Pregnant Women and Children from Six European Countries, Environ. Health Perspect. 127 (2019) 107005. https://doi.org/10.1289/EHP5324.

[54] Y.-C. Chen, J.-F. Hsu, C.-W. Chang, S.-W. Li, Y.-C. Yang, M.-R. Chao, H.-J.C. Chen, P.-C. Liao, Connecting chemical exposome to human health using high-resolution mass spectrometry-based biomonitoring: Recent advances and future perspectives, Mass Spectrom. Rev. 42 (2023) 2466–2486. https://doi.org/10.1002/mas.21805.

[55] E.M. Ostrea, D.M. Bielawski, N.C. Posecion, M. Corrion, E. Villanueva-Uy, R.C. Bernardo, Y. Jin, J.J. Janisse, J.W. Ager, 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. 109 (2009) 116–122. https://doi.org/10.1016/j.envres.2008.09.004.

[56] C. Dereumeaux, A. Saoudi, M. Pecheux, B. Berat, P. de Crouy-Chanel, C. Zaros, S. Brunel, C. Delamaire, A. le Tertre, A. Lefranc, S. Vandentorren, L. Guldner, Biomarkers of exposure to environmental contaminants in French pregnant women from the Elfe cohort in 2011, Environ. Int. 97 (2016) 56–67. https://doi.org/10.1016/j.envint.2016.10.013.

[57] W. Föllmann, N. Ali, M. Blaszkewicz, G.H. Degen, Biomonitoring of Mycotoxins in Urine: Pilot Study in Mill Workers, J. Toxicol. Environ. Health A 79 (2016) 1015–1025. https://doi.org/10.1080/15287394.2016.1219540.

[58] K. Inoue, A. Yamaguchi, M. Wada, Y. Yoshimura, T. Makino, H. Nakazaw, Quantitative detection of bisphenol A and bisphenol A diglycidyl ether metabolites in human plasma by liquid chromatography-electrospray mass spectrometry, J. Chromatogr. B. Biomed. Sci. App. 765 (2001) 121–126. https://doi.org/10.1016/s0378-4347(01)00393-0.

[59] K. Inoue, M. Kawaguchi, Y. Funakoshi, H. Nakazawa, Size-exclusion flow extraction of bisphenol A in human urine for liquid chromatography-mass spectrometry, J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci. 798 (2003) 17–23. https://doi.org/10.1016/j.jchromb.2003.08.042.

[60] F. Artacho-Cordón, M. Fernández-Rodríguez, C. Garde, E. Salamanca, L.M. Iribarne-Durán, P. Torné, J. Expósito, L. Papay-Ramírez, M.F. Fernández, N. Olea, J.P. Arrebola, Serum and adipose tissue as matrices for assessment of exposure to persistent organic pollutants in breast cancer patients, Environ. Res. 142 (2015) 633–643. https://doi.org/10.1016/j.envres.2015.08.020.

[61] T. Tian, F. Liu, Y. Fu, J. Ao, S. Lin, Q. Cheng, K. Kayim, F. Kong, L. Wang, X. Long, Y. Wang, J. Qiao, Environmental exposure patterns to 94 current-use pesticides in women of reproductive age who are preparing for pregnancy, Sci. Total Environ. 948 (2024) 174624. https://doi.org/10.1016/j.scitotenv.2024.174624.

[62] E.M. Hardy, C. Dereumeaux, L. Guldner, O. Briand, S. Vandentorren, A. Oleko, C. Zaros, B.M.R. Appenzeller, Hair versus urine for the biomonitoring of pesticide exposure: Results from a pilot cohort study on pregnant women, Environ. Int. 152 (2021) 106481. https://doi.org/10.1016/j.envint.2021.106481.

[63] R. Rodríguez-Gómez, J. Martín, A. Zafra-Gómez, E. Alonso, J.L. Vílchez, A. Navalón, Biomonitoring of 21 endocrine disrupting chemicals in human hair samples using ultra-high performance liquid chromatography–tandem mass spectrometry, Chemosphere 168 (2017) 676–684. https://doi.org/10.1016/j.chemosphere.2016.11.008.

[64] L. Wang, A.G. Asimakopoulos, K. Kannan, Accumulation of 19 environmental phenolic and xenobiotic heterocyclic aromatic compounds in human adipose tissue, Environ. Int. 78 (2015) 45–50. https://doi.org/10.1016/j.envint.2015.02.015.

[65] S. Sousa, M.L. Maia, C. Delerue-Matos, C. Calhau, V.F. Domingues, The role of adipose tissue analysis on Environmental Pollutants Biomonitoring in women: The European scenario, Sci. Total Environ. 806 (2022) 150922. https://doi.org/10.1016/j.scitotenv.2021.150922.

[66] K. Croes, A. Colles, G. Koppen, E. Govarts, L. Bruckers, E. Van de Mieroop, V. Nelen, A. Covaci, A.C. Dirtu, C. Thomsen, L.S. Haug, G. Becher, M. Mampaey, G. Schoeters, N. Van Larebeke, W. Baeyens, Persistent organic pollutants (POPs) in human milk: A biomonitoring study in rural areas of Flanders (Belgium), Chemosphere 89 (2012) 988–994. https://doi.org/10.1016/j.chemosphere.2012.06.058.

[67] A. Goutelle, J. Viseur, K.Z. Boudjeltia, V. Nuyens, E. Cavatorta, P.V. Antwerpen, Y. Maréchal, Mass spectrometry analysis of environmental pollutants in breast and artificial milk for newborns, Heliyon 10 (2024). https://doi.org/10.1016/j.heliyon.2024.e32350.

[68] B. Warth, D. Braun, C.N. Ezekiel, P.C. Turner, G.H. Degen, D. Marko, Biomonitoring of Mycotoxins in Human Breast Milk: Current State and Future Perspectives, Chem. Res. Toxicol. 29 (2016) 1087–1097. https://doi.org/10.1021/acs.chemrestox.6b00125.

[69] G. Sousa, C. Delerue-Matos, X. Wang, F. Rodrigues, M. Oliveira, Potential of Saliva for Biomonitoring of Occupational Exposure: Collection of Evidence from the Literature, in: P.M. Arezes, J.S. Baptista, R.B. Melo, J. Castelo Branco, P. Carneiro, A. Colim, N. Costa, S. Costa, J. Duarte, J.C. Guedes, G. Perestrelo (Eds.), Occup. Environ. Saf. Health IV, Springer International Publishing, Cham, 2023: pp. 587–598. https://doi.org/10.1007/978-3-031-12547-8_47.

