Development of a novel universal proxy to assess the environmental fate and impact of complex (bio)pesticides by Mass Spectrometry-based Metabolomics

Hikmat Ghosson

THÈSE

Pour obtenir le grade de

Docteur

Délivré par

UNIVERSITE DE PERPIGNAN VIA DOMITIA

Préparée au sein de l'école doctorale

ED 305 Energie et Environnement

Et de l'unité de recherche

USR 3278 – CRIOBE – EPHE-CNRS-UPVD

Spécialité :

Chimie Analytique

Présentée par

Hikmat GHOSSON

TITRE DE LA THESE

Développement d'un nouveau proxy universel pour évaluer le devenir et l'impact environnemental de (bio)pesticides complexes par Métabolomique basée sur la Spectrométrie de Masse

Soutenue le **04/12/2020** devant le jury composé de

Mme. Carole CALAS-BLANCHARD, Professeur des Universités *Université de Perpignan Via Domitia*	Président
Mme. Emmanuelle VULLIET, Directeur de recherche, H.D.R. *Centre National de la Recherche Scientifique*	Rapporteur
M. Laurent DEBRAUWER, Ingénieur de recherche, H.D.R. *Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement*	Rapporteur
M. Julien BOCCARD, Collaborateur scientifique *Université de Genève*	Examinateur
M. Yann GUITTON, Ingénieur de recherche *École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation, Nantes-Atlantique*	Examinateur
M. Cédric BERTRAND, Professeur des Universités *Université de Perpignan Via Domitia*	Directeur de thèse
Mme. Marie-Virginie SALVIA, Maître de conférences, H.D.R. *Université de Perpignan Via Domitia*	Co-directeur

A Thesis submitted for the degree of

Doctor of Philosophy

Delivered by

UNIVERSITE DE PERPIGNAN VIA DOMITIA

Prepared at the doctoral school

ED 305 Energie et Environnement

And the research unit

USR 3278 – CRIOBE – EPHE-CNRS-UPVD

Major

Analytical Chemistry

By

Hikmat GHOSSON

THESIS TITLE

Development of a novel universal proxy to assess the environmental fate and impact of complex (bio)pesticides by Mass Spectrometry-based Metabolomics

Defended on **December 04, 2020** before the jury composed of

Mme. Carole CALAS-BLANCHARD, Professeur des Universités *Université de Perpignan Via Domitia*	President
Mme. Emmanuelle VULLIET, Directeur de recherche, H.D.R. *Centre National de la Recherche Scientifique*	Rapporteur
M. Laurent DEBRAUWER, Ingénieur de recherche, H.D.R. *Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement*	Rapporteur
M. Julien BOCCARD, Collaborateur scientifique *Université de Genève*	Examiner
M. Yann GUITTON, Ingénieur de recherche *École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation, Nantes-Atlantique*	Examiner
M. Cédric BERTRAND, Professeur des Universités *Université de Perpignan Via Domitia*	Thesis Director
Mme. Marie-Virginie SALVIA, Maître de conférences, H.D.R. *Université de Perpignan Via Domitia*	Co-director

Dedication

To those who sacrificed their lives
when struggling for the foundation of
the Lebanese National University…

“Aut viam inveniam aut faciam”

“We will either find a way, or make one”

Hannibal Barca, the Canaanite Commander of Carthage

Acknowledgments

I would like to thank every person who contributed, supported, or helped realizing the present work and fulfilling the missions required for achieving the present dissertation, and I apologize in advance to any single person who I may forget to mention… You were many, colleagues and friends, to participate in this experience and I am very grateful to you and to your appreciated efforts and support.

First, I would like to acknowledge the members of the Jury **Dr. Emmanuelle VULLIET** and **Dr. Laurent DEBRAUWER**, to have given me the honor of being the Rapporteurs of this dissertation, and to have given their precious time to read and assess of the present manuscript. I also acknowledge **Pr. Carole CALAS-BLANCHARD**, **Dr. Yann GUITTON** and **Dr. Julien BOCCARD** for their acceptance to examine the present work and to participate in the Jury of my thesis. This is a great honor for me.

I would like to express my gratitude for the two persons who gave me the opportunity to access this distinct experience, and supervised my work during those three years. They let me take the initiative, helped me focusing the targets, and were always open to discuss new ideas and propositions… And despite all the difficulties and the critical moments, they always supported my work by transferring their knowledge and experience to me, which improved my scientific, technical and personal skills and maturity. They gave me their trust, and finally guided me towards the accomplishment of this dissertation with a lot of satisfaction from my part, so I am very grateful for them, and I had the pleasure to work under their guidance, and still… Thus, I express my deepest thanks to my supervisors **Pr. Cédric BERTRAND** and **Dr. Marie-Virginie SALVIA**.

I would like express my deepest recognition to **Dr. Frédérique COURANT**, **Dr. Anne-Emmanuelle HAY-DE BETTIGNIES**, **Pr. Thierry NOGUER**, and once again, **Dr. Yann GUITTON** for their acceptance to guide and participate in my "Comité de Suivi Individuel" (CSI), and for their appreciated support, guidance and advices that helped me achieving my targets and orientating my future perspectives.

I acknowledge the **Doctoral School "ED 305"**, represented by **Pr. Nicolas INGUIMBERT**, the Doctoral School Director, and the Doctoral School Council represented by all its members, for granting me the doctoral fellowship, and for supporting the pursuit of my learning cursus and training activities. I also acknowledge the **"Réseau Francophone de Métabolomique et Fluxomique" (RFMF)**, and the **"Société Française de Spectrométrie de Masse" (SFSM)** for the grants that they afforded to me during the three years of the doctoral work, and for allowing me to participate and communicate in the workshops, the schools and the congresses that they organized. These participations helped me improving my knowledge, my skills and my network that were necessary for the accomplishing of the present thesis, and for my aimed future career in scientific research as well.

I would like to express my acknowledgments to all the collaborators who were involved in all or parts of the present thesis. They will be mentioned in the subsequent chapters of the present manuscript.

I express all my sincere gladness to have worked with the CRIOBE team, so I thank all the former and actual **"Criobian" colleagues**, without any exception, for being a graceful part from this experience. I express my special amical gratitude for my colleagues **Chandra**, **Mathieu**, **Amani**, **Slimane**, **Mélina**, **Mélanie**, **Laurine**, **Christian**, and **Christelle**, for the great time that we spent together when trying to "joyfully" struggle with our theses and research topics. Also, my special thanks to **Delphine**, **Jennifer** and **Peter** who were always here to help and afford their great support.

Otherwise, I would like to express my heartfelt thanks to my former supervisors and teachers **Dr. Nicolas LE YONDRE**, **Dr. Adrián SCHWARZENBERG** and **Dr. Wafaa NOUN** for teaching me and for encouraging my will to take this distinct challenge, and for their permanent and precious support before and during those three years. They always afforded a great help for me, especially when I faced many difficult moments and obstacles, and they have never stopped dedicating their time, their support and their knowledge whenever I asked.

I also warmly thank my friend-colleague **Dr. Rim JABER**, for affording her great help and precious time when I was preparing for the doctoral entrance exam and for all the professional and personal advices that she gave me along those three years.

I am very grateful and should acknowledge my teachers **Mr. Imad DIB**, **Mr. Rami ABOU SAMRA**, **Mr. Joseph HLEIHEL**, **Mr. Naji MAKHOUL**, **Mrs. Maria KAZZI** and **Mrs. Rola ATALLAH** for baptizing my early passion in Science. I also acknowledge their essential role in teaching me how to assume my responsibilities and how to respect my professional, social and personal duties. I was and will be always honored for being one of your students…

My deepest acknowledgments to my friends-colleagues **Dr. Ihsan EL MASRI**, **Dr. Mohamad EL MAOUCH**, **Dr. May TAHA**, **Dr. Ali EL ZEIN**, **Dr. Racha ZGHEIB** and **Dr. Jamal EL HACHEM** for their encouragement and permanent support before and during those three years.

My warmest thanks to the dearest **Mira**, my classmate-partner, my colleague, and my great friend who was always here to encourage me, and for being the person who understood the most my difficulties, as well as my gladness, my perspectives and my ambitions.

My distinct thanks to **Sara**, for always being here, especially during those last and most difficult moments. You were a great unforgettable support. I owe you one so when you will make your decision and take your way towards your own adventure, be sure, I will be there…

And last but not least, all my profound, warm, and grateful thanks to my small family… **Maha**, **Ibrahim**, **Samar** and **Ward**.. No words can express my gratitude to you… Your ultimate confidence and infinite support were and will be always the reason why I can overcome difficulties, achieve success, and aim beyond…

Résumé

Malgré la prise de conscience écologique et sanitaire, la consommation mondiale des pesticides est en augmentation. Étant donné que ces produits chimiques présentent de nombreux effets néfastes sur la santé humaine et l'environnement, des mesures doivent être prises afin de limiter leurs effets. Les produits de biocontrôle sont proposés comme une solution alternative aux produits synthétiques. En effet, ces « biopesticides » sont présumés être moins nocifs et relativement moins persistants. Toutefois, cet *a priori* doit être examiné et une évaluation stricte des risques de ces nouvelles substances doit être envisagée.

Le développement de solutions de biocontrôle passe d'abord par les protocoles proposés pour étudier leur activité, leur devenir et leur impact environnemental. Actuellement, le temps de demi-vie (t½) est utilisé pour évaluer le devenir environnemental des pesticides synthétiques. Cependant, l'approche t½ ne donne que des informations sur la persistance des pesticides dans l'environnement, mais aucune indication concernant la formation de produits de dégradation ou son impact sur la biodiversité n'est apportée. De plus, les produits de biocontrôle sont des mélanges (bio)chimiques complexes. La t½ n'est pas applicable pour ce type des produits. Par conséquent, de nouvelles approches analytiques doivent être envisagées afin de surmonter ces limitations.

Une nouvelle approche basée sur la méta-métabolomique non-ciblée et la Spectrométrie de Masse ; nommée « Empreinte Métabolique Environnementale » (EMF), a été récemment introduite. Elle offre un nouveau proxy universel et intégratif; le « temps de résilience », dédié à l'évaluation du devenir environnemental et l'impact des (bio)pesticides complexes dans des matrices environnementales (ex. sol, sédiments). Cette approche vise à analyser le méta-métabolome d'une matrice polluée et le comparer à celui d'une matrice non-polluée. Le méta-métabolome d'une matrice polluée est composé de deux parties principales : 1) « le xénométabolome » contenant la substance active, les ingrédients de formulation et les sous-produits de transformation du pesticide, et 2) « l'endometabolome » produit par les microorganismes de la matrice. Un méta-métabolome d'une matrice non-polluée comprend uniquement l'endometabolome. Le devenir du pesticide sera ainsi étudié par le suivi de l'évolution du xénométabolome au cours du temps. Par ailleurs, la comparaison des profils endométaboliques des échantillons pollués à ceux des échantillons non-pollués permet d'évaluer l'impact du pesticide sur la biodiversité de la matrice.

Néanmoins, le développement et l'évaluation d'une telle approche méta-métabolomique récente et complexe doivent être effectués en profondeur. Plusieurs problématiques doivent alors être abordées : 1) des protocoles d'extractions performantes et des méthodes analytiques de pointe doivent être mis en place, 2) les pipelines de traitement de données et les outils chimiométriques appropriés doivent être développés pour maitriser la complexité des ensembles des données générées, 3) l'impact de la complexité du méta-métabolome sur les analyses basées sur la Spectrométrie de Masse doit être évalué, et 4) l'étude des résidus volatiles doit être envisagée et nécessite donc le développement de nouvelles méthodologies analytiques.

Ainsi, le travail a été mené sur 3 axes principaux. Le premier axe, décrit dans le premier chapitre expérimental de la thèse, portait sur deux points étroitement liés. Le premier est le développement des protocoles d'extraction et d'une méthode LC-HRMS pour analyser à la fois les xénométabolites des pesticides et les endométabolites du sol. Ces extractions et méthode doivent aussi être capables de couvrir une gamme moléculaire relativement large en terme de polarité. Le deuxième point est le développement d'une nouvelle approche chimiométrique visant à évaluer les performances des protocoles d'extraction développés. Cette évaluation est basée sur des critères analytiques et contextuels : 1) la capacité du protocole à couvrir la plus large gamme possible en terme de polarité et de famille de métabolites, 2) le compromis entre cette large couverture moléculaire et le rendement quantitative de l'extraction, 3) la reproductibilité du protocole, et 4) la capacité de discriminer les échantillons pollués de ceux non-pollués à partir de leurs profils méta-métaboliques. Dans ces buts, deux nouveaux protocoles d'extraction ont été développés et comparés à trois autres protocoles référencés dans la bibliographie. Ces cinq protocoles ont été appliqués sur deux types de sols de propriétés physico-chimiques dissemblables, pollués par deux pesticides de natures et complexités différentes. Les cinq différentes extractions ont été également appliquées sur des groupes d'échantillons non-pollués servant de contrôle (150 échantillons de microcosmes en total). D'autre part, une méthode UHPLC-ESI-Q/ToF et une approche chimiométrique utilisant les outils computationnels et statistiques de la métabolomique non-ciblée ont été mises en place. La méthode analytique a été bien adaptée pour l'analyse et la détection des petites molécules avec une capacité relativement acceptable pour la couverture d'une large gamme de polarité (en utilisant une colonne RPLC modifiée). Par ailleurs, l'approche chimiométrique a permis de décrypter les larges et complexes jeux de données LC-HRMS multifactoriels. Elle a donc fourni un outil permettant d'analyser et interpréter les résultats qui ont démontrés les nouveaux protocoles développés comme optimaux pour l'EMF et le contexte étudié. Ces nouveaux

protocoles sont principalement basés sur deux étapes rapides impliquant des mélanges quaternaires de solvants polaires et apolaires miscibles. En comparaison avec les trois autres protocoles de simple étape, ils ont été capables 1) de couvrir la gamme de polarité la plus large avec des rendements acceptables, 2) de montrer les meilleures performances pour l'extraction et la détection des xénométabolites et des endométabolites à la fois, avec 3) des reproductibilités acceptables, et 4) des fortes capacités de discrimination entre des sols pollués et non-pollués.