[70] R. Cassoulet, L. Haroune, N. Abdelouahab, V. Gillet, A.A. Baccarelli, H. Cabana, L. Takser, J.-P. Bellenger, Monitoring of prenatal exposure to organic and inorganic contaminants using meconium from an Eastern Canada cohort, Environ. Res. 171 (2019) 44–51. https://doi.org/10.1016/j.envres.2018.12.044.

[71] E.M. Ostrea, D.M. Bielawski, N.C. Posecion, M. Corrion, E. Villanueva-Uy, Y. Jin, J.J. Janisse, J.W. Ager, A comparison of infant hair, cord blood and meconium analysis to detect fetal exposure to environmental pesticides, Environ. Res. 106 (2008) 277–283. https://doi.org/10.1016/j.envres.2007.08.014.

[72] E.M. Ostrea, O. Matias, C. Keane, E. Mac, R. Utarnachitt, A. Ostrea, M. Mazhar, Spectrum of gestational exposure to illicit drugs and other xenobiotic agents in newborn infants by meconium analysis, J. Pediatr. 133 (1998) 513–515. https://doi.org/10.1016/s0022-3476(98)70059-9.

[73] M. Musatadi, A. Andrés-Maguregi, F. De Angelis, A. Prieto, E. Anakabe, M. Olivares, N. Etxebarria, O. Zuloaga, The role of sample preparation in suspect and non-target screening for exposome analysis using human urine, Chemosphere 339 (2023) 139690. https://doi.org/10.1016/j.chemosphere.2023.139690.

[74] O. Golge, Validation of Quick Polar Pesticides (QuPPe) Method for Determination of Eight Polar Pesticides in Cherries by LC-MS/MS, Food Anal. Methods 14 (2021) 1432–1437. https://doi.org/10.1007/s12161-021-01966-w.

[75] H. Guo, H. Wang, J. Zheng, W. Liu, J. Zhong, Q. Zhao, Sensitive and rapid determination of glyphosate, glufosinate, bialaphos and metabolites by UPLC–MS/MS using a modified Quick Polar Pesticides Extraction method, Forensic Sci. Int. 283 (2018) 111–117. https://doi.org/10.1016/j.forsciint.2017.12.016.

[76] T. Diallo, Y. Makni, A. Lerebours, H. Thomas, T. Guérin, J. Parinet, Wide-scope screening of multi-class contaminants in seafood using a novel sample preparation (QuEChUP) procedure coupled with UHPLC-Q-TOF-MS: Application for semi-quantitation of real seafood samples, Food Chem. 426 (2023) 136572. https://doi.org/10.1016/j.foodchem.2023.136572.

[77] P. Schippers, S. Rasheed, Y.M. Park, T. Risch, L. Wagmann, S. Hemmer, S.K. Manier, R. Müller, J. Herrmann, M.R. Meyer, Evaluation of extraction methods for untargeted metabolomic studies for future applications in zebrafish larvae infection models, Sci. Rep. 13 (2023) 7489. https://doi.org/10.1038/s41598-023-34593-y.

[78] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat. Protoc. 6 (2011) 1060–1083. https://doi.org/10.1038/nprot.2011.335.

[79] E. de Hoffmann, V. Stroobant, Mass Spectrometry: Principles and Applications, John Wiley & Sons, 2007.

[80] L. Han, Y. Sapozhnikova, S.J. Lehotay, Method validation for 243 pesticides and environmental contaminants in meats and poultry by tandem mass spectrometry coupled to low-pressure gas chromatography and ultrahigh-performance liquid chromatography, Food Control 66 (2016) 270–282. https://doi.org/10.1016/j.foodcont.2016.02.019.

[81] C. Chen, S. Kim, LC-MS-based Metabolomics of Xenobiotic-induced Toxicities, Comput. Struct. Biotechnol. J. 4 (2013) e201301008. https://doi.org/10.5936/csbj.201301008.

[82] G.A. Theodoridis, H.G. Gika, E.J. Want, I.D. Wilson, Liquid chromatography-mass spectrometry based global metabolite profiling: a review, Anal. Chim. Acta 711 (2012) 7–16. https://doi.org/10.1016/j.aca.2011.09.042.

[83] S. Boudah, M.-F. Olivier, S. Aros-Calt, L. Oliveira, F. Fenaille, J.-C. Tabet, C. Junot, 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. 966 (2014) 34–47. https://doi.org/10.1016/j.jchromb.2014.04.025.

[84] L. Perez de Souza, S. Alseekh, F. Scossa, A.R. Fernie, Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research, Nat. Methods 18 (2021) 733–746. https://doi.org/10.1038/s41592-021-01116-4.

[85] R. Ramautar, Capillary Electrophoresis–Mass Spectrometry for Metabolomics, Royal Society of Chemistry, 2018.

[86] T.O. Metz, E.S. Baker, E.L. Schymanski, R.S. Renslow, D.G. Thomas, T.J. Causon, I.K. Webb, S. Hann, R.D. Smith, J.G. Teeguarden, Integrating ion mobility spectrometry into

mass spectrometry-based exposome measurements: what can it add and how far can it go?, Bioanalysis 9 (2016) 81. https://doi.org/10.4155/bio-2016-0244.

[87] G. Paglia, J.P. Williams, L. Menikarachchi, J.W. Thompson, R. Tyldesley-Worster, S. Halldórsson, O. Rolfsson, A. Moseley, D. Grant, J. Langridge, B.O. Palsson, G. Astarita, Ion mobility derived collision cross sections to support metabolomics applications, Anal. Chem. 86 (2014) 3985–3993. https://doi.org/10.1021/ac500405x.

[88] B. Habchi, S. Alves, A. Paris, D.N. Rutledge, E. Rathahao-Paris, How to really perform high throughput metabolomic analyses efficiently?, TrAC Trends Anal. Chem. 85 (2016) 128–139. https://doi.org/10.1016/j.trac.2016.09.005.

[89] Atmospheric pressure ion sources - Covey - 2009 - Mass Spectrometry Reviews - Wiley Online Library, (n.d.). https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/mas.20246 (accessed November 5, 2024).

[90] M.M. Wolff, W.E. Stephens, A Pulsed Mass Spectrometer with Time Dispersion, Phys Rev 69 (1946) 691.

[91] W. Paul, H. Steinwedel, Notizen: Ein neues Massenspektrometer ohne Magnetfeld, Z. Für Naturforschung A 8 (1953) 448–450. https://doi.org/10.1515/zna-1953-0710.

[92] W. Paul, H. Steinwedel, A NEW MASS SPECTROMETER WITHOUT A MAGNETIC FIELD, in: 1953. https://www.semanticscholar.org/paper/A-NEW-MASS-SPECTROMETER-WITHOUT-A-MAGNETIC-FIELD-Paul-Steinwedel/8a18ab7bba031d36e6a66b44f669111731147b8c (accessed December 1, 2024).

[93] M.B. Comisarow, A.G. Marshall, Fourier transform ion cyclotron resonance spectroscopy, Chem. Phys. Lett. 25 (1974) 282–283. https://doi.org/10.1016/0009-2614(74)89137-2.

[94] Q. Hu, R.J. Noll, H. Li, A. Makarov, M. Hardman, R. Graham Cooks, The Orbitrap: a new mass spectrometer, J. Mass Spectrom. JMS 40 (2005) 430–443. https://doi.org/10.1002/jms.856.

[95] K.H. Kingdon, A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures, Phys. Rev. 21 (1923) 408–418. https://doi.org/10.1103/PhysRev.21.408.

[96] D. Cochran, R. Powers, Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Applications for Metabolomics, Biomedicines 12 (2024) 1786. https://doi.org/10.3390/biomedicines12081786.

[97] N. Caballero-Casero, L. Belova, P. Vervliet, J.-P. Antignac, A. Castaño, L. Debrauwer, M.E. López, C. Huber, J. Klanova, M. Krauss, A. Lommen, H.G.J. Mol, H. Oberacher, O. Pardo, E.J. Price, V. Reinstadler, C.M. Vitale, A.L.N. van Nuijs, A. Covaci, 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 Anal. Chem. 136 (2021) 116201. https://doi.org/10.1016/j.trac.2021.116201.

[98] S. Lennon, J. Chaker, E.J. Price, J. Hollender, C. Huber, T. Schulze, L. Ahrens, F. Béen, N. Creusot, L. Debrauwer, G. Dervilly, C. Gabriel, T. Guérin, B. Habchi, E.L. Jamin, J. Klánová, T. Kosjek, B. Le Bizec, J. Meijer, H. Mol, R. Nijssen, H. Oberacher, N. Papaioannou, J. Parinet, D. Sarigiannis, M.A. Stravs, Ž. Tkalec, E.L. Schymanski, M. Lamoree, J.-P. Antignac, A. David, Harmonized quality assurance/quality control provisions to assess completeness and robustness of MS1 data preprocessing for LC-HRMS-based suspect screening and non-targeted analysis, TrAC Trends Anal. Chem. 174 (2024) 117674. https://doi.org/10.1016/j.trac.2024.117674.

[99] D.M. Avtonomov, A. Raskind, A.I. Nesvizhskii, BatMass: a Java Software Platform for LC–MS Data Visualization in Proteomics and Metabolomics, J. Proteome Res. 15 (2016) 2500–2509. https://doi.org/10.1021/acs.jproteome.6b00021.

[100] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Orešič, Algorithms and tools for the preprocessing of LC–MS metabolomics data, Chemom. Intell. Lab. Syst. 108 (2011) 23–32. https://doi.org/10.1016/j.chemolab.2011.03.010.

[101] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification, Anal. Chem. 78 (2006) 779–787. https://doi.org/10.1021/ac051437y.

[102] J. Pfeuffer, T. Sachsenberg, O. Alka, M. Walzer, A. Fillbrunn, L. Nilse, O. Schilling, K. Reinert, O. Kohlbacher, OpenMS – A platform for reproducible analysis of mass spectrometry data, J. Biotechnol. 261 (2017) 142–148. https://doi.org/10.1016/j.jbiotec.2017.05.016.

[103] A. Rafiei, L. Sleno, Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis, Rapid Commun. Mass Spectrom. 29 (2015) 119–127. https://doi.org/10.1002/rcm.7094.

[104] E. Lange, R. Tautenhahn, S. Neumann, C. Gröpl, Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements, BMC Bioinformatics 9 (2008) 375. https://doi.org/10.1186/1471-2105-9-375.

[105] W. Tu, Zero-Inflated Data, in: Encycl. Environmetrics, John Wiley & Sons, Ltd, 2006. https://doi.org/10.1002/9780470057339.vaz000g.

[106] M. Rusilowicz, M. Dickinson, A. Charlton, S. O'Keefe, J. Wilson, A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples, Metabolomics 12 (2016) 56. https://doi.org/10.1007/s11306-016-0972-2.

[107] M. Krauss, H. Singer, J. Hollender, LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns, Anal. Bioanal. Chem. 397 (2010) 943–951. https://doi.org/10.1007/s00216-010-3608-9.

[108] D. Li, W. Liang, X. Feng, T. Ruan, G. Jiang, Recent advances in data-mining techniques for measuring transformation products by high-resolution mass spectrometry, TrAC Trends Anal. Chem. 143 (2021) 116409. https://doi.org/10.1016/j.trac.2021.116409.

[109] M. Musatadi, I. Baciero-Hernández, A. Prieto, M. Olivares, N. Etxebarria, O. Zuloaga, Development and evaluation of a comprehensive workflow for suspect screening of exposome-related xenobiotics and phase II metabolites in diverse human biofluids, Chemosphere 351 (2024) 141221. https://doi.org/10.1016/j.chemosphere.2024.141221.

[110] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, D.S. Wishart, BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification, J. Cheminformatics 11 (2019) 2. https://doi.org/10.1186/s13321-018-0324-5.

[111] L. Vergeynst, H. Van Langenhove, K. Demeestere, Balancing the False Negative and Positive Rates in Suspect Screening with High-Resolution Orbitrap Mass Spectrometry Using Multivariate Statistics, Anal. Chem. 87 (2015) 2170–2177. https://doi.org/10.1021/ac503426k.

[112] K. Goryński, B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, R. Kaliszan, Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds, Anal. Chim. Acta 797 (2013) 13–19. https://doi.org/10.1016/j.aca.2013.08.025.

[113] J. Chaker, E. Gilles, C. Monfort, C. Chevrier, S. Lennon, A. David, Scannotation: A Suspect Screening Tool for the Rapid Pre-Annotation of the Human LC-HRMS-Based Chemical Exposome, Environ. Sci. Technol. 57 (2023) 19253–19262. https://doi.org/10.1021/acs.est.3c04764.