Le deuxième axe, évoqué dans le deuxième chapitre expérimental, a visé l'évaluation de l'effet des complexités hétérogènes des méta-métabolomes sur la détermination des biomarqueurs environnementaux. Cette évaluation a été conduite suite à des observations questionnables lors d'une étude cinétique menée sur du sédiment pollué par un biopesticide complexe basé sur des microorganismes. En effet, les analyses multivariées et les statistiques ont démontré une diminution significative d'intensités de certains endométabolites dans les groupes pollués. Cependant, la vérification des données brutes LC-HRMS a révélé que dans les échantillons pollués, ces candidats de biomarqueurs co-éluent avec des xénométabolites qui génèrent des amas importants d'ions multichargés en Electrospray (des macromolécules issues des ingrédients de formulation). Cette observation a suscité des doutes sur la possibilité de l'occurrence du phénomène de la « suppression d'ions » qui provoque une perte du signal. Cela a mené à considérer que la diminution significative du signal des candidats de biomarqueurs est potentiellement due à la suppression d'ions, et non pas à un effet biologique. Pour tester cette hypothèse, une approche pragmatique basée sur la dilution des échantillons a été mise en place. Elle a pu révéler que pour la majorité des candidats sélectionnés, la significativité de la diminution du signal a été perdue après les dilutions, et a donc permis de conclure que la suppression d'ions issue des complexités hétérogènes des méta-métabolomes pourrait entrainer des faux-positifs. Pour cela, la vérification des chromatogrammes et des spectres de masse doit être systématique dans le cadre de l'EMF, afin de vérifier la qualité des données et éviter les interprétations biaisées des résultats. En outre, l'approche pragmatique basée sur les dilutions d'échantillons pourrait servir comme outil de filtration des « vrais » biomarqueurs environnementaux.

Le troisième axe (troisième chapitre expérimental) visait à mettre en place une nouvelle méthodologie pour analyser les résidus volatils de pesticides complexes. Le suivi de ces résidus et l'étude de leur évolution au cours du temps servent comme information complémentaire pour l'évaluation du devenir des pesticides, ainsi que pour l'estimation de l'exposition du milieu à leurs xénométabolites potentiellement toxiques. Pour se faire, une conception d'un microcosme

de sol vivant a été mise en place, et une méthode automatisée d'extraction de l'espace de tête basée sur la micro-extraction en phase solide (HS-SPME) a été développée et optimisée. Elle a été couplée en ligne avec la GC-MS pour l'analyse des métabolites volatiles. Ce montage du complexe « échantillon-extraction-analyse » a assuré une méthodologie moins laborieuse, verte et non-destructive. Cela permet d'analyser les mêmes échantillons de sols pour plusieurs fois, ce qui diminue le cout et le temps nécessaire pour la préparation d'échantillons. Il limite également le biais « entre-échantillons » causé par la variabilité biologique. De plus, l'automatisation de la méthode permet d'éviter les biais analytiques liés au manipulateur. Suite à la mise en place de cette méthode, une étude cinétique de 38 jours a été menée pour la preuve du concept. Elle impliquait des sols pollués par un biopesticide typique et complexe et des contrôles non-pollués qui ont été analysés pour 8 points cinétiques (8 extractions effectuées sur les mêmes lots d'échantillons). Cette étude propose la métabolomique non-ciblée et ces outils computationnels et statistiques comme alternative pour étudier ce type des pesticides complexes. Elle a pu démontrer la capacité de la méthode et de la stratégie globale à expliquer l'évolution temporelle des profils métaboliques volatiles, leur dissipation, et à filtrer l'information pertinente permettant l'identification putative rapide de 96 xénométabolites, dont 63 signalés pour la première fois pour ce biopesticide, et 20 sous-produits de transformation. Ces identifications ont été faites grâce à la sélectivité avancée de la GC-MS, permettant de calculer les Indices de Rétention de Kováts et la recherche des spectres EI-MS dans les bases de données spectrales. Par ailleurs, les outils chimiométriques et statistiques ont permis d'estimer une reproductibilité et une sensibilité de la méthode qui ont été considérées acceptables.

En conclusion, une avancée significative a été apportée à l'approche « EMF ». Elle a été consolidée pour les applications en laboratoire et sur le terrain, en abordant des différentes problématiques étudiées sur plusieurs types de pesticides et de matrices, ce qui prouve son aspect « universel ». Néanmoins, de nombreux points restent à évaluer et développer, tels que l'optimisation des plans d'expériences des larges études cinétiques, l'étude profonde de la sensibilité de l'approche, et l'évaluation des modèles statistiques convenables pour la détermination du « temps de résilience ».

Development of a novel universal proxy to assess the environmental fate and impact of complex (bio)pesticides by Mass Spectrometry-based Metabolomics

Table of Contents

List of Publications and Communications

1. Research articles

Ghosson, H.; Guitton, Y.; Ben Jrad, A.; Patil, C.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. **(2021)**. Electrospray ionization and heterogeneous matrix effects in liquid chromatography/mass spectrometry based meta-metabolomics: A biomarker or a suppressed ion? *Rapid Communications in Mass Spectrometry*, 35(2):e8977. doi:10.1002/rcm.8977.

Ghosson, H.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. **(2020)**. Online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted volatile metabolomics for studying emerging complex biopesticides: A proof of concept. *Analytica Chimica Acta*, 1134:58-74. doi:10.1016/j.aca.2020.08.016.

Ghosson, H.; Brunato, Y; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Liquid Chromatography-High Resolution Mass Spectrometry-based untargeted profiling as a tool for analytical development: Assessment of novel extraction protocols for herbicide-polluted soil meta-metabolomics. *In-preparation*.

2. Oral presentations

Ghosson, H.; Brunato, Y.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. In-depth assessment of exhaustive extraction protocols for pesticide-polluted soil meta-metabolomics by LC-HRMS-based untargeted metabolic profiling (Oral communication). *1ères journées scientifiques numériques du RFMF 2020 (1JSRFMFnumériques)*. GoToMeeting, Online. **(2020, June 25-26)**. ⟨hal-03149337⟩.

Ghosson, H.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. HS-SPME-GC-MS-based untargeted metabolomics for kinetics tracking of natural herbicides' volatile residues: can we "footprint" the "Volatilome"?? (Oral communication). *Journée Métabolomique en Occitanie (R-Oc-Me)*. Perpignan, France. **(2019, October 11)**. ⟨hal-02706143⟩.

Ghosson, H.; Patil, C.; Ben Jrad, A.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Electrospray Ionization and co-elution in Meta-metabolomics: a biomarker or a suppressed ion? (Oral communication). *XXIV^{èmes} Rencontres du Club Jeune de la Société Française de Spectrométrie de Masse (RCJSM2019)*. Saint-Pierre-Quiberon, France. **(2019, March 18-22)**. ⟨hal-02707111⟩.

Patil, C.; Ben Jrad, A.; **Ghosson, H.**; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Metabolic foot-printing approach to assess the environmental fate and impact of bioinsecticides (Oral communication). *Natural Products & Biocontrol 2018 (Biocontrol2018)*. Perpignan, France. **(2018, September 25-28)**.

Patil, C.; Ben Jrad, A.; **Ghosson, H.**; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Study of the environmental impact of *Bti vs.* chemical insecticides by metabolic foot-printing approach (Oral communication). *11èmes journées scientifiques du RFMF 2018 (11JSRFMF)*. Liège, Belgium. **(2018, May 23-25)**.

Patil, C.; Ben Jrad, A.; **Ghosson, H.**; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Mapping the Bti triggered environmental metabolic landscape (Oral communication). *Workshop on Environmental and socioeconomic effects of Bti mosquito control*. Kurhaus Annweiler, Germany. **(2018, March 26-28)**.

3. Posters

Ghosson, H.; Patil, C.; Ben Jrad, A.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Electrospray Ionization and samples complexity in Meta-metabolomics: a biomarker or a suppressed ion? (Poster communication). *European RFMF Metabomeeting 2020 (MetaboEU2020)*. Toulouse, France. **(2020, January 22-24)**. ⟨hal-02456148⟩.

Ghosson, H.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Headspace-Solid Phase Micro Extraction-Gas Chromatography-Quadrupole Mass Spectrometry-based metabolomics for kinetics tracking of natural herbicides' volatile residues: a simple non-destructive method (Poster communication). *12èmes journées scientifiques du RFMF 2019 (12JSRFMF)*. Clermont-Ferrand, France. **(2019, May 21-23)**. ⟨hal-02711869⟩.

Ghosson, H.; Suryawanshi, R.; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Introducing High Resolution Mass Spectrometry in the Environmental Metabolic Footprinting metabolomics approach; application on extracted natural herbicide "Myrigalone A" (Poster communication). *Natural Products & Biocontrol 2018 (Biocontrol2018)*. Perpignan, France. **(2018, September 25-28)**. ⟨hal-02712543⟩.

Patil, C.; Ben Jrad, A.; **Ghosson, H.**; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Study of the environmental impact of insecticides by metabolomic foot-printing approach - An introduction of the concept (Poster communication). *EuroScience Open Forum (ESOF 2018)*. Toulouse, France. **(2018, July 09-14)**.

Ghosson, H.; Lazarus, M.; Raviglione, D.; Patil, C.; Salvia, M.-V.; Bertrand, C. Untargeted metabolic profiling of environmental volatile compounds by Headspace-Solid Phase Micro Extraction-Gas Chromatography-Mass Spectrometry: an introduction to environmental "volatilomics" (Poster communication). *11èmes journées scientifiques du RFMF 2018 (11JSRFMF)*. Liège, Belgium. **(2018, May 23-25)**. ⟨hal-02713060⟩.

Patil, C.; Ben Jrad, A.; **Ghosson, H.**; Raviglione, D.; Salvia, M.-V.; Bertrand, C. Study of the environmental impact of insecticides by metabolic foot-printing approach - An introduction to the concept (Poster communication). *5th MCAA Annual Conference and General Assembly*. Leuven, Belgium. **(2018, February 02-03)**. ⟨hal-02713858⟩.

Introduction

"State of the Art"

The State of the Art

1. Pesticides: general aspects

According to The Food and Agriculture Organization of the United Nations (FAO), pesticide is *"any substance or mixture of substances intended for preventing, destroying or controlling any pest, including vectors of human or animal disease, unwanted species of plants or animals causing harm during or otherwise interfering with the production, processing, storage, transport, or marketing of food, agricultural commodities, wood and wood products, or animal feedstuffs, or which may be administered to animals for the control of insects, arachnids or other pests in or on their bodies. The term includes substances intended for use as a plant growth regulator, defoliant, desiccant, or agent for thinning fruit or preventing the premature fall of fruit, and substances applied to crops either before or after harvest to protect the commodity from deterioration during storage and transport"* [1].

Pesticides are thus agents that target a living organism in order to control it or eradicate it. They are in fact classified according to their targets: insecticides for insects, herbicides for weeds, fungicide for fungus, etc. Their use is essential for several anthropogenic and economic activities. For instance, herbicides are applied in order to enhance the yield and the quality of agricultural production, by eliminating competitive weeds that occupy arable areas and consume soil nutrients. Insecticides are used to eradicate or control various types of insects, as mosquito, aphid, or pine processionary. These insects can be vectors of human and animal diseases, phytopathogenic, or devastators of food and agricultural production. Fungicides and bactericides are mainly used for hygienic reasons, and for controlling devastators of food and agricultural production as well. As shown in Figure Int. 1, pesticides use in the World has increased by a fold of ≈ 2 during the last three decades (between 1990 and 2018), with a constant tendency that started to appear since 2011. For the European Union[1] and France, the use of pesticides seems to be constant with a slight decreasing tendency between 1990 and 2018 (Figure Int. 2). These products are still largely applied however. For instance, in 2018, the quantities of active ingredients of pesticides used in or sold to the agricultural sector for crops and seeds were 367794 tons and 85072 tons, which correspond to 3.14 Kg and 4.45 Kg per ha of arable land and land under permanent crops for E.U. and France, respectively [2].

[1] United Kingdom included.

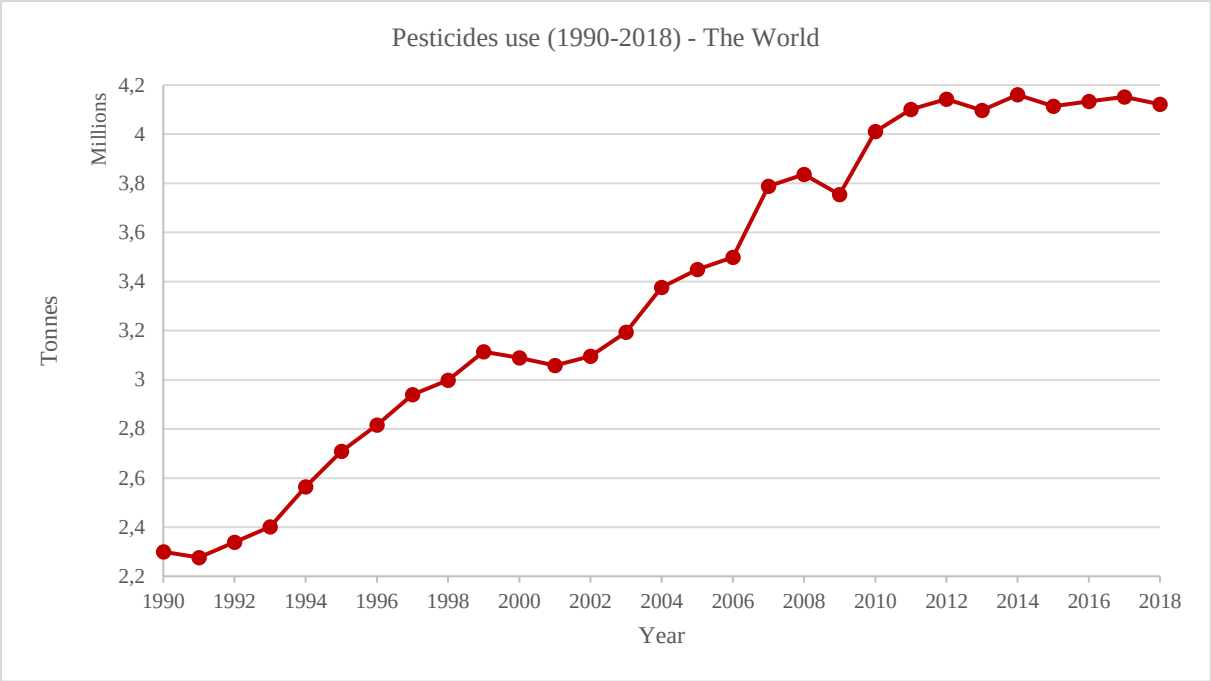
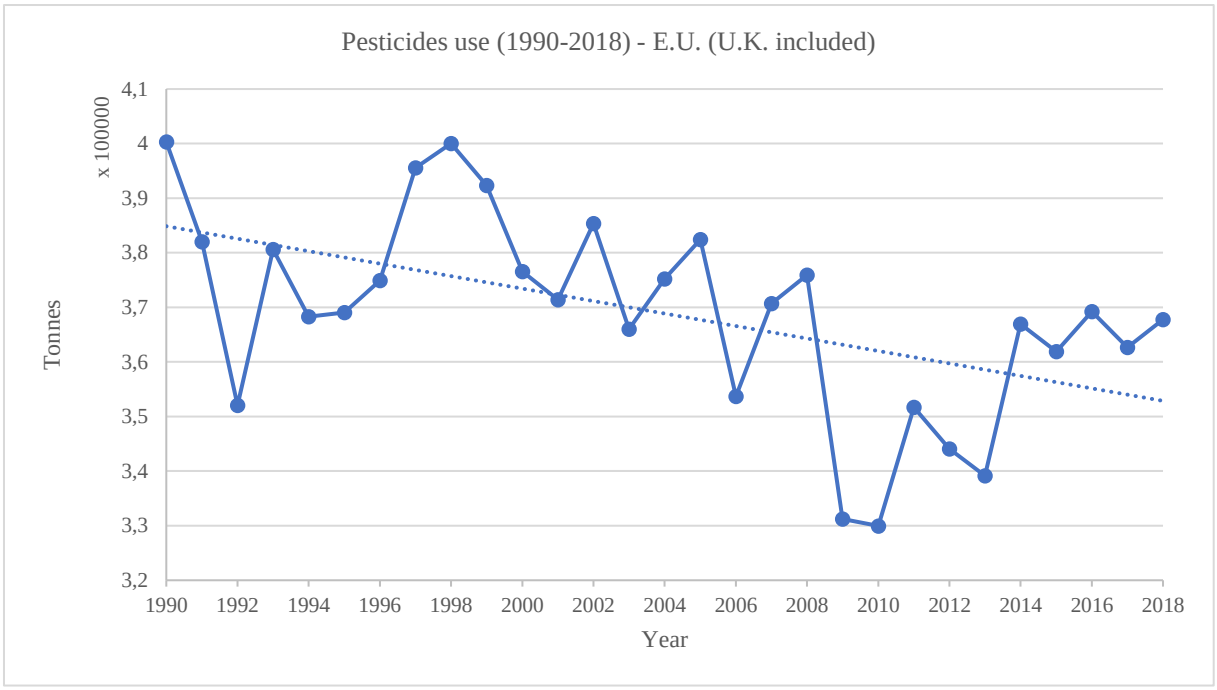