[114] J. Parinet, Predicting reversed-phase liquid chromatographic retention times of pesticides by deep neural networks, Heliyon 7 (2021). https://doi.org/10.1016/j.heliyon.2021.e08563.

[115] L. Sleno, The use of mass defect in modern mass spectrometry, J. Mass Spectrom. 47 (2012) 226–236. https://doi.org/10.1002/jms.2953.

[116] C.A. Hughey, C.L. Hendrickson, R.P. Rodgers, A.G. Marshall, K. Qian, Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra, Anal. Chem. 73 (2001) 4676–4681. https://doi.org/10.1021/ac010560w.

[117] S. Kim, R.W. Kramer, P.G. Hatcher, Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram, Anal. Chem. 75 (2003) 5336–5344. https://doi.org/10.1021/ac034415p.

[118] Ò. Aznar-Alemany, B. Sala, K. Jobst, E. Reiner, A. Borrell, A. Aguilar, E. Eljarrat, Temporal trends of halogenated and organophosphate contaminants in striped dolphins from the Mediterranean Sea, Sci. Total Environ. 753 (2021) 142205. https://doi.org/10.1016/j.scitotenv.2020.142205.

[119] A. Léon, R. Cariou, S. Hutinet, J. Hurel, Y. Guitton, C. Tixier, C. Munschy, J.-P. Antignac, G. Dervilly-Pinel, B. Le Bizec, HaloSeeker 1.0: A User-Friendly Software to Highlight Halogenated Chemicals in Nontargeted High-Resolution Mass Spectrometry Data Sets, Anal. Chem. 91 (2019) 3500–3507. https://doi.org/10.1021/acs.analchem.8b05103.

[120] I. Oesterle, M. Pristner, S. Berger, M. Wang, V. Verri Hernandes, A. Rompel, B. Warth, Exposomic Biomonitoring of Polyphenols by Non-Targeted Analysis and Suspect Screening, Anal. Chem. 95 (2023) 10686–10694. https://doi.org/10.1021/acs.analchem.3c01393.

[121] H. Zhang, D. Zhang, K. Ray, A software filter to remove interference ions from drug metabolites in accurate mass liquid chromatography/mass spectrometric analyses, J. Mass Spectrom. JMS 38 (2003) 1110–1112. https://doi.org/10.1002/jms.521.

[122] H. Zhang, D. Zhang, K. Ray, M. Zhu, Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry, J. Mass Spectrom. 44 (2009) 999–1016. https://doi.org/10.1002/jms.1610.

[123] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.-M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), Metabolomics Off. J. Metabolomic Soc. 3 (2007) 211–221. https://doi.org/10.1007/s11306-007-0082-2.

[124] M. Sud, E. Fahy, D. Cotter, A. Brown, E.A. Dennis, C.K. Glass, A.H. Merrill Jr, R.C. Murphy, C.R.H. Raetz, D.W. Russell, S. Subramaniam, LMSD: LIPID MAPS structure database, Nucleic Acids Res. 35 (2007) D527–D532. https://doi.org/10.1093/nar/gkl838.

[125] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying Small Molecules *via* High Resolution Mass Spectrometry: Communicating Confidence, Environ. Sci. Technol. 48 (2014) 2097–2098. https://doi.org/10.1021/es5002105.

[126] J. Parinet, Y. Makni, T. Diallo, T. Guerin, Liquid Chromatographic Retention Time Prediction Models to Secure and Improve the Feature Annotation Process in High-Resolution Mass Spectrometry, (2023). https://doi.org/10.2139/ssrn.4501990.

[127] I. Blaženović, T. Kind, J. Ji, O. Fiehn, Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics, Metabolites 8 (2018) 31. https://doi.org/10.3390/metabo8020031.

[128] M. Berdi, B. de Lauzon-Guillain, A. Forhan, F.A. Castelli, F. Fenaille, M.-A. Charles, B. Heude, C. Junot, K. Adel-Patient, EDEN Mother-Child Cohort Study Group, Immune components of early breastmilk: Association with maternal factors and with reported food allergy in childhood, Pediatr. Allergy Immunol. Off. Publ. Eur. Soc. Pediatr. Allergy Immunol. 30 (2019) 107–116. https://doi.org/10.1111/pai.12998.

[129] N. Bekhti, F. Castelli, A. Paris, B. Guillon, C. Junot, C. Moiron, F. Fenaille, K. Adel-Patient, The Human Meconium Metabolome and Its Evolution during the First Days of Life, Metabolites 12 (2022) 414. https://doi.org/10.3390/metabo12050414.

[130] M. Berdi, Développement de méthodes pour l'analyse de la composition globale du lait maternel précoce dans la cohorte mère-enfant EDEN : impact des facteurs maternels et environnementaux (sur cette composition) et identification de biomarqueurs prédictifs d'une allergie alimentaire, phdthesis, Université Sorbonne Paris Cité, 2017. https://theses.hal.science/tel-04461984 (accessed November 14, 2024).

[131] R. Adusumilli, P. Mallick, Data Conversion with ProteoWizard msConvert, Methods Mol. Biol. Clifton NJ 1550 (2017) 339–368. https://doi.org/10.1007/978-1-4939-6747-6_23.

[132] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36. https://doi.org/10.1021/ci00057a005.

[133] B. Winter, C. Winter, J. Schilling, A. Bardow, A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing, Digit. Discov. 1 (2022) 859–869. https://doi.org/10.1039/D2DD00058J.

[134] Z.-M. Win, A.M.Y. Cheong, W.S. Hopkins, Using Machine Learning To Predict Partition Coefficient (Log P) and Distribution Coefficient (Log D) with Molecular Descriptors and Liquid Chromatography Retention Time, J. Chem. Inf. Model. 63 (2023) 1906–1913. https://doi.org/10.1021/acs.jcim.2c01373.

[135] A. James, D. Weininger, Daylight Inc. 4. SMARTS—a language for describing molecular patterns, (n.d.). https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed December 3, 2024).

[136] G. Landrum, RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, (n.d.).

[137] Les études de l'Alimentation Totale (EAT), Anses - Agence Natl. Sécurité Sanit. L'alimentation L'environnement Trav. (2019). https://www.anses.fr/fr/content/les-etudes-de-lalimentation-totale-eat (accessed November 27, 2024).

[138] J.L.C.M. Dorne, J. Richardson, A. Livaniou, E. Carnesecchi, L. Ceriani, R. Baldin, S. Kovarich, M. Pavan, E. Saouter, F. Biganzoli, L. Pasinato, M. Zare Jeddi, T.P. Robinson, G.E.N. Kass, A.K.D. Liem, A.A. Toropov, A.P. Toropova, C. Yang, A. Tarkhov, N. Georgiadis, M.R. Di Nicola, A. Mostrag, H. Verhagen, A. Roncaglioni, E. Benfenati, A. Bassan, EFSA's OpenFoodTox: An open source toxicological database on chemicals in food and feed and its future developments, Environ. Int. 146 (2021) 106293. https://doi.org/10.1016/j.envint.2020.106293.

[139] F. Hernández, J.V. Sancho, M. Ibáñez, E. Abad, T. Portolés, L. Mattioli, Current use of high-resolution mass spectrometry in the environmental sciences, Anal. Bioanal. Chem. 403 (2012) 1251–1264. https://doi.org/10.1007/s00216-012-5844-7.

[140] M. Castro-Puyana, R. Pérez-Míguez, L. Montero, M. Herrero, Reprint of: Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability, TrAC Trends Anal. Chem. 96 (2017) 62–78. https://doi.org/10.1016/j.trac.2017.08.007.

[141] E.J. Finehout, K.H. Lee, An introduction to mass spectrometry applications in biological research, Biochem. Mol. Biol. Educ. Bimon. Publ. Int. Union Biochem. Mol. Biol. 32 (2004) 93–100. https://doi.org/10.1002/bmb.2004.494032020331.

[142] R.A. Yost, C.G. Enke, Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation, Anal. Chem. 51 (1979) 1251–1264. https://doi.org/10.1021/ac50048a002.

[143] Linear ion traps in mass spectrometry - Douglas - 2005 - Mass Spectrometry Reviews - Wiley Online Library, (n.d.). https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/mas.20004 (accessed February 23, 2024).

[144] Z.-J. Zhu, A.W. Schultz, J. Wang, C.H. Johnson, S.M. Yannone, G.J. Patti, G. Siuzdak, Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database, Nat. Protoc. 8 (2013) 451–460. https://doi.org/10.1038/nprot.2013.004.

[145] R.M.A. Heeren, A.J. Kleinnijenhuis, L.A. McDonnell, T.H. Mize, A mini-review of mass spectrometry using high-performance FTICR-MS methods, Anal. Bioanal. Chem. 378 (2004) 1048–1058. https://doi.org/10.1007/s00216-003-2446-4.

[146] R.H. Perry, R.G. Cooks, R.J. Noll, Orbitrap mass spectrometry: Instrumentation, ion motion and applications, Mass Spectrom. Rev. 27 (2008) 661–699. https://doi.org/10.1002/mas.20186.

[147] O. Coskun, Separation Techniques: CHROMATOGRAPHY, North. Clin. Istanb. (2016). https://doi.org/10.14744/nci.2016.32757.

[148] K.D. Bartle, P. Myers, History of gas chromatography, TrAC Trends Anal. Chem. 21 (2002) 547–557. https://doi.org/10.1016/S0165-9936(02)00806-3.

[149] J.G. Dorsey, W.T. Cooper, J.F. Wheeler, H.G. Barth, J.P. Foley, Liquid Chromatography: Theory and Methodology, Anal. Chem. 66 (1994) 500–546. https://doi.org/10.1021/ac00084a019.

[150] Y. Ni, M. Su, Y. Qiu, M. Chen, Y. Liu, A. Zhao, W. Jia, Metabolic profiling using combined GC–MS and LC–MS provides a systems understanding of aristolochic acid-induced nephrotoxicity in rat, FEBS Lett. 581 (2007) 707–711. https://doi.org/10.1016/j.febslet.2007.01.036.

[151] K. Segers, S. Declerck, D. Mangelings, Y.V. Heyden, A.V. Eeckhaut, Analytical techniques for metabolomic studies: a review, Bioanalysis 11 (2019) 2297–2318. https://doi.org/10.4155/bio-2019-0014.

[152] X. Han, A. Aslanian, J.R. Yates, Mass spectrometry for proteomics, Curr. Opin. Chem. Biol. 12 (2008) 483–490. https://doi.org/10.1016/j.cbpa.2008.07.024.

[153] D. Aurich, O. Miles, E.L. Schymanski, Historical exposomics and high resolution mass spectrometry, Exposome 1 (2021) osab007. https://doi.org/10.1093/exposome/osab007.

[154] B.G. Katzung, ed., Basic & clinical pharmacology, Fourteenth edition, McGraw-Hill Education, New York Chicago San Francisco Athens London Madrid Mexico City Milan New Delhi Singapore Sydney Toronto, 2018.

[155] Genetic basis of drug metabolism | American Journal of Health-System Pharmacy | Oxford Academic, (n.d.). https://academic.oup.com/ajhp/article-abstract/59/21/2061/5157950 (accessed February 25, 2024).

[156] K.K. Murray, Comment on: “Nominal Mass?” by Athula B. Attygalle and Julius Pavlov, J. Am. Soc. Mass Spectrom. 28, 1737-1738 (2017), J. Am. Soc. Mass Spectrom. 28 (2017) 2724–2725. https://doi.org/10.1007/s13361-017-1801-1.

[157] Edward. Kendrick, A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds., Anal. Chem. 35 (1963) 2146–2154. https://doi.org/10.1021/ac60206a048.

[158] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics 8 (2007) 105. https://doi.org/10.1186/1471-2105-8-105.

[159] E. Rathahao-Paris, A. Paris, J. Bursztyka, J.-P. Jaeg, J.-P. Cravedi, L. Debrauwer, Identification of xenobiotic metabolites from biological fluids using flow injection analysis high-resolution mass spectrometry and post-acquisition data filtering, Rapid Commun. Mass Spectrom. 28 (2014) 2713–2722. https://doi.org/10.1002/rcm.7066.

[160] E. Rathahao-Paris, S. Alves, L. Debrauwer, J.-P. Cravedi, A. Paris, An efficient data-filtering strategy for easy metabolite detection from the direct analysis of a biological fluid using Fourier transform mass spectrometry, Rapid Commun. Mass Spectrom. RCM 31 (2017) 485–494. https://doi.org/10.1002/rcm.7812.