Figure Int. 1: Pesticides use in the World between 1990 and 2018.
The Pesticides Use database includes data on the use of major pesticide groups (Insecticides, Herbicides, Fungicides, Plant growth regulators and Rodenticides) and of relevant chemical families. Data report the quantities (in tons of active ingredients) of pesticides used in or sold to the agricultural sector for crops and seeds. Information on quantities applied to single crops is not available. Source: FAO Statistics [2].



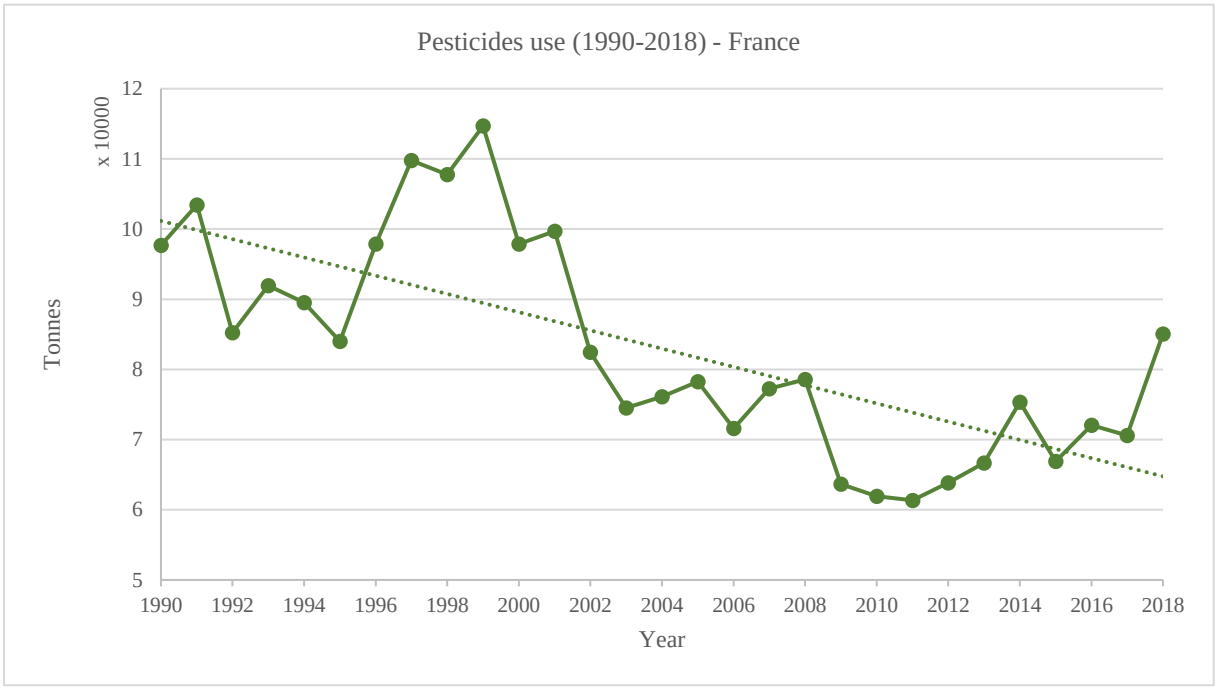


Figure Int. 2: Pesticides use in the E.U. and France between 1990 and 2018.
The Pesticides Use database includes data on the use of major pesticide groups (Insecticides, Herbicides, Fungicides, Plant growth regulators and Rodenticides) and of relevant chemical families. Data report the quantities (in tons of active ingredients) of pesticides used in or sold to the agricultural sector for crops and seeds. Information on quantities applied to single crops is not available. Source: FAO Statistics [2].

A pesticide consists of two main components: the active substance and the formulation ingredients.

The active substance assures the principal role of the pesticide, *i.e.* the regulation of the targeted organism. It is mainly consisting of one or several molecules that deliver the toxic action against the targeted organism. In principle, the toxic activity of this/those molecule(s) should be limited to the targeted organism [3].

On the other hand, the formulation ingredients consist of a mix of molecules and/or macromolecules that render the practical use of the pesticide optimal. For instance, they enhance the solubility of the active substance in water, or its stability in solution, and they may assure its effective physical delivery to the targeted organism in order to improve the performance of the product. The formulation ingredients are mostly polymer-based emulsifiers and surfactants. They are considered inert compounds that should not affect the targeted organism or the other organisms present in the environment [3].

2. Pesticides toxicity to non-target organisms

The use of pesticides is historically associated with health, environmental, and social concerns [4]. Different types of pesticides have been known for their impact on non-target organisms. This was proven by several studies and extended reviews [5–9].

Pesticides risks on human health are a subject for several open scientific debates. For instance, concerns about pesticides potential neurotoxic risks [10–12] and their role as potential endocrine disruptors [13–15] have been reported. Moreover, the correlation between the exposure to pesticides and cancer has been broadly debated in the middle of the scientific and research communities [16–19]. For instance, several pesticides are declared by the International Agency for Research on Cancer (IARC) as “potential carcinogenic agents” for human beings [20–22]. One popular example is the well-known Glyphosate herbicide [20]. Its carcinogenic effect is still a subject for a scientific and regulatory debate [23–25].

On the other hand, pesticides ecotoxicological risks have also been reported. For instance, pesticides impact on bees have been notably documented [26–35] and present a trendy topic in Ecology. Daphnia is well known for undergoing pesticides toxicity [5,36–38], as well as fishes [4,5], corals [39], and many other terrestrial and aquatic species [4,5,40]. Besides, diverse studies demonstrated that the formulation ingredients present in the commercial solutions of pesticides can accentuate the ecotoxicity of the active substance [37,41–45]. It is important to note that pesticides transformation products can also have non-negligible adverse effects on non-target organisms [46–51].

All these concerns banished the *a priori* claiming that pesticides only act on the targeted organisms. The impact of a pesticide can result from several and interlaced factors as its applied quantity (dose), its formulation, and the environment where it is applied. In fact, the impact and/or the toxicity are not only related to the active substance, but can also be issued from the type and the composition of the formulation ingredients, and the transformation pathways of the pesticide. These pathways are mainly determined by the environment and its dominating biosystems, but can also be affected by the applied concentration of the pesticide, and the type of its formulation ingredients.

Hence, with such facts alerting about the potential negative effects of pesticides, questions concerning the environmental impact arise. This impact is more expected in the environments that are highly exposed to pesticides application, mainly, the arable soil of the agricultural fields, and its microbial communities as bacteria, actinomycetes, fungi, algae, protozoa, and nematodes [52], as well as the sediments of lagoons and lakes, where insecticides reproduction takes place.

3. Importance of soil microorganisms for agriculture

Microorganisms of soil constitute one of the essential components that determine soil quality. They assure the decomposition and the re-elaboration of organic matter, and play an important role for plant health, as they can be pathogens and/or beneficial agents. The beneficial role of microorganisms for plants has been widely described and detailed in the literature. This role appears in various aspects of relations between microorganisms and plants.

For instance, several microbes and fungi assure Nitrogen and Phosphorus provision for plants through mutualistic relations with plant roots that give carbohydrates for the microbes and the fungi [53]. This role allows some microorganisms to be qualified as biofertilizers, as the commonly-known Azospirillum, Herbaspirillum, Acetobacter, Azotobacter and Azoarocus that assure the unavailable Nitrogen provision by fixing it from the atmosphere, or as the *Bacillus megaterium* as Phosphorus solubilizer, and *Bacillus mucilaginosus* as K solubilizer [54]. On another hand, microorganisms as *Pseudomonas trivialis* and *Burkholderia cepacia* can play the role of biopesticides by producing enzymes, nitric oxide, osmolytes, siderophores, organic acids or antibiotics to kill pathogens [55–57]. Others can be biostimulants that induce plants systemic resistance [58], as *Pseudomonas* spp. and *Bacillus* spp. [59]. Microbes as Azospirillum, Pseudomonas and Bacillus can help plants to resist biotic or abiotic stress as drought, salinity or metal toxicity [60,61]. Furthermore, several microorganisms as Azospirillum, Rhizobium,

Bacillus, Pseudomonas, Serratia, Stenotrophomonas, Streptomyces, Ampelomyces, Coniothyrium and Trichoderma promote plant growth [62] by producing specific phytohormones as auxins, cytokinins, gibberellins, abscisic acid, ethylene [63,64] or other secondary metabolites [60,65].

4. Impact of pesticides on soil microorganisms

As mentioned in the previous paragraph, soil microorganisms are essential for the assurance of soil quality and agricultural productivity for crops. However, pesticides can have adverse effects on these microbial populations. The impact of pesticides on soil microorganisms have been broadly documented by Stanley & Preetha [5]. They report that pesticides with specific mode of action are unlikely to be harmful or directly affecting soil microbes [66]. However, pesticides with pertinent action can directly harm those microorganisms by killing them, or reducing their population or their activities. They can also affect them by changing the biochemical and physiological attributes in soil or the food sources and chains. The reported effects can be summarized as the following; effects on: microbial biomass, microbial population and growth, microbial diversity, microbial community, microbial biomass carbon, soil/microbial respiration, microbial activities, microbial enzymatic activities, and pesticide resistance *vs.* metabolism by microbes. These effects were assessed by different experimental scales and designs as Laboratory/Microcosm, Semifield/Mesocosm, and Field studies. They must be assessed on community basis and not at an individual microorganism scale, as the toxic effects of pesticides can be significantly different when it comes to microbial communities (as within the biofilms [67]). It should be mentioned that toxic effects of pesticides transformation products on soil microbes have been also reported [46].

5. Emergence of "Biopesticides"

As described previously, pesticides adverse effects can deteriorate soil health and thus can have negative consequences on agricultural productivity. To limit the damage caused by those products, one of the proposed solutions is to limit their use and to find alternatives. This policy has been adopted in several countries and regions of the World. For instance, the French government initiated in 2008 the "Écophyto I" program that was updated by the "Écophyto II", and the "Écophyto II+" programs [68]. Those programs are a part of a European policy mainly defined by the Directive 2009/128/EC in order to establish a framework for community action to achieve the sustainable use of pesticides [69].

One of the major aims of the French plans is to reduce the use of "phytopharmaceuticals" (*i.e.* synthetic pesticides) by 50 % by 2025. This reduction should be achieved by accelerating the withdrawal of the most "risky" substances, including the Glyphosates that must be definitively withdrawn by 2022. Moreover, the reduction of phytopharmaceuticals must be accelerated by promoting recognition and dissemination of the "less risky" pesticides that originate from natural preparations.

Hence, one of the proposed solutions is to replace synthetic pesticides by pesticides originating from nature, *i.e.* the "biopesticides".

The definition of biopesticides can differ between various references and works. It can also interfere with the definition of "biocontrol agents". To avoid the confusion, the adopted definition of "biopesticides" will be based on the statements of Glare *et al.* [70]. The definition can be summarized as the following:

"Biopesticides are preparations containing living microorganisms such as bacteria, viruses, fungi, protozoa and nematodes and/or bioactive compounds (such as metabolites) produced directly from these microbes, which are used to suppress populations of pests, including insects, pathogens and weeds. Biopesticides can also be plant extracts and other naturally sourced materials as botanical compounds, essential oils and semiochemicals (e.g. pheromones). These products need to be repeatedly applied to the pest-infested areas because their populations are not self-sustaining for more than one or a few growing seasons, and they are not capable of spreading beyond the area of application" [70].

Biopesticides types, categories and examples were documented by Copping & Menn [71]. According to their review, these products can be divided into two principal categories: natural products, and microorganisms. Natural products consist of molecules or mixes of molecules originating from natural and/or biological sources. They can be divided into three sub-categories: microorganism-derived products, compounds derived from higher plants, and animal-derived products. On the other hand, microorganisms can be divided into five sub-categories: viruses, bacteria, fungi, nematodes, and protozoa.

It should be mentioned that beneficial macroorganisms and transgenic plants expressing plant protection compounds are not included in the above definitions and categories. They will not be considered for the present thesis.

Apart from their ambiguous definition, and despite their questionable efficiency and high costs, the biopesticides are increasingly emerging as competitive products (Figure Int. 3) that present a promising alternative solution to replace synthetic pesticides [70]. The main reason supporting

their emergence is that they are supposed relatively less persistent in nature, and less harmful for the health and the environment [70–75]. This latter hypothesis is supported by the fact that the majority of biopesticides target single pathogens [70,75]. Some exceptions can be given; mainly the active substances from natural products such as the Azadirachtin A and B insecticides [72], and the Nonanoic acid-based herbicides that attack cellular membranes, or other microorganisms-based products such as the Serenade® fungicide based on the *Bacillus subtilis* strain QST [70,76,77]. Thus, as biopesticides target single pathogens, their risks on non-target species are presumed to be narrow.

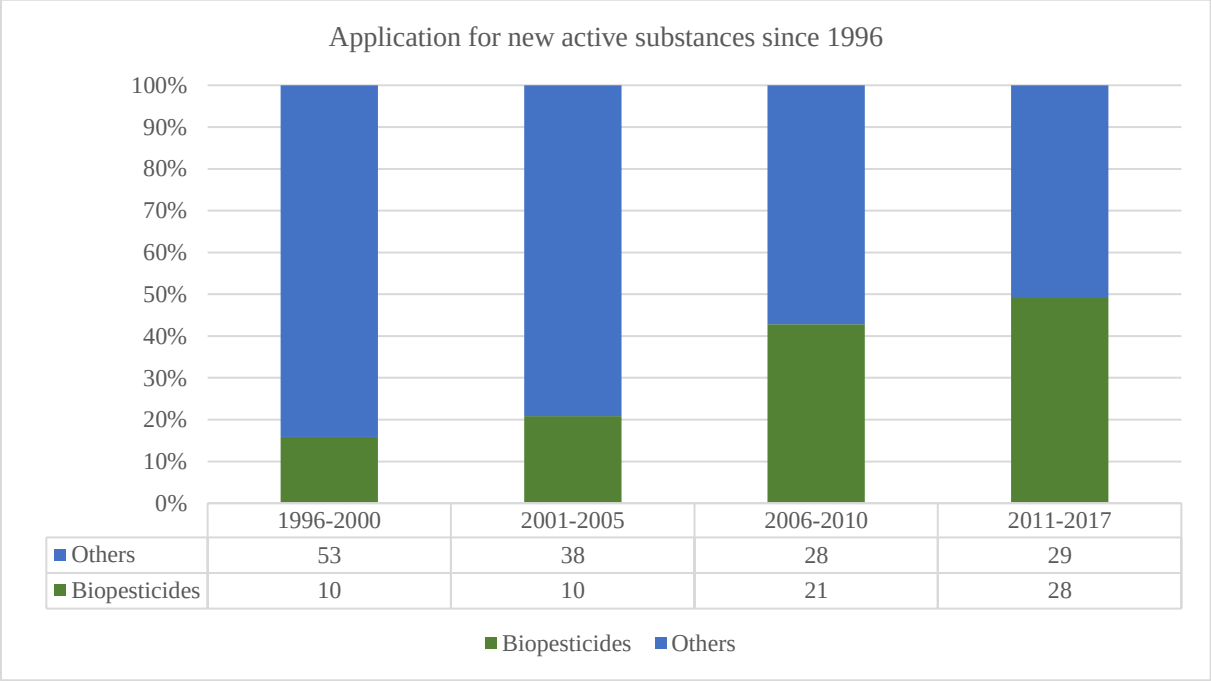


Figure Int. 3: Evolution of the number of applications for new biopesticides (active substances), compared to synthetic pesticides, in E.U. (Source: European Commission [78]).

Nevertheless, the *a priori* claiming that biopesticides are less persistent in the environment is questionable, particularly for living microorganisms. The hypothesis is still however not widely examined, mainly due to difficulties facing the analytical tools and methods. In fact, biopesticides are majorly complex products consisting of organelles, unicellular, or multicellular organisms, and/or contain diverse molecules and/or macromolecules that are partially non-characterized (this problematic will be evoked in details in the subsequent sections). On the other hand, questions concerning the toxic and ecotoxic risks of biopesticides were raised in the middle of scientific and regulation communities [79]. The given argument behind these questions is that precautions regarding the use of the novel emerging natural substances, their transformation products and their formulations should be taken in consideration in order to avoid their potential risks on human health and environment. In fact,

and as a confirmation of these concerns, recent safety auditing and research works proved that several biopesticides could present risks on non-target organisms. For instance, concerns were raised by the European Food Safety Authority (EFSA) regarding the potential toxicity of Spinosad insecticide (originating from *Saccharopolyspora spinose*) on human and animal health [80]. Azadiractine (an active ingredient of neem oil) was reported to be a potential endocrine disruptor [75]. On the other hand, direct negative effects of the anti-mosquito bioinsecticide *Bacillus thuringiensis* subsp. *israelensis* on non-target species as Chlorophyta, Diptera, Lepidoptera and Plecoptera were observed in laboratory experiments [81]. Besides, recent preliminary studies investigated a potential impact of the Leptospermone (a natural β-Triketone herbicide) on soil microbial communities [82,83].

Therefore, measures regarding risk assessment of these emerging products should be considered as an essential step that determines the authorization of the use of such products. However, the applied regulations and risk assessment guidelines are still limited to the same regulations and protocols dedicated for the synthetic pesticides. For instance, in the E.U., the Regulation (EC) No 1107/2009 [84] is so far the determinant text that regulates the authorization of the use of biopesticides' active substances [72,75]. However, several issues regarding this regulation have been reported. Indeed, there is an ambiguous and non-harmonized definition of biopesticides on the E.U. scale [72,75]. This issue caused several practical difficulties regarding the achievement of the required trials that aim to assess the risks in order to obtain the marketing authorization, particularly, for the microorganisms category [75,85]. After, the regulation has evolved by emitting specific guidelines for the different categories [86–89]. Nevertheless, the major remaining issue is that the guidelines are so far based on the classical protocols that were implemented basically to assess separately the fate and the impact of the synthetic pesticides [72,75]. However, as the biopesticides consist of products with different and more complex biological-chemical natures, the present adopted guidelines for their risk assessment seem to be insufficient and non-adapted.

6. Regulations on pesticides in the E.U.: an overview

The principal law that organizes pesticides use in the E.U., and defines the procedures that allow the acquisition of marketing authorization for a Plant Protection Product (PPP) (*i.e.*, a pesticide) is the Regulation (EC) No 1107/2009 [84]. It is applicable either on synthetic pesticides, or on biopesticides, as mentioned previously. In brief, the application for the use of a novel active substance should first start by its evaluation by the applicant of the authorization (mostly the

manufacturer). The evaluation should respect the guidelines defined by the European Commission [90] (that will be described in brief subsequently). After, the dossier should be submitted to a Member State of his choice (the Rapporteur Member State; RMS). The RMS notifies the other member states, the European Commission and the European Food Safety Authority (EFSA). The RMS shall prepare a draft assessment report, assessing whether the active substance can be expected to meet the approval criteria. Then, the EFSA conducts an in-depth assessment of the dossier. Following the assessment report issued by the EFSA, the European Commission takes the decision regarding the approval (or the rejection) of the marketing authorization of the active substance in query [91]. If approved, the active substance can be classified in one of the four classes defined by the Regulation (EC) No 1107/2009: 1) a "base active substance" with an unlimited-time approval, 2) a "low-risk active substance" approved for a maximum duration of 15 years, 3) a "standard active substance" approved for a maximum duration of 10 years, or 4) an "active substance considered for substitution" approved for a maximum duration of 7 years [84].

The Regulation (EC) No 1107/2009 also defines the criteria and the indicators that should be examined in order to approve and classify the active substance. The European Commission provides guidelines to define those criteria and indicators, and to allow their assessment [90]. Among all the required information and studies for the evaluation; the accumulation, the transformation and the persistence of the PPP and its residues should be investigated in several different matrices as plants, plant products, foodstuffs (of plant and animal origin), feedingstuffs, soil, water, air, body fluids and tissues [92]. The persistence and the dissipation of the active substances are based on the DT50 and DT90. The DT50 is defined as the "disappearance time of 50 % of the substance" (DT90 is for 90 % of the substance) [93]. The disappearance is defined as "*processes that result in transformation, degradation and eventually mineralization of the substance, including microbial degradation, chemical hydrolysis, and photochemical reactions, or other processes, such as leaching, volatilization and uptake by plants*" [93]. In soil, the tolerated laboratory experiments DT50 is < 60 days at a temperature equal to 20 °C (< 90 days if the temperature is 10 °C – if the active substance must be used in cold areas). In the field, the tolerated DT50 is < 3 months and the tolerated DT90 is < 1 year. The field experiments must be carried out on 4 different soil types. It is worth mentioning that the DT50/DT90 indicators does not distinguish between the chemical degradation mechanisms (*e.g.* mineralization, microbial degradation, chemical hydrolysis, or photochemical reactions), and the loss by mechanisms as leaching, volatilization, or uptake by

plants. The DT50/DT90 values assess the whole dissipation of the PPP including both mechanisms. In order to exclusively target the degradation of the compound, another parameter was included in the guidelines: the "DegT50" [94,95]. In addition, the adsorption of the active substance on soil should also be assessed [95]. Its measuring indicators are the "K_d", *i.e.*, the soil/water distribution coefficient[2], or the "K_{oc}": the soil organic Carbon adsorption coefficient, *i.e.*, the K_d standardized to the organic Carbon percentage in soil[3] [96]. Those indicators are essential to determine the mobility of the substance in soil, and thus the potential pollution by leaching that they may engender to groundwater.

The recommended methods and the criteria for methods validation and for the classification of PPP transformation products (*e.g.* metabolites) are also defined by the European Commission [97]. To study the transformation and the dissipation of the products and its residues, targeted multi-residues methods are recommended. The Mass Spectrometry-based methods are also recommended for those experiments [97].

The toxicity and the ecotoxicity should also be assessed on different species. For instance, the terrestrial species that should be examined are summarized in the Guidance Document on Terrestrial Ecotoxicology Under Council Directive 91/414/EEC (SANCO/10329/2002 rev 2 final) [98]. They include soil microorganisms. The defined indicator to assess the ecotoxicity of a PPP on soil microorganisms is the microbial activity that should not be affected by > ±25 %. It should be mentioned that the toxicity and ecotoxicity tests should be carried out on the active substance (*a priori* risk assessment) and on the formulation ingredients (*a posteriori* risk assessment). The formulation ingredients should be tested particularly when a "homologation" of the PPP is demanded. The "homologation" is the use of the same active substance in order to control a different pathogen.

7. Limitations of the existing methodologies

As described previously, the classic methodologies and concepts based on the DT50/DT90 only target defined chemical compounds, *i.e.*, the active substance and the known transformation products. From an analytical chemistry point of view, these approaches can present limitations when it comes to the emerging alternatives, *i.e.*, the biopesticides.

[2] (Concentration of the chemical substance in soil/concentration of the chemical substance in water)

[3] $K_{oc} = \left[\left({K_d}/{\%_{(oc)}}\right) \times 100\right]$

In fact, some of those products are complex mixtures consisting of various known and unknown molecules. These molecules can act in a synergic and/or pleotropic mode of actions. An example can be given for the bioherbicide based on the *Myrica gale* methanolic extract [99]. This bioherbicide contains a wide variety of compounds that were partially identified [99–101]. Among several others, one major active compound is the "Myrigalone A". However, this active compound has been proven as more efficient when applied within the integral extract mixture [102]. This mixture contains several known and unknown polyphenolic compounds that help protecting the Myrigalone A from being rapidly degraded by photolysis [103]. Hence, as the DT50/DT90 approaches are based on targeted known molecules, they are not applicable when the composition of the PPP is partially unknown. Furthermore, the transformation products that issue from such complex mixtures are complex to be identified.

On the other hand, the DT50/DT90 approaches are not applicable when it comes to biopesticides based on microorganisms. In fact, this type of biopesticides is not based on defined molecular active substances. The quantification of their active compounds and their transformation by-products are thus not possible. This problematic has been in fact raised during the OECD 9th Biopesticides Expert Group Seminar on Test Methods for Microorganisms. The report issued following the seminar indicated that the microorganisms-based biopesticides are regulated following the same guidelines and approaches that were basically dedicated for chemical pesticides. Thus, it acknowledges the limitation of such approaches for this type of biopesticides, as their nature and mechanisms of actions are different. The report also considers that these approaches are outdated as the recent scientific advances have carried out novel techniques and proxies that can be exploited for microorganisms-based pesticides risk assessment. Therefore, new alternatives are recommended for this task [104].

8. Metabolomics: an alternative tool?

Metabolomics is a branch of Analytical Chemistry that considers studying "small molecules" with molecular mass < 1000-1500 Da. This branch is still increasingly growing and advancing since more than two decades (according to PubMed database [105], the first paper holding the term "metabolome" was in 1998 by H. Tweeddale and her colleagues [106]). Metabolomics has been developed and applied in a wide range of fields that includes medicine, pharmacology, toxicology, forensics, nutrition and food security, livestock, marine biology, plant biology, agriculture, and environment. Its approaches, and particularly the untargeted metabolomics, has

been proven capable for high throughput analyses that provide a big amount of data, using performant analytical chemistry techniques, relatively short-time analyses and computational data processing tools.

The involvement of metabolomics in the environmental research is increasingly considered in the recent past years [107–112]. The main aims of environmental metabolomics is to study the abiotic stress, the anthropogenic influence, and the impact of environmental diseases on nature and its existing species [108]. Environmental biomarkers and the modifications in the metabolic pathways are searched following factors as temperature or acidity change in the studied environment, or following the pollution of the environment by chemicals or its invasion by intruder species. Nonetheless, the environmental metabolomics is still a "minor" interest if compared to the global state of metabolomics research. Figure Int. 4 shows that between 2010 and 2020, papers holding both terms "environmental" and "metabolomics" did not surpass 18.50 % of the total number of papers holding the term "metabolomics" (the maximum reached was 18.42 % in 2018) [105,113]. In fact, environmental metabolomics still confront several challenging problematics as sample designing, extraction protocols, data processing and the significant lack of environmental metabolome databases [108].

On the other hand, untargeted metabolomics applied for pesticides research was also reported [114]. It is mostly targeting the elucidation of the modes-of-action of bioactive compounds, ecotoxicological and toxicological risks assessment, and the investigation of disruptions in plant metabolic pathways as response to the application of pesticides.