[161] B. Heude, A. Forhan, R. Slama, L. Douhaud, S. Bedel, M.-J. Saurel-Cubizolles, R. Hankard, O. Thiebaugeorges, M. De Agostini, I. Annesi-Maesano, M. Kaminski, M.-A. Charles, EDEN mother-child cohort study group, Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development, Int. J. Epidemiol. 45 (2016) 353–363. https://doi.org/10.1093/ije/dyv151.

[162] S. Boudah, M.-F. Olivier, S. Aros-Calt, L. Oliveira, F. Fenaille, J.-C. Tabet, C. Junot, 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. 966 (2014) 34–47. https://doi.org/10.1016/j.jchromb.2014.04.025.

[163] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria., n.d.

[164] F. Cuyckens, R. Hurkmans, J.M. Castro-Perez, L. Leclercq, R.J. Mortishire-Smith, Extracting metabolite ions out of a matrix background by combined mass defect, neutral loss and isotope filtration, Rapid Commun. Mass Spectrom. RCM 23 (2009) 327–332. https://doi.org/10.1002/rcm.3881.

[165] A. Roy-Lachapelle, M. Solliec, M. Sinotte, C. Deblois, S. Sauvé, High resolution/accurate mass (HRMS) detection of anatoxin-a in lake water using LDTD–APCI coupled to a Q-Exactive mass spectrometer, Talanta 132 (2015) 836–844. https://doi.org/10.1016/j.talanta.2014.10.021.

[166] Y. Xu, J.-F. Heilier, G. Madalinski, E. Genin, E. Ezan, J.-C. Tabet, C. Junot, Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-orbitrap mass spectrometer for further metabolomics database building, Anal. Chem. 82 (2010) 5490–5501. https://doi.org/10.1021/ac100271j.

[167] A. McMillan, J.B. Renaud, G.B. Gloor, G. Reid, M.W. Sumarah, Post-acquisition filtering of salt cluster artefacts for LC-MS based human metabolomic studies, J. Cheminformatics 8 (2016). https://doi.org/10.1186/s13321-016-0156-0.

[168] R. Cariou, E. Omer, A. Léon, G. Dervilly-Pinel, B. Le Bizec, Screening halogenated environmental contaminants in biota based on isotopic pattern and mass defect

provided by high resolution mass spectrometry profiling, Anal. Chim. Acta 936 (2016) 130–138. https://doi.org/10.1016/j.aca.2016.06.053.

[169] H. Zhang, M. Zhu, K.L. Ray, L. Ma, D. Zhang, Mass defect profiles of biological matrices and the general applicability of mass defect filtering for metabolite detection, Rapid Commun. Mass Spectrom. RCM 22 (2008) 2082–2088. https://doi.org/10.1002/rcm.3585.

[170] A. Bérard, F. Abbas-Chorfa, B. Kassai, T. Vial, K.A. Nguyen, O. Sheehy, A.-M. Schott, The French Pregnancy Cohort: Medication use during pregnancy in the French population, PLoS ONE 14 (2019) e0219095. https://doi.org/10.1371/journal.pone.0219095.

[171] S. Lamy, E. Houivet, J. Benichou, S. Marret, F. Thibaut, Perinatal network of Upper-Normandy, Caffeine use during pregnancy: prevalence of use and newborn consequences in a cohort of French pregnant women, Eur. Arch. Psychiatry Clin. Neurosci. 271 (2021) 941–950. https://doi.org/10.1007/s00406-020-01105-2.

[172] B. Pierrot, G. Legendre, J. Riou, A. Gentil, B. Molle-Guiliani, A. Petit, Pregnancy and tobacco: Practice and knowledge of French midwives, Midwifery 129 (2024) 103886. https://doi.org/10.1016/j.midw.2023.103886.

[173] E. Jauniaux, B. Gulbis, G. Acharya, P. Thiry, C. Rodeck, Maternal tobacco exposure and cotinine levels in fetal fluids in the first half of pregnancy, Obstet. Gynecol. 93 (1999) 25–29. https://doi.org/10.1016/s0029-7844(98)00318-4.

[174] T. Nagao, D. Yukihira, Y. Fujimura, K. Saito, K. Takahashi, D. Miura, H. Wariishi, Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: *In silico* evaluation and metabolomic application, Anal. Chim. Acta 813 (2014) 70–76. https://doi.org/10.1016/j.aca.2014.01.032.

[175] L. Debrauwer, L. Mervant, O. Laprevote, E.L. Jamin, Pivotal Role of Mass Spectrometry for the Assessment of Exposure to Reactive Chemical Contaminants: From the Exposome to the Adductome, Mass Spectrom. Rev. (2024). https://doi.org/10.1002/mas.21917.

.

Annexes

Annexe 1. Liste d'exclusion des composés endogènes

Nom
Adenosine 3',5'-diphosphate
Glutathione
NADP
NADPH
Uridine 5'-diphosphate
Adenosine triphosphate
NAD
Uridine diphosphate glucuronic acid
Phosphoadenosine phosphosulfate
FAD
Oxygen
Formaldehyde
Phosphate
NADH
Water
Hydrogen peroxide
Heme
Hydrogen Ion
Adenosine diphosphate
Zinc
5-Deoxyribose-1-phosphate
Coenzyme A
Propionyl-CoA
Pentanoyl-CoA
Pyrophosphate
Uridine triphosphate
Phosphorylcholine

Annexe 2.

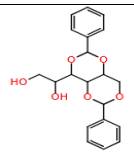
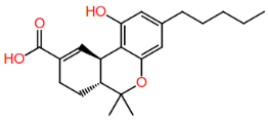
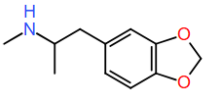
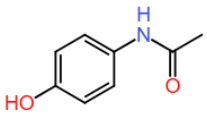
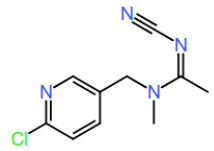
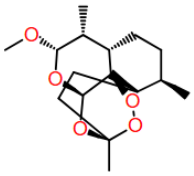
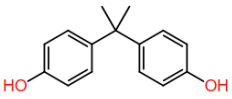
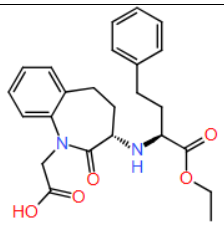
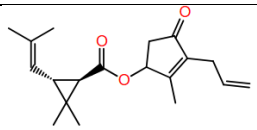
Annexe 2.1. : Liste des composés utilisés comme étalons internes (3,75 µg/mL) lors la préparation des échantillons