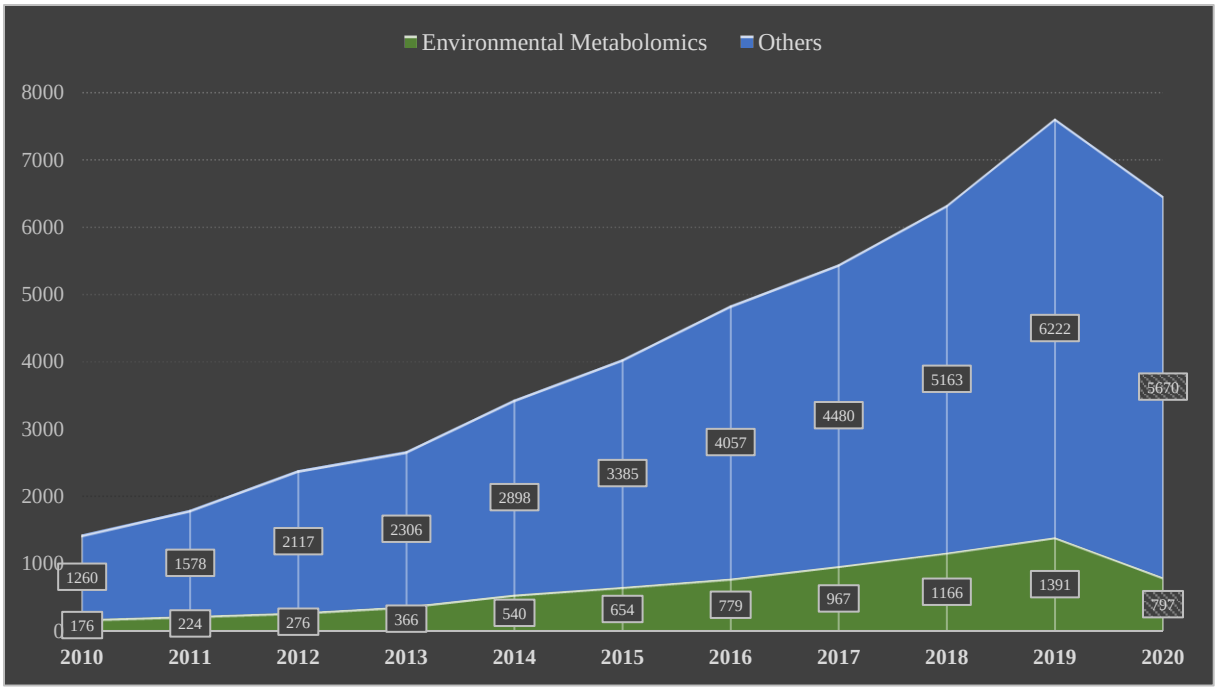


Figure Int. 4: Evolution of the numbers of articles holding the term "metabolomics", or both terms "environmental" and "metabolomics" between 2010 and 2020, according to PubMed database (October 11, 2020) [105,113].

Hence, as the untargeted metabolomics seems to be a promising tool that can provide high throughput generation of biochemical information, it was recently considered in order to suggest a novel "universal" tool for assessing both the environmental fate and impact of complex (bio)pesticides in the environmental matrices. The suggested approach was called "Environmental Metabolic Footprinting" (EMF), introduced in 2016-2017 by Patil *et al.* [115] and Salvia *et al.* [116].

9. Environmental Metabolic Footprinting: the concept, objectives, and challenges

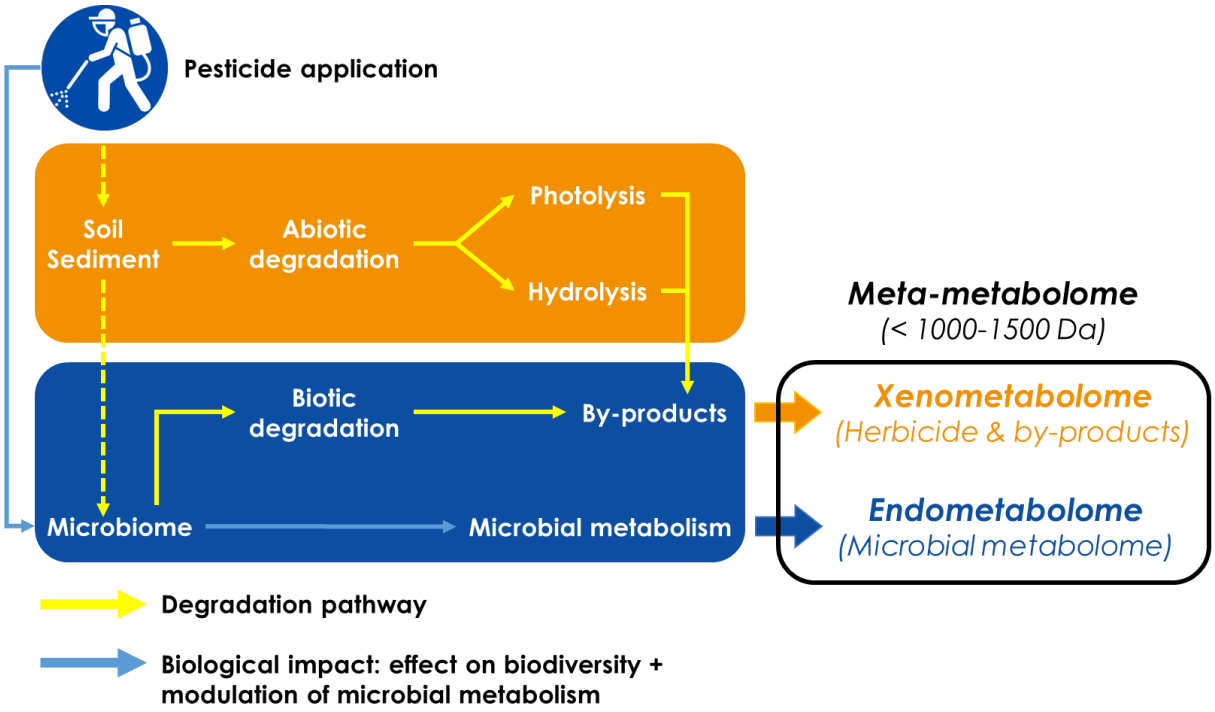


Figure Int. 5: The Environmental Metabolic Footprinting concept.
Figure readapted from Patil *et al.* [115], with permission from authors.

As shown in Figure Int. 5 the EMF is based on a meta-metabolomics approach (*i.e.* metabolomics of a whole community [117,118]) that aims to extract, analyze and detect both the xenometabolome of an applied pesticide, and the endometabolome of the environmental matrix. The xenometabolome consists of the active substance, the formulation ingredients, and the transformation products issued either from the active substance or from the formulation ingredients. The endometabolome consists of metabolites produced by the different microorganisms living in the studied environmental matrix, as the primary metabolites and the secondary metabolites. The xenometabolome and the endometabolome will then constitute the meta-metabolome that will be the target of the extraction, the chemical analyses, and the data processing.

The EMF seeks to investigate integrally the persistence, the transformation and the impact of an applied pesticide to an environmental matrix. As an untargeted approach, it can consider either known compounds or unknown compounds. Thus, it must be applicable for both the synthetic pesticides and the biopesticides. In addition, as it aims to analyze the whole meta-metabolome, the fate and the impact of the complex formulated products or microorganisms-based products can be studied. The EMF should also detect **new transformation by-products**,

and determine **new biomarkers of impact**. This can be done by comparative kinetics experiments, where the polluted (spiked) matrix is compared to an unpolluted control matrix.

The ultimate aim of EMF is to introduce a **novel indicator** called the **"resilience time"**. This indicator is defined as the time needed for the difference between the meta-metabolome of the spiked matrix and that of the control matrix to be statistically non-significant (Figure Int. 6). In this case, the biochemical state of the environmental matrix and its community is considered as normalized, as it is comparable to the control matrix, *i.e.*, the "natural" state of the matrix. Thus, at the resilience time, the whole xenometabolome including all pollutants issued from the pesticide are considered dissipated (or more precisely, below the detection limit of the approach), and the observable impact of the pollution is surpassed.

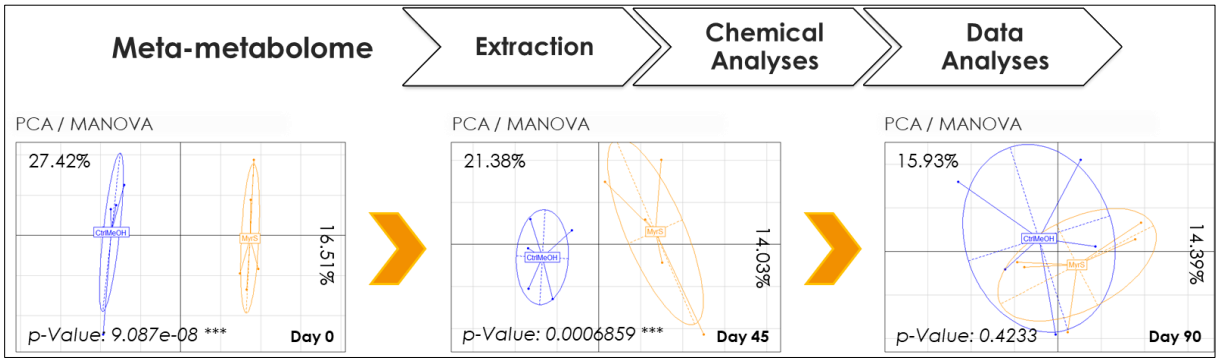


Figure Int. 6: The "resilience time" indicator.
The present example shows that the resilience time is reached between day 45 and day 90.

The previously mentioned research works that introduced the EMF aimed to prove the concept of the approach. The first by Patil *et al.* [115] was carried out on soil microcosms. It investigated the fate and the impact of two pure β-Triketone herbicides by 90-days kinetics: the Sulcotrione as a synthetic product, and the Leptospermone as its natural alternative. The second work by Salvia *et al.* [116] proved the applicability of the EMF on a complex microorganism-based bioinsecticide: the *Bacillus thuringiensis israelensis* (Bti). The study was performed on two kinetics time points after the application of the pesticide on sediment microcosm samples. It showed a significant statistical difference between control and spiked meta-metabolomes after 8 days of the spiking. This approach seems sufficiently sensitive for such microorganisms-based pesticides.

These two studies were proofs of concept that allowed going forward with the EMF approach. Nevertheless, several challenging problematics and questions are confronting this novel strategy.

The starting step is being the **extraction of the meta-metabolome**. In fact, wider is the extracted meta-metabolome; wider is the collected information on the fate and impact.

On the other hand, **performant analytical methods** should be set-up and adapted to analyze such complex samples. The analytical technologies that will be considered for development are the Liquid Chromatography (LC) coupled to the Electrospray Ionization-High Resolution Mass Spectrometry (ESI-HRMS), and the Gas Chromatography (GC) couples to the Electron Impact-Mass Spectrometry (EI-MS). The Chromatography is the common technology used to separate the compounds of a complex mixture. It helps detecting, quantifying, characterizing and enhancing the sensitivity of the metabolites detection. The Mass Spectrometry is a powerful technology that allows the detection, the characterization and the quantification of metabolites after their ionization, based on their mass-to-charge ratios (*m/z*). The HRMS is able to measure the exact *m/z* ratios with high precision. This precision allows determining metabolites elemental compositions with high reliability, which improves the selectivity within the analyzed meta-metabolic profile and thus enhances the reliability of the information. Tandem Mass Spectrometry (MS/MS) experiments are an available option for LC-HRMS analyses. As for the EI fragmentation, they allow identifying the metabolites by structural elucidation and spectral data search, which afford a tool for characterizing environmental biomarkers and pesticides transformation by-products.

Besides, such advanced analytical methods provide big and complex data. To deal with their complexity, **computational data preprocessing, chemometrics, and statistical data analyses** should be applied. These tools need however to be developed, adapted and assessed for the targeted question, mainly, to reliably determine the resilience time, and to differentiate between reliable biomarkers and analytical artefacts.

For all these objectives and problematics, the present thesis will be concentrated on three main axes (Figure Int. 7).

The first, exhibited in Chapter I, is to develop **novel extraction protocols for the EMF**. The development will be simultaneously performed with the set-up of **adapted LC-HRMS/MS methods**, and the introduction of **a novel chemometric approach** that helps assessing the optimal extraction protocol.

The second axis exhibited in Chapter II will be focusing on the determination and the characterization of **environmental biomarkers**. It will evoke the problematics of heterogeneous sample complexities, and its impact on ESI. This problematic issue may lead to determine false “biomarkers” due to the **Ion Suppression** phenomenon. A pragmatic approach will be then exhibited in order to avoid such false positives.

The third axis (Chapter III) will introduce a novel EMF-based methodology dedicated to analyze the **volatile residues** of complex herbicides applied on soil. The methodology will be based on a green, non-destructive **Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted approach**. It will be complementary to the EMF and will consider the volatile xenometabolome. The methodology will also establish the basis of **the chemometric and statistical analyses** that allow prioritizing the relevant pesticide substances and their transformation by-products, as well as determining the dissipation time and the sensitivity of the method.

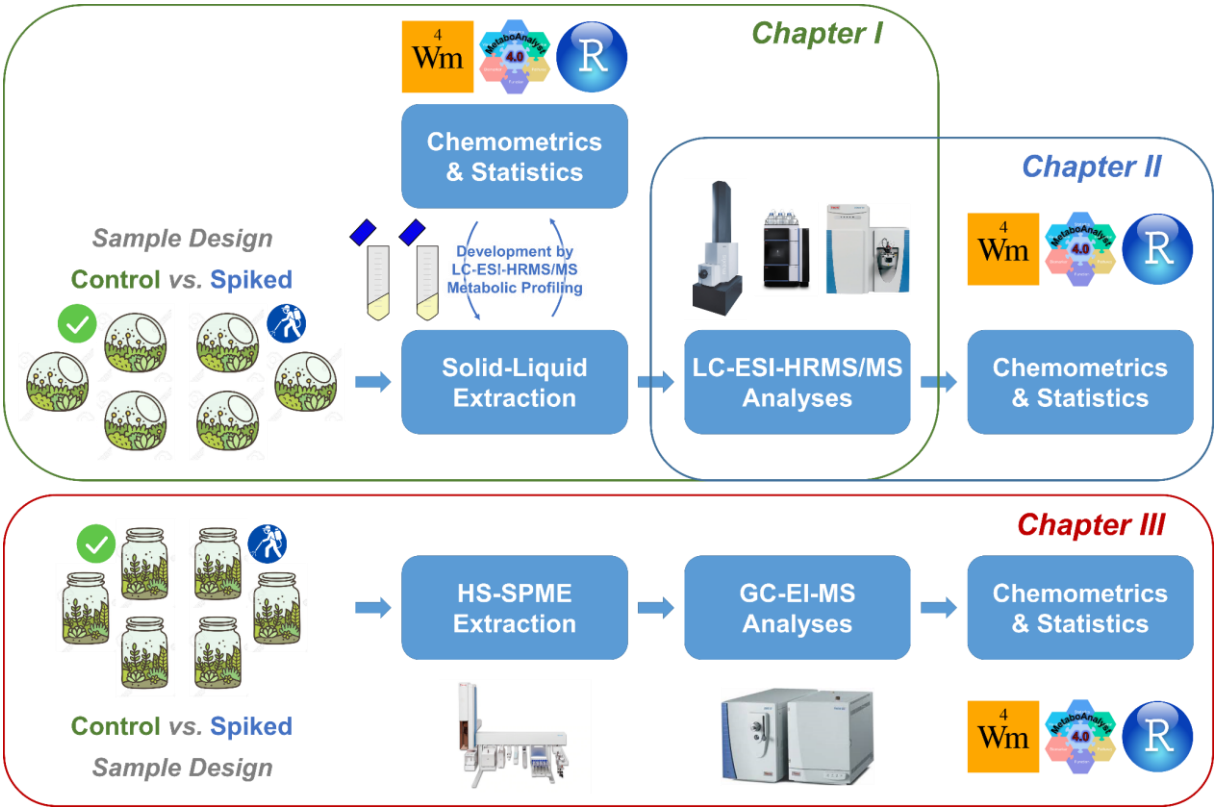


Figure Int. 7: The three axes that will be addressed in the present thesis.

It should be mentioned that the three different axes were studied in a context of different supporting projects. These projects are interested in different types of phytochemicals, matrices and experimental designs. Nonetheless, all of them have in common the need for the development and the application of the EMF. Therefore, this multi-actioners and multi-projects environment provided a suitable collaborative context for the research work that led to the achievement of the current thesis.

The first axis was supported by the European Regional Development Fund (ERDF) **Interreg POCTEFA PALVIP** European project [119]. This project aims 1) to evaluate the efficiency of novel biopesticides, 2) to study their modes of action and 3) to assess their environmental

dissipation and impact. Those products are mainly novel bioherbicides applied on soil and novel fungicides applied on plants, fruits and fruits products. The experiments are performed mainly in field and in the laboratory as well.

The second axis was held in the context of the Marie Skłodowska-Curie Actions **EnvFate** European project [120] that aims to assess the environmental fate and impact of the Bti bioinsecticide in sediments. Its main experiments are performed in the laboratory.

The third axis was supported by the **Interreg POCTEFA PALVIP** project and the French National Research Agency (ANR) **TRICETOX** national project [121]. This latter is interested in assessing the fate and the impact of both the synthetic and the natural β-Triketone herbicides applied on soil. Its main experiments are performed in the laboratory.

List of Abbreviations

ANR: Agence Nationale de la Recherche

Bti: *Bacillus thuringiensis israelensis*

EC: European Commission

EFSA: European Food Safety Authority

EI: Electron Impact

EMF: Environmental Metabolic Footprinting

ERDF: European Regional Development Fund

ESI: Electrospray Ionization

E.U.: European Union

FAO: The Food and Agriculture Organization of the United Nations

GC: Gas Chromatography

HRMS: High Resolution Mass Spectrometry

IARC: International Agency for Research on Cancer

LC: Liquid Chromatography

MS/MS: Tandem Mass Spectrometry

MS: Mass Spectrometry

OECD: Organization for Economic Co-operation and Development

PPP: Plant Protection Product

RMS: Rapporteur Member State

U.K.: United Kingdom

References

[1] The Food and Agriculture Organization of the United Nations (FAO), Guidelines for Legislation on the Control of Pesticides, United Nations, Rome, Italy, 1989. http://www.fao.org/fileadmin/templates/agphome/documents/Pests_Pesticides/Code/Old_guidelines/LEGIS.pdf (accessed October 1, 2020).

[2] FAO Statistics, The Food and Agriculture Organization of the United Nations (FAO). (n.d.). http://www.fao.org/faostat/en/ (accessed October 1, 2020).

[3] World Health Organization (WHO), The Food and Agriculture Organization of the United Nations (FAO), The International Code of Conduct on Pesticide Management, United Nations, Rome, Italy, 2014. http://www.fao.org/agriculture/crops/thematic-sitemap/theme/pests/code/en/ (accessed October 1, 2020).

[4] D. Pimentel, 'Environmental and Economic Costs of the Application of Pesticides Primarily in the United States,' Environ Dev Sustain. 7 (2005) 229–252. https://doi.org/10.1007/s10668-005-7314-2.

[5] J. Stanley, G. Preetha, Pesticide Toxicity to Non-target Organisms, Springer, Dordrecht, Netherlands, 2016. https://doi.org/10.1007/978-94-017-7752-0 (accessed October 4, 2020).

[6] G.W. Ware, Effects of pesticides on nontarget organisms, in: F.A. Gunther (Ed.), Residue Reviews, 1st ed., Springer, New York, NY, 1980: p. 218. https://doi.org/10.1007/978-1-4612-6107-0_9 (accessed October 4, 2020).

[7] T.M. Smith, G.W. Stratton, Effects of synthetic pyrethroid insecticides on nontarget organisms, in: F.A. Gunther (Ed.), Residue Reviews, 1st ed., Springer, New York, NY, 1986: p. 151. https://doi.org/10.1007/978-1-4612-4934-4_4 (accessed October 4, 2020).

[8] M.H. Fulton, P.B. Key, Acetylcholinesterase inhibition in estuarine fish and invertebrates as an indicator of organophosphorus insecticide exposure and effects, Environ Toxicol Chem. 20 (2001) 37–45. https://doi.org/10.1002/etc.5620200104.

[9] C. D. Milam, J. L. Farris, J. D. Wilhi, Evaluating Mosquito Control Pesticides for Effect on Target and Nontarget Organisms, Archives of Environmental Contamination and Toxicology. 39 (2000) 324–328. https://doi.org/10.1007/s002440010111.

[10] J.R. Richardson, V. Fitsanakis, R.H.S. Westerink, A.G. Kanthasamy, Neurotoxicity of pesticides, Acta Neuropathol. 138 (2019) 343–362. https://doi.org/10.1007/s00401-019-02033-9.

[11] M. Bjørling-Poulsen, H.R. Andersen, P. Grandjean, Potential developmental neurotoxicity of pesticides used in Europe, Environ Health. 7 (2008) 50. https://doi.org/10.1186/1476-069X-7-50.

[12] M.C. Keifer, J. Firestone, Neurotoxicity of Pesticides, Journal of Agromedicine. 12 (2007) 17–25. https://doi.org/10.1300/J096v12n01_03.

[13] Y. Combarnous, Endocrine Disruptor Compounds (EDCs) and agriculture: The case of pesticides, Comptes Rendus Biologies. 340 (2017) 406–409. https://doi.org/10.1016/j.crvi.2017.07.009.

[14] W. Mnif, A.I.H. Hassine, A. Bouaziz, A. Bartegi, O. Thomas, B. Roig, Effect of Endocrine Disruptor Pesticides: A Review, International Journal of Environmental Research and Public Health. 8 (2011) 2265–2303. https://doi.org/10.3390/ijerph8062265.

[15] M. Lacasaña, I. López-Flores, M. Rodríguez-Barranco, C. Aguilar-Garduño, J. Blanco-Muñoz, O. Pérez-Méndez, R. Gamboa, S. Bassol, M.E. Cebrian, Association between organophosphate pesticides exposure and thyroid hormones in floriculture workers, Toxicology and Applied Pharmacology. 243 (2010) 19–26. https://doi.org/10.1016/j.taap.2009.11.008.

[16] K.Z. Guyton, D. Loomis, Y. Grosse, F. El Ghissassi, L. Benbrahim-Tallaa, N. Guha, C. Scoccianti, H. Mattock, K. Straif, Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate, The Lancet Oncology. 16 (2015) 490–491. https://doi.org/10.1016/S1470-2045(15)70134-8.

[17] M.C.R. Alavanja, Pesticides and Lung Cancer Risk in the Agricultural Health Study Cohort, American Journal of Epidemiology. 160 (2004) 876–885. https://doi.org/10.1093/aje/kwh290.

[18] M.C.R. Alavanja, Use of Agricultural Pesticides and Prostate Cancer Risk in the Agricultural Health Study Cohort, American Journal of Epidemiology. 157 (2003) 800–814. https://doi.org/10.1093/aje/kwg040.

[19] K. Wiklund, J. Dich, L.E. Holm, G. Eklund, Risk of cancer in pesticide applicators in Swedish agriculture., Occupational and Environmental Medicine. 46 (1989) 809–814. https://doi.org/10.1136/oem.46.11.809.

[20] International Agency for Research on Cancer (IARC), Some Organophosphate Insecticides and Herbicides, World Health Organization (WHO), Lyon, France, 2017. https://publications.iarc.fr/549 (accessed October 3, 2020).

[21] International Agency for Research on Cancer (IARC), Occupational Exposures in Insecticide Application, and Some Pesticides, World Health Organization (WHO), Lyon, France, 1991. https://publications.iarc.fr/71 (accessed October 3, 2020).

[22] International Agency for Research on Cancer (IARC), Some Organochlorine Pesticides, World Health Organization (WHO), Lyon, France, 1974. https://publications.iarc.fr/23 (accessed October 3, 2020).

[23] G.M. Williams, M. Aardema, J. Acquavella, S.C. Berry, D. Brusick, M.M. Burns, J.L.V. de Camargo, D. Garabrant, H.A. Greim, L.D. Kier, D.J. Kirkland, G. Marsh, K.R. Solomon, T. Sorahan, A. Roberts, D.L. Weed, A review of the carcinogenic potential of glyphosate by four independent expert panels and comparison to the IARC assessment, Critical Reviews in Toxicology. 46 (2016) 3–20. https://doi.org/10.1080/10408444.2016.1214677.

[24] G.M. Williams, C. Berry, M. Burns, J.L.V. de Camargo, H. Greim, Glyphosate rodent carcinogenicity bioassay expert panel review, Critical Reviews in Toxicology. 46 (2016) 44–55. https://doi.org/10.1080/10408444.2016.1214679.

[25] International Agency for Research on Cancer (IARC), IARC response to criticisms of the Monographs and the glyphosate evaluation, United Nations, Lyon, France, 2018. https://www.iarc.fr/wp-content/uploads/2018/07/IARC_response_to_criticisms_of_the_Monographs_and_the_glyphosate_evaluation.pdf (accessed October 4, 2020).

[26] C. Sandrock, L.G. Tanadini, J.S. Pettis, J.C. Biesmeijer, S.G. Potts, P. Neumann, Sublethal neonicotinoid insecticide exposure reduces solitary bee reproductive success: Loss of pollinator fitness, Agr Forest Entomol. 16 (2014) 119–128. https://doi.org/10.1111/afe.12041.

[27] J.S. Pettis, E.M. Lichtenberg, M. Andree, J. Stitzinger, R. Rose, D. vanEngelsdorp, Crop Pollination Exposes Honey Bees to Pesticides Which Alters Their Susceptibility to the Gut Pathogen Nosema ceranae, PLoS ONE. 8 (2013) e70182. https://doi.org/10.1371/journal.pone.0070182.

[28] M. Henry, M. Beguin, F. Requier, O. Rollin, J.-F. Odoux, P. Aupinel, J. Aptel, S. Tchamitchian, A. Decourtye, A Common Pesticide Decreases Foraging Success and Survival in Honey Bees, Science. 336 (2012) 348–350. https://doi.org/10.1126/science.1215039.

[29] P.R. Whitehorn, S. O'Connor, F.L. Wackers, D. Goulson, Neonicotinoid Pesticide Reduces Bumble Bee Colony Growth and Queen Production, Science. 336 (2012) 351–352. https://doi.org/10.1126/science.1215025.

[30] C. Brittain, S.G. Potts, The potential impacts of insecticides on the life-history traits of bees and the consequences for pollination, Basic and Applied Ecology. 12 (2011) 321–331. https://doi.org/10.1016/j.baae.2010.12.004.

[31] V. Mommaerts, S. Reynders, J. Boulet, L. Besard, G. Sterk, G. Smagghe, Risk assessment for side-effects of neonicotinoids against bumblebees with and without impairing foraging behavior, Ecotoxicology. 19 (2010) 207–215. https://doi.org/10.1007/s10646-009-0406-2.

[32] N. Desneux, A. Decourtye, J.-M. Delpuech, The Sublethal Effects of Pesticides on Beneficial Arthropods, Annu. Rev. Entomol. 52 (2007) 81–106. https://doi.org/10.1146/annurev.ento.52.110405.091440.

[33] D.G. Alston, V.J. Tepedino, B.A. Bradley, T.R. Toler, T.L. Griswold, S.M. Messinger, Effects of the Insecticide Phosmet on Solitary Bee Foraging and Nesting in Orchards of Capitol Reef National Park, Utah, Environmental Entomology. 36 (2007) 811–816. https://doi.org/10.1093/ee/36.4.811.

[34] E. Ladurner, J. Bosch, W.P. Kemp, S. Maini, Assessing delayed and acute toxicity of five formulated fungicides to *Osmia lignaria* Say and *Apis mellifera*, Apidologie. 36 (2005) 449–460. https://doi.org/10.1051/apido:2005032.

[35] P.G. Kevan, Forest application of the insecticide fenitrothion and its effect on wild bee pollinators (Hymenoptera: Apoidea) of lowbush blueberries (Vaccinium SPP.) in Southern New Brunswick, Canada, Biological Conservation. 7 (1975) 301–309. https://doi.org/10.1016/0006-3207(75)90045-2.

[36] H. Toumi, M. Boumaiza, M. Millet, C.M. Radetski, V. Felten, C. Fouque, J.F. Férard, Effects of deltamethrin (pyrethroid insecticide) on growth, reproduction, embryonic development and sex differentiation in two strains of Daphnia magna (Crustacea, Cladocera), Science of The Total Environment. 458–460 (2013) 47–53. https://doi.org/10.1016/j.scitotenv.2013.03.085.

[37] G. Domínguez-Cortinas, J.M. Saavedra, G.E. Santos-Medrano, R. Rico-Martínez, Analysis of the toxicity of glyphosate and Faena® using the freshwater invertebrates *Daphnia magna* and *Lecane quadridentata*, Toxicological & Environmental Chemistry. 90 (2008) 377–384. https://doi.org/10.1080/02772240701529038.

[38] T.H. Andersen, R. Tjørnhøj, L. Wollenberger, T. Slothuus, A. Baun, Acute and chronic effects of pulse exposure of Daphnia magna to dimethoate and pirimicarb, Environ Toxicol Chem. 25 (2006) 1187–1195. https://doi.org/10.1897/05-465R1.1.

[39] M. Besson, C. Gache, F. Bertucci, R.M. Brooker, N. Roux, H. Jacob, C. Berthe, V.A. Sovrano, D.L. Dixson, D. Lecchini, Exposure to agricultural pesticide impairs visual lateralization in a larval coral reef fish, Sci Rep. 7 (2017) 9165. https://doi.org/10.1038/s41598-017-09381-0.

[40] M.E. DeLorenzo, G.I. Scott, P.E. Ross, Toxicity of pesticides to aquatic microorganisms: A review, Environ Toxicol Chem. 20 (2001) 84–98. https://doi.org/10.1002/etc.5620200108.

[41] J.L. Pereira, S.C. Antunes, B.B. Castro, C.R. Marques, A.M.M. Gonçalves, F. Gonçalves, R. Pereira, Toxicity evaluation of three pesticides on non-target aquatic and soil organisms: commercial formulation versus active ingredient, Ecotoxicology. 18 (2009) 455–463. https://doi.org/10.1007/s10646-009-0300-y.

[42] C. Cox, M. Surgan, Unidentified Inert Ingredients in Pesticides: Implications for Human and Environmental Health, Environmental Health Perspectives. 114 (2006) 1803–1806. https://doi.org/10.1289/ehp.9374.