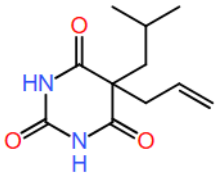
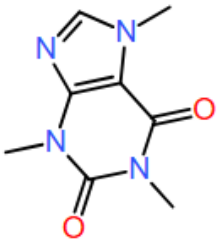
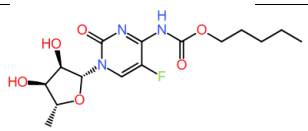
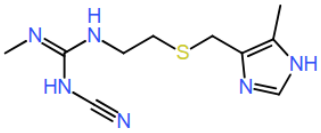
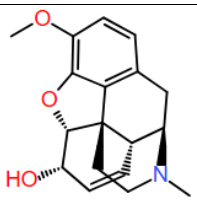
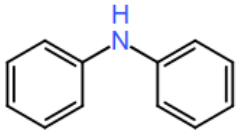
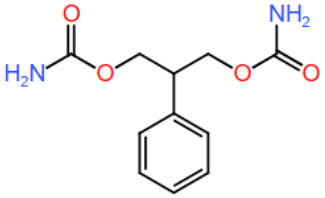
Etalons Internes	Composition élémentaire	Masse précise	$[M+H]^+$/adduit *m/z*	$[M-H]^-$/adduit *m/z*
Dimetridazole	$C_5H_7N_3O_2$	141,05383	142,06111	**Nd**
AMPA 2-amino-3-(3-hydroxy-5-methyl-isoxazol-4-yl)propanoic acid	$C_7H_{10}N_2O_4$	186,06406	187,07134	185,05678
MCPA 2-methyl-4-chlorophenoxyacetic acid	$C_9H_9ClO_3$	200,02402	**nd**	199,01674
Dinoseb	$C_{10}H_{12}N_2O_5$	240,07462	**nd**	239,06734

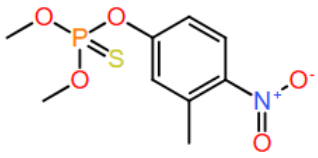
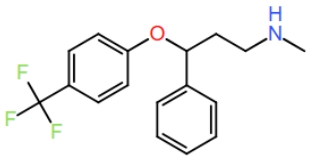
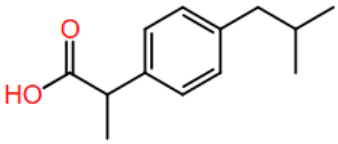
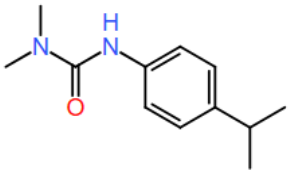
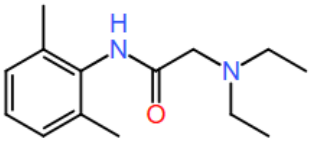
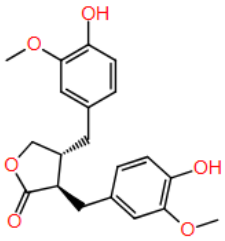
Annexe 2.2. : Liste des composés utilisés comme étalons externes pour des analyses LC-HRMS

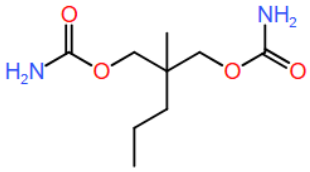
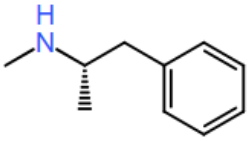
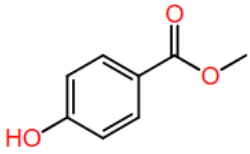
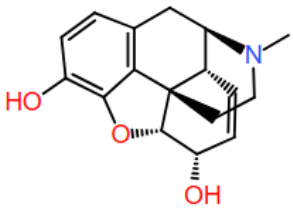
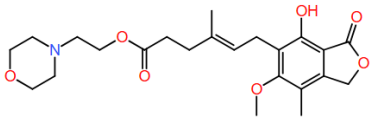
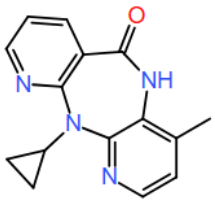
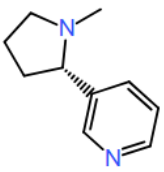
Etalons externes	Composition élémentaire	Masse précise	$[M+H]^+$/adduit *m/z*	$[M-H]^-$/adduit *m/z*
Alanine ^{13}C	$^{13}C_1\,^{12}C_2\,H_7\,N\,O_2$	90,05103	91,05831	**nd**
Metformin	$C_4\,H_{11}\,N_5$	129,10145	130,10873	**nd**
Ethylmalonic acid	$C_5\,H_8\,O_4$	132,04226	**nd**	131,03498
Aspartate ^{15}N	$C_4\,H_7\,^{15}N_1\,O_4$	134,03454	135,04182	133,02726
Glucose ^{13}C	$^{13}C_1\,^{12}C_5\,H_{12}\,O_6$	181,06674	204.05596 (+Na)	216.03614 (+HCl)
2-Aminoanthracene	$C_{14}\,H_{11}\,N$	193,08915	194,09643	**nd**
Amiloride	$C_6\,H_8\,Cl\,N_7\,O$	229,04789	230,05517	228.04061 et 264.01729 (+HCl)
Imipramine	$C_{19}\,H_{24}\,N_2$	280,19395	281,20123	**nd**
Atropine	$C_{17}\,H_{23}\,N\,O_3$	289,16779	290,17507	**nd**
Prednisone	$C_{21}\,H_{26}\,O_5$	358,17802	359,1853	403.17482 (+HCOOH) et 393.14742 (+HCl)
Colchicine	$C_{22}\,H_{25}\,N\,O_6$	399,16819	400,17547	**nd**
Dihydrostreptomycin	$C_{21}\,H_{41}\,N_7\,O_{12}$	583,28132	584.2886 et 292.64794	618.25072 (+HCl)
Roxithromycin (fragment)	$C_{29}\,H_{54}\,O_{10}\,N_2$	590,37784	591,38512	625.34725 (+HCl)