[43] K.A. Krogh, B. Halling-Sørensen, B.B. Mogensen, K.V. Vejrup, Environmental properties and effects of nonionic surfactant adjuvants in pesticides: a review, Chemosphere. 50 (2003) 871–901. https://doi.org/10.1016/S0045-6535(02)00648-3.

[44] T. Pereira, M.J. Cerejeira, J. Espírito-Santo, Use of microbiotests to compare the toxicity of water samples fortified with active ingredients and formulated pesticides, Environmental Toxicology. 15 (2000) 401–405. https://doi.org/10.1002/1522-7278(2000)15:5<401::AID-TOX7>3.0.CO;2-H.

[45] D.J. Oakes, J.K. Pollak, The in vitro evaluation of the toxicities of three related herbicide formulations containing ester derivatives of 2,4,5-T and 2,4-D using sub-mitochondrial particles, Toxicology. 151 (2000) 1–9. https://doi.org/10.1016/S0300-483X(00)00244-4.

[46] D. Virág, Z. Naár, A. Kiss, Microbial Toxicity of Pesticide Derivatives Produced with UV-photodegradation, Bull Environ Contam Toxicol. 79 (2007) 356–359. https://doi.org/10.1007/s00128-007-9230-7.

[47] P. Drzewicz, G. Nałęcz-Jawecki, M. Gryz, J. Sawicki, A. Bojanowska-Czajka, W. Głuszewski, K. Kulisa, S. Wołkowicz, M. Trojanowicz, Monitoring of toxicity during degradation of selected pesticides using ionizing radiation, Chemosphere. 57 (2004) 135–145. https://doi.org/10.1016/j.chemosphere.2004.05.019.

[48] C.J. Sinclair, A.B.A. Boxall, Assessing the Ecotoxicity of Pesticide Transformation Products, Environ. Sci. Technol. 37 (2003) 4617–4625. https://doi.org/10.1021/es030038m.

[49] O. Osano, W. Admiraal, H.J.C. Klamer, D. Pastor, E.A.J. Bleeker, Comparative toxic and genotoxic effects of chloroacetanilides, formamidines and their degradation products on Vibrio fischeri and Chironomus riparius, Environmental Pollution. 119 (2002) 195–202. https://doi.org/10.1016/S0269-7491(01)00334-7.

[50] O. Osano, W. Admiraal, D. Otieno, Developmental disorders in embryos of the frog *Xenopus laevis* induced by chloroacetanilide herbicides and their degradation products, Environ Toxicol Chem. 21 (2002) 375–379. https://doi.org/10.1002/etc.5620210221.

[51] G.W. Stratton, C.T. Corke, Toxicity of the insecticide permethrin and some degradation products towards algae and cyanobacteria, Environmental Pollution Series A, Ecological and Biological. 29 (1982) 71–80. https://doi.org/10.1016/0143-1471(82)90055-1.

[52] T. Ramesh, N.S. Bolan, M.B. Kirkham, H. Wijesekara, M. Kanchikerimath, C. Srinivasa Rao, S. Sandeep, J. Rinklebe, Y.S. Ok, B.U. Choudhury, H. Wang, C. Tang, X. Wang, Z. Song, O.W. Freeman II, Chapter One - Soil organic carbon dynamics: Impact of land use changes and management practices: A review, in: D.L. Sparks (Ed.), Advances in Agronomy, Academic Press, 2019: pp. 1–107. https://doi.org/10.1016/bs.agron.2019.02.001.

[53] S. Smith, D. Read, Mycorrhizal Symbiosis, 2nd ed., Academic Press, London, U.K., 1996. https://www.elsevier.com/books/mycorrhizal-symbiosis/smith/978-0-12-652840-4 (accessed October 5, 2020).

[54] S.C. Wu, Z.H. Cao, Z.G. Li, K.C. Cheung, M.H. Wong, Effects of biofertilizer containing N-fixer, P and K solubilizers and AM fungi on maize growth: a greenhouse trial, Geoderma. 125 (2005) 155–166. https://doi.org/10.1016/j.geoderma.2004.07.003.

[55] R.A. Sikora, K. Schäfer, A.A. Dababat, Modes of action associated with microbially induced ***in planta*** suppression of plant-parasitic nematodes, Austral. Plant Pathol. 36 (2007) 124–134. https://doi.org/10.1071/AP07008.

[56] U. Chakraborty, B. Chakraborty, M. Basnet, Plant growth promotion and induction of resistance inCamellia sinensis byBacillus megaterium, J. Basic Microbiol. 46 (2006) 186–195. https://doi.org/10.1002/jobm.200510050.

[57] J.W. Kloeppe, R. Rodríguez-Kábana, A.W. Zehnder, J.F. Murphy, E. Sikora, C. Fernández, Plant root-bacterial interactions in biological control of soilborne diseases and potential extension to systemic and foliar diseases, Australasian Plant Pathology. 28 (1999) 21–26. https://doi.org/10.1071/AP99003.

[58] L.C. van Loon, P.A.H.M. Bakker, C.M.J. Pieterse, Systemic Resistance Induced By Rhizosphere Bacteria, Annu. Rev. Phytopathol. 36 (1998) 453–483. https://doi.org/10.1146/annurev.phyto.36.1.453.

[59] G. Wei, J.W. Kloepper, S. Tuzun, Induced systemic resistance to cucumber diseases and increased plant growth by plant growth-promoting rhizobacteria under field conditions., Phytopathology. 86 (1996) 221–224. https://doi.org/10.1094/Phyto-86-221.

[60] C. Dimkpa, T. Weinand, F. Asch, Plant-rhizobacteria interactions alleviate abiotic stress conditions: Plant-rhizobacteria interactions, Plant, Cell & Environment. 32 (2009) 1682–1694. https://doi.org/10.1111/j.1365-3040.2009.02028.x.

[61] J. Yang, J.W. Kloepper, C.-M. Ryu, Rhizosphere bacteria help plants tolerate abiotic stress, Trends in Plant Science. 14 (2009) 1–4. https://doi.org/10.1016/j.tplants.2008.10.004.

[62] G. Berg, Plant–microbe interactions promoting plant growth and health: perspectives for controlled use of microorganisms in agriculture, Appl Microbiol Biotechnol. 84 (2009) 11–18. https://doi.org/10.1007/s00253-009-2092-7.

[63] T.N. Arkhipova, E. Prinsen, S.U. Veselov, E.V. Martinenko, A.I. Melentiev, G.R. Kudoyarova, Cytokinin producing bacteria enhance plant growth in drying soil, Plant Soil. 292 (2007) 305–315. https://doi.org/10.1007/s11104-007-9233-5.

[64] D. Perrig, M.L. Boiero, O.A. Masciarelli, C. Penna, O.A. Ruiz, F.D. Cassán, M.V. Luna, Plant-growth-promoting compounds produced by two agronomically important strains of Azospirillum brasilense, and implications for inoculant formulation, Appl Microbiol Biotechnol. 75 (2007) 1143–1150. https://doi.org/10.1007/s00253-007-0909-9.

[65] C.-M. Ryu, C.-H. Hu, R.D. Locy, J.W. Kloepper, Study of mechanisms for plant growth promotion elicited by rhizobacteria in Arabidopsis thaliana, Plant Soil. 268 (2005) 285–292. https://doi.org/10.1007/s11104-004-0301-9.

[66] E.P. de B. Ferreira, A.N. Dusi, J.R. Costa, G.R. Xavier, N.G. Rumjanek, Assessing insecticide and fungicide effects on the culturable soil bacterial community by analyses of variance of their DGGE fingerprinting data, European Journal of Soil Biology. 45 (2009) 466–472. https://doi.org/10.1016/j.ejsobi.2009.07.003.

[67] A. Tlili, A. Bérard, J.-L. Roulier, B. Volat, B. Montuelle, PO43− dependence of the tolerance of autotrophic and heterotrophic biofilm communities to copper and diuron, Aquatic Toxicology. 98 (2010) 165–177. https://doi.org/10.1016/j.aquatox.2010.02.008.

[68] Écophyto, Ministère de l'Agriculture et de l'Alimentation. (2015). https://agriculture.gouv.fr/ecophyto (accessed October 8, 2020).

[69] European Parliament, Council Of The European Union, Directive 2009/128/EC: establishing a framework for Community action to achieve the sustainable use of pesticides, 2009. https://eur-lex.europa.eu/eli/dir/2009/128/oj (accessed October 8, 2020).

[70] T. Glare, J. Caradus, W. Gelernter, T. Jackson, N. Keyhani, J. Köhl, P. Marrone, L. Morin, A. Stewart, Have biopesticides come of age?, Trends in Biotechnology. 30 (2012) 250–258. https://doi.org/10.1016/j.tibtech.2012.01.003.

[71] L.G. Copping, J.J. Menn, Biopesticides: a review of their action, applications and efficacy, Pest Management Science. 56 (2000) 651–676. https://doi.org/10.1002/1526-4998(200008)56:8<651::AID-PS201>3.0.CO;2-U.

[72] J.J. Villaverde, B. Sevilla-Morán, P. Sandín-España, C. López-Goti, J.L. Alonso-Prados, Biopesticides in the framework of the European Pesticide Regulation (EC) No. 1107/2009: Biopesticides within the European Pesticide Regulation (EC) No. 1107/2009, Pest. Manag. Sci. 70 (2014) 2–5. https://doi.org/10.1002/ps.3663.

[73] R. Pavela, G. Benelli, Essential Oils as Ecofriendly Biopesticides? Challenges and Constraints, Trends in Plant Science. 21 (2016) 1000–1007. https://doi.org/10.1016/j.tplants.2016.10.005.

[74] C. Bertrand, C. Prigent-Combaret, A. Gonzales-Coloma, Chemistry, activity, and impact of plant biocontrol products, Environ Sci Pollut Res. 25 (2018) 29773–29774. https://doi.org/10.1007/s11356-018-3209-2.

[75] D.C. Robin, P.A. Marchand, Evolution of the biocontrol active substances in the framework of the European Pesticide Regulation (EC) No. 1107/2009: Evolution of BCA within the framework of Regulation (EC) No. 1107/2009, Pest. Manag. Sci. 75 (2019) 950–958. https://doi.org/10.1002/ps.5199.

[76] P.G. Marrone, An effective biofungicide with novel modes of action, Pestic. Outlook. 13 (2002) 193–194. https://doi.org/10.1039/B209431M.

[77] Crop Science - Bayer, Key Pests Controlled by Serenade ASO, Crop Science. (n.d.). https://www.cropscience.bayer.us/products/fungicides/serenade-aso/pests (accessed October 15, 2020).

[78] P. Pitton, Low risk active substances in plant protection – State of play, (2018). https://ec.europa.eu/food/sites/food/files/safety/docs/adv-grp_plenary_20180427_pres_09a.pdf (accessed October 10, 2020).

[79] M. Amichot, P. Joly, F. Martin-Laurent, D. Siaussat, A.-V. Lavoir, Biocontrol, new questions for Ecotoxicology?, Environ Sci Pollut Res. 25 (2018) 33895–33900. https://doi.org/10.1007/s11356-018-3356-5.

[80] European Food Safety Authority (EFSA), M. Arena, D. Auteri, S. Barmaz, A. Brancato, D. Brocca, L. Bura, L. Carrasco Cabrera, A. Chiusolo, D. Court Marques, F. Crivellente, C. De Lentdecker, M. Egsmose, G. Fait, L. Ferreira, M. Goumenou, L. Greco, A. Ippolito, F. Istace, S. Jarrah, D. Kardassi, R. Leuschner, C. Lythgo, J.O. Magrans, P. Medina, I. Miron, T. Molnar, A. Nougadere, L. Padovani, J.M. Parra Morte, R. Pedersen, H. Reich, A. Sacchi, M. Santos, R. Serafimova, R. Sharp, A. Stanek, F. Streissl, J. Sturma, C. Szentes, J. Tarazona, A. Terron, A. Theobald, B. Vagenende, L. Villamar-Bouza, Peer review of the pesticide risk assessment of the active substance spinosad, EFSA Journal. 16 (2018) e05252. https://doi.org/10.2903/j.efsa.2018.5252.

[81] C.A. Brühl, L. Després, O. Frör, C.D. Patil, B. Poulin, G. Tetreau, S. Allgeier, Environmental and socioeconomic effects of mosquito control in Europe using the biocide Bacillus thuringiensis subsp. israelensis (Bti), Science of The Total Environment. 724 (2020) 137800. https://doi.org/10.1016/j.scitotenv.2020.137800.

[82] C. Mallet, S. Romdhane, C. Loiseau, J. Béguet, F. Martin-Laurent, C. Calvayrac, L. Barthelmebs, Impact of Leptospermone, a Natural β-Triketone Herbicide, on the Fungal Composition and Diversity of Two Arable Soils, Front. Microbiol. 10 (2019) 1024. https://doi.org/10.3389/fmicb.2019.01024.

[83] S. Romdhane, M. Devers-Lamrani, J. Beguet, C. Bertrand, C. Calvayrac, M.-V. Salvia, A.B. Jrad, F.E. Dayan, A. Spor, L. Barthelmebs, F. Martin-Laurent, Assessment of the ecotoxicological impact of natural and synthetic β-triketone herbicides on the diversity and activity of the soil bacterial community using omic approaches, Science of The Total Environment. 651 (2019) 241–249. https://doi.org/10.1016/j.scitotenv.2018.09.159.

[84] European Parliament, Council Of The European Union, Regulation (EC) No 1107/2009: concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC, 2009. https://eur-lex.europa.eu/eli/reg/2009/1107/oj (accessed October 8, 2020).

[85] I. Sundh, M.S. Goettel, Regulating biocontrol agents: a historical perspective and a critical examination comparing microbial and macrobial agents, BioControl. 58 (2013) 575–593. https://doi.org/10.1007/s10526-012-9498-3.

[86] European Commission – Health & Consumer Protection Directorate-General, Guidance Document on the assessment of new isolates of Baculovirus species already included in Annex I of Council Directive 91/414/EEC (SANCO/0253/2008 rev. 2), European Commission, Brussels, Belgium, 2008. https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_aas_guidance_baculovirus.pdf (accessed October 8, 2020).

[87] European Commission – Health & Consumer Protection Directorate-General, Guidance Document on botanical active substances used in plant protection products (SANCO/11470/2012– rev. 8), European Commission, Brussels, Belgium, 2014. https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_doss_botanicals-rev-8.pdf (accessed October 8, 2020).

[88] European Commission – Health & Consumer Protection Directorate-General, Guidance Document on semiochemical active substances and plant protection products (SANTE/12815/2014 rev. 5.2), European Commission, Brussels, Belgium, 2016. https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_doss_semiochemicals-201605.pdf (accessed October 8, 2020).