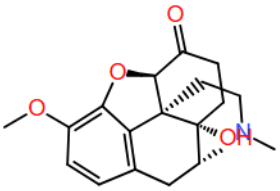
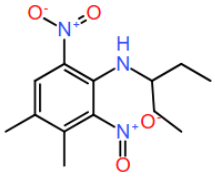
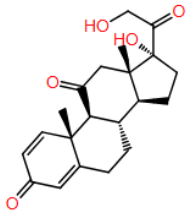
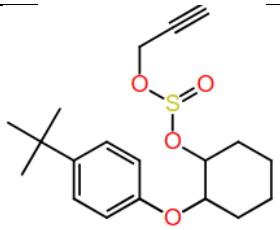
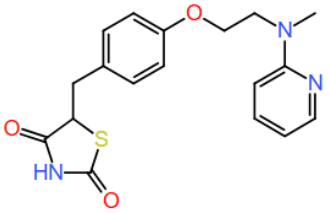
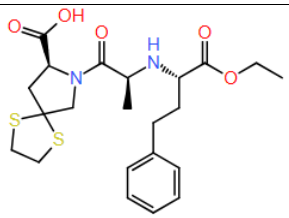
Annexe 3. Liste des molécules putatives détectées dans les échantillons de méconium en mode positif

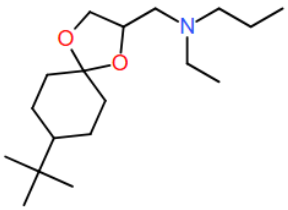
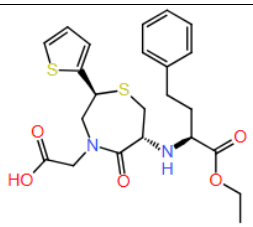
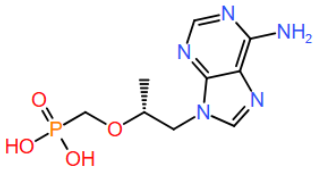
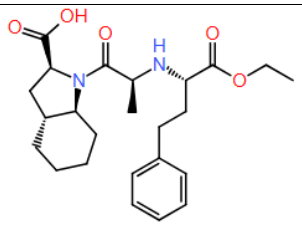
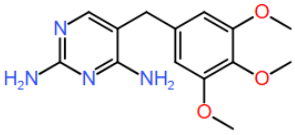
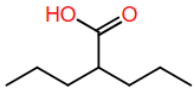
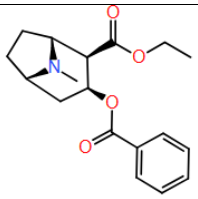
Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Dibenzylidène sorbitol		$C_{20}H_{22}O_6$	359.1489
11-nor-9-carboxy-Delta(9)-tétrahydrocannabinol		$C_{21}H_{28}O_4$	345.2060
3,4-méthylènedioxy-N-méthylamphétamine		$C_{11}H_{15}NO_2$	194.1176
Acétaminophène		$C_8H_9NO_2$	152.0706
Acétamipride		$C_{10}H_{11}ClN_4$	223.0745
Artéméther		$C_{16}H_{26}O_5$	299.1853
Bisphénol A		$C_{15}H_{16}O2$	229.1223
Bénazépril		$C_{24}H_{28}N_2O_5$	425.2071
Bioalléthrine		$C_{19}H_{26}O_3$	303.1955

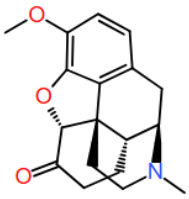
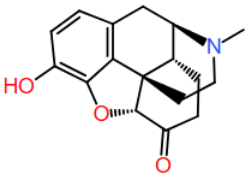
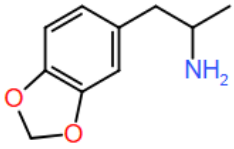
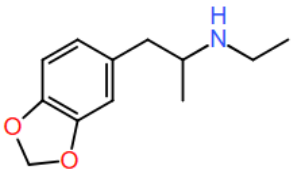
Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Butalbital		$C_{11}H_{16}N_2O_3$	225.1234
Caféine		$C_8H_{10}N_4O_2$	195.0877
Capécitabine		$C_{15}H_{22}FN_3O_6$	360.1565
Cimétidine		$C_{10}H_{16}N_6S$	253.1230
Codéine		$C_{18}H_{21}NO_3$	300.1594
Diphénylamine		C12H11N	170.0964
Felbamate		$C_{11}H_{14}N_2O_4$	239.1026

Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Fénitrothion		$C_9H_{12}NO_5PS$	278.0247
Fluoxétine		$C_{17}H_{18}F_3NO$	310.1413
Ibuprofène		$C_{13}H_{18}O_2$	207.1380
Isoproturon		$C_{12}H_{18}N_2O$	207.1492
Lidocaïne		$C_{14}H_{22}N_2O$	235.1805
Matairésinol		$C_{20}H_{22}O_6$	359.1489

Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Méprobamate		$C_9H_{18}N_2O_4$	219.1339
Méthamphétamine		$C_{10}H_{15}N$	150.1277
Méthylparabène		$C_8H_8O_3$	153.0546
Morphine		$C_{17}H_{19}NO_3$	286.1438
Mycophénolate mofétil		$C_{23}H_{31}NO_7$	434.2173
Névirapine		$C_{15}H_{14}N_4O$	267.1240
Nicotine		$C_{10}H_{14}N_2$	163.1230

Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Oxycodone		$C_{18}H_{21}NO_4$	316.1543
Pendiméthaline		$C_{13}H_{19}N_3O_4$	282.1448
Prednisone		$C_{21}H_{26}O_5$	359.1853
Propargite		$C_{19}H_{26}O_4S$	351.1625
Rosiglitazone		$C_{18}H_{19}N_3O_3S$	358.1220
Spirapril		$C_{22}H_{30}N_2O_5S_2$	467.1669

Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Spiroxamine		$C_{18}H_{35}NO_2$	298.2741
Temocapril		$C_{23}H_{28}N_2O_5S_2$	477.1512
Ténofovir		$C_9H_{14}N_5O_4P$	288.0856
Trandolapril		$C_{24}H_{34}N_2O_5$	431.2540
Triméthoprime		$C_{14}H_{18}N_4O_3$	291.1452
Acide valproïque		$C_8H_{16}O_2$	145.1223
Cocaéthylène		$C_{18}H_{23}NO_4$	318.1700

Composé	Structure	Formule	$[M+H]^+$ m/z théorique
Hydrocodone		$C_{18}H_{21}NO_3$	300.1594
Hydromorphone		$C_{17}H_{19}NO_3$	286.1438
3,4-Méthylènedioxyamphétamine		$C_{10}H_{13}NO_2$	180.1019
3,4-méthylènedioxy-N-éthylamphétamine		$C_{12}H_{17}NO_2$	208.1332