[89] European Commission – Health & Consumer Protection Directorate-General, Guidance Document for applicants on preparing dossiers for the approval or renewal of approval of a micro-organisms including viruses according to Regulation (EU) No 283/2013 and Regulation (EU) No 284/2013 (SANCO/12545/2014– rev. 2), European Commission, Brussels, Belgium, 2016. https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_applicants-microbial_en.pdf (accessed October 8, 2020).

[90] Guidelines on Active Substances and Plant Protection Products, European Commission. (n.d.). https://ec.europa.eu/food/plant/pesticides/approval_active_substances/guidance_documents_en (accessed October 11, 2020).

[91] H. Fontier, Procedure for the approval of an a.s. under Regulation (EC) No 1107/2009, (2011). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_efsa-proc.pdf (accessed October 12, 2020).

[92] European Commission – Health & Consumer Protection Directorate-General, Guidance for generating and reporting methods of analysis in support of pre-registration data requirements for Annex II (part A, Section 4) and Annex III (part A, Section 5) of Directive 91/414. (SANCO/3029/99 rev.4), (2000). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_res_pre-reg-cont-monitor.pdf (accessed October 12, 2020).

[93] European Commission – Directorate General for Agriculture, Guidance Document on Persistence in Soil (9188/VI/97 rev. 8), (2000). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_fate_soil-persistance.pdf (accessed October 12, 2020).

[94] European Commission – Health & Consumer Protection Directorate-General, Guidance Document for evaluating laboratory and field dissipation studies to obtain DegT50 values of active substances of plant protection products and transformation products of these active substances in soil (SANCO/12117/2014 – final), (2014). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_fate_efsa_degt50.pdf (accessed October 12, 2020).

[95] European Food Safety Authority, EFSA Guidance Document for evaluating laboratory and field dissipation studies to obtain DegT50 values of active substances of plant protection products and transformation products of these active substances in soil, EFSA Journal. 12 (2014) 3662. https://doi.org/10.2903/j.efsa.2014.3662.

[96] OECD, Guidance Document for Conducting Pesticide Terrestrial Field Dissipation Studies (ENV/JM/MONO(2016)6), OECD, Paris, France, 2016. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO%282016%296&doclanguage=en (accessed October 15, 2020).

[97] European Commission – Health & Consumer Protection Directorate-General, Guidance document on pesticide residue analytical methods (SANCO/825/00 rev. 8.1), (2010). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_res_post-reg-cont-monitor.pdf (accessed October 12, 2020).

[98] European Commission – Health & Consumer Protection Directorate-General, Guidance Document on Terrestrial Ecotoxicology Under Council Directive 91/414/EEC (SANCO/10329/2002 rev 2 final), (2002). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_ecotox_terrestrial.pdf (accessed October 12, 2020).

[99] J. Popovici, C. Bertrand, G. Comte, Use of a Myrica gale plant for producing a herbicide agent, US008734858B2, 2014. https://patents.google.com/patent/US8734858B2/en.

[100] K.P. Svoboda, A. Inglis, J. Hampson, B. Galambosi, Y. Asakawa, Biomass production, essential oil yield and composition of Myrica gale L. harvested from wild populations in Scotland and Finland, Flavour and Fragrance Journal. 13 (1998) 6. https://doi.org/10.1002/(SICI)1099-1026(199811/12)13:6<367::AID-FFJ724>3.0.CO;2-M.

[101] R.R. Carlton, P.G. Waterman, A.I. Gray, Variation of leaf gland volatile oil within a population of sweet gale (Myrica gale) (Myricaceae), Chemoecology. 3 (1992) 45–54. https://doi.org/10.1007/BF01261456.

[102] K. Oracz, A. Voegele, D. Tarkowská, D. Jacquemoud, V. Turečková, T. Urbanová, M. Strnad, E. Sliwinska, G. Leubner-Metzger, Myrigalone A Inhibits Lepidium sativum Seed Germination by Interference with Gibberellin Metabolism and Apoplastic Superoxide Production Required for Embryo Extension Growth and Endosperm Rupture, Plant and Cell Physiology. 53 (2012) 81–95. https://doi.org/10.1093/pcp/pcr124.

[103] A. Khaled, M. Sleiman, E. Darras, A. Trivella, C. Bertrand, N. Inguimbert, P. Goupil, C. Richard, Photodegradation of Myrigalone A, an Allelochemical from *Myrica gale*: Photoproducts and Effect of Terpenes, J. Agric. Food Chem. 67 (2019) 7258–7265. https://doi.org/10.1021/acs.jafc.9b01722.

[104] OECD, Report of the 9th Biopesticides Expert Group Seminar on Test Methods for Microorganisms (ENV/JM/MONO(2019)8), 2019. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2019)8&doclanguage=en (accessed October 10, 2020).

[105] Search for the term "Metabolomics," PubMed. (2020). https://pubmed.ncbi.nlm.nih.gov/?term=(metabolomics) (accessed October 11, 2020).

[106] H. Tweeddale, L. Notley-McRobb, T. Ferenci, Effect of Slow Growth on Metabolism of Escherichia coli, as Revealed by Global Metabolite Pool ("Metabolome") Analysis, J. Bacteriol. 180 (1998) 5109–5116. https://doi.org/10.1128/JB.180.19.5109-5116.1998.

[107] C.Y. Lin, M.R. Viant, R.S. Tjeerdema, Metabolomics: Methodologies and applications in the environmental sciences, J. Pestic. Sci. 31 (2006) 245–251. https://doi.org/10.1584/jpestics.31.245.

[108] J.G. Bundy, M.P. Davey, M.R. Viant, Environmental metabolomics: a critical review and future perspectives, Metabolomics. 5 (2009) 3–21. https://doi.org/10.1007/s11306-008-0152-0.

[109] M.R. Viant, U. Sommer, Mass spectrometry based environmental metabolomics: a primer and review, Metabolomics. 9 (2013) 144–158. https://doi.org/10.1007/s11306-012-0412-x.

[110] B.P. Lankadurai, E.G. Nagato, M.J. Simpson, Environmental metabolomics: an emerging approach to study organism responses to environmental stressors, Environ. Rev. 21 (2013) 180–205. https://doi.org/10.1139/er-2013-0011.

[111] K. Longnecker, J. Futrelle, E. Coburn, M.C. Kido Soule, E.B. Kujawinski, Environmental metabolomics: Databases and tools for data analysis, Marine Chemistry. 177 (2015) 366–373. https://doi.org/10.1016/j.marchem.2015.06.012.

[112] C. Bedia, P. Cardoso, N. Dalmau, E. Garreta-Lara, C. Gómez-Canela, E. Gorrochategui, M. Navarro-Reig, E. Ortiz-Villanueva, F. Puig-Castellví, R. Tauler, Chapter Nineteen - Applications of Metabolomics Analysis in Environmental Research, in: J. Jaumot, C. Bedia, R. Tauler (Eds.), Comprehensive Analytical Chemistry, Elsevier, 2018: pp. 533–582. https://doi.org/10.1016/bs.coac.2018.07.006.

[113] Search for the terms "Metabolomics" and "Environmental," PubMed. (2020). https://pubmed.ncbi.nlm.nih.gov/?term=(metabolomics)%20AND%20(environmental) (accessed October 11, 2020).

[114] K.A. Aliferis, M. Chrysayi-Tokousbalides, Metabolomics in pesticide research and development: review and future perspectives, Metabolomics. 7 (2011) 35–53. https://doi.org/10.1007/s11306-010-0231-x.

[115] C. Patil, C. Calvayrac, Y. Zhou, S. Romdhane, M.-V. Salvia, J.-F. Cooper, F.E. Dayan, C. Bertrand, Environmental Metabolic Footprinting: A novel application to study the impact of a natural and a synthetic β-triketone herbicide in soil, Science of The Total Environment. 566–567 (2016) 552–558. https://doi.org/10.1016/j.scitotenv.2016.05.071.

[116] M.-V. Salvia, A. Ben Jrad, D. Raviglione, Y. Zhou, C. Bertrand, Environmental Metabolic Footprinting (EMF) vs. half-life: a new and integrative proxy for the discrimination between control and pesticides exposed sediments in order to further characterise pesticides' environmental impact, Environ Sci Pollut Res. 25 (2018) 29841–29847. https://doi.org/10.1007/s11356-017-9600-6.

[117] M.A. O'Malley, Metametabolomics, in: W. Dubitzky, O. Wolkenhauer, H. Yokota, K.-H. Cho (Eds.), Encyclopedia of Systems Biology, 1st ed., Springer-Verlag New York, New York, NY, 2013: pp. 1296–1297. https://doi.org/10.1007/978-1-4419-9863-7_903.

[118] W. Dubitzky, O. Wolkenhauer, H. Yokota, K.-H. Cho, eds., Encyclopedia of Systems Biology, 1st ed., Springer-Verlag New York, New York, NY, 2013. https://www.springer.com/gp/book/9781441998620 (accessed August 15, 2020).

[119] European Regional Development Fund, Protection Alternative Des Productions Végétales Interrégionale Pyrénéenne (Interreg POCTEFA PALVIP project), POCTEFA. (2018). https://www.poctefa.eu/fr/listes-de-projets/detail-du-projet/?IdProyecto=63932bfe-f1de-461b-adc5-664211b79add (accessed February 16, 2021).

[120] Marie Skłodowska-Curie Actions, Study of the environmental impact of insecticides by metabolomic foot-printing approach (EnvFate project), CORDIS. (2017). https://cordis.europa.eu/project/id/746656 (accessed February 16, 2021).

[121] Agence Nationale de la Recherche, Environmental impact of synthetic and natural β-triketone herbicides: detection, microbial adaptation, biodegradation and toxicity (TRICETOX project), Agence Nationale de La Recherche. (2013). https://anr.fr/Project-ANR-13-CESA-0002 (accessed February 16, 2021).

Materials and Methods

"Fundamental Aspects"

Preamble

In order to achieve the defined objectives of the present thesis, diverse Analytical Chemistry approaches, techniques and instrumentations will be coupled with Chemometrics and Statistics. They will be exploited in a metabolomics-based framework. The present section aims to introduce the used analytical and statistical tools with a brief historical background, and to describe their fundamental aspects. Moreover, several examples and diagrams will be given in order to explain the main principles of the highlighted tools, and will introduce their proper theoretical terms and formulas.

1. Extraction

The extraction is a physical action that aims to transfer molecules from one material to another, *e.g.*, from the solid matrix of soil to a liquid solvent. In analytical chemistry and metabolomics, it is used in order to transfer the molecules of interest (analytes) from the sample to the analytical system. The extraction constitutes one of the most critical steps in metabolomics workflow. It could influence the detection, the quantification, and the reproducibility of a given experiment. Thus, it should be rigorously optimized and in-depth examined in order to assure high-quality results and conclusions.

The main rule determining the mechanisms and the performance of the extraction is the Thermodynamics **law of partition**, *i.e.*, the exchange equilibrium of a given molecule between the different materials involved in the extraction process. This equilibrium can be influenced by different factors, mainly the **non-covalent bonding interactions** that can take place between the molecule and the involved materials (*e.g.* electrostatic interactions, Hydrogen bond, van der Waals force, or dipole-dipole interactions). Those interactions are determined by the polarity of both the molecule and the involved materials. In fact, molecules containing partially charged heteroatoms (O, N, S, P, Halogens, etc.) or π-bonds in certain positions and within a specific stereochemistry have an affinity with molecules presenting the same properties. Otherwise, molecules with long aliphatic carbon chains or cyclic structures tend to assemble with the similar molecules. **The potential of Hydrogen (pH), the temperature, the gas pressure**[4]**, the liquid viscosity and the solid particle size** can also influence the extraction mechanisms by displacing the equilibrium of partition of the analyte between the two materials, following the principle of Le Chatelier [2]. Thus, the physical-chemical nature, the atomic composition and the stereochemistry, as well as the state of both the analyte and the involved materials should be taken into consideration for extractions development.

In the present thesis, two types of extractions will be developed and applied: the Soil-Liquid Extraction, and the Headspace-Solid Phase Microextraction.

[4] *Cf.* Henry's law [1].

1.1. Solid-Liquid Extraction

The Solid-Liquid Extraction[5] is a common technique that is widely used in analytical chemistry and metabolomics. It is based on extracting analytes (solutes) from a solid matrix by dissolving them in a liquid solvent. Therefore, the partition of the analyte depends on its affinity to the solid matrix, and its solubility in the solvent. Several factors can influence this type of extractions, mainly, **the polarity of both the analyte and the solvent**, as well as the **physical-chemical properties of the solid**. In fact, as described previously, if the analyte and the solvent are similar in term of polarity, the dissolution should be favored. If the analyte and the solid present similar or complementary natures, a high affinity can be expected, *e.g.*, if the solid contains metal elements and the molecule is polar or contain donor atoms (as N, O), the molecule presents a high affinity to the solid due to the electrostatic attractions, or the coordination complexation, respectively.

In addition, **the pH** is another factor that can influence the equilibrium and the extraction efficiency, particularly, when the analytes present acid or basic properties. In fact, if an organic molecule is acid, the increase of pH leads to its deprotonation. Thus, it will be transformed to a negative-charged anion (the conjugate base) and could be highly dissolved in polar solvents as H_2O. For base organic compounds, the decrease of pH leads to their protonation, so they will be transformed to positive-charged cations (acid conjugates) and thus, they will be more soluble in polar solvents. On another hand, the cationic exchange is a well-known adsorption process that can occur in matrices like soil. Cationic compounds can be highly adsorbed on such matrices, which renders their extraction less efficient [3]. Aggressive pH values should however not be applied in order to avoid potential analyte degradation by hydrolysis, dehydration or reactions with solvents such as Methanol that can attack and protect Carbonyl group in basic conditions.

On the other hand, the influence of **the temperature** on the equilibrium depends on the nature of both the analyte and the solvent. Theoretically, if the desolvation is exothermic, a low-temperature extraction can favor the extraction efficiency. Otherwise, a high-temperature extraction favors the endothermic desolvation. However, practically, the quantity of the solvent is mostly applied with a relatively large excess. Thus, different mechanisms will dominate and in this case, increasing the temperature can increase the solubility of the analytes (solutes). For

[5] The “SLE” abbreviation for Solid-Liquid Extraction will not be adopted in order to avoid the confusion with the “Supported Liquid Extraction”.

instance, increasing temperature leads to the increase of the kinetic energies of both the solvent molecules and the solute molecules, which renders solvent molecules more effective to break the solute intermolecular attractions, and increases the vibration of solute molecules. This increase of vibration favors their desolvation due to the decrease of their ability to hold together. Otherwise, the decrease in solvent viscosity and the melting of some solid-unsolvable analytes by temperature increase are other possible mechanisms that can favor the extraction efficiency. Nonetheless, high-temperature can negatively affect thermolabile analytes that might be degraded in such conditions. It might also engender a loss of volatile metabolites during the extraction.

Other physical and mechanical factors can influence the Solid-Liquid Extraction. For instance, small/fine solid particle size favors the extraction efficiency, as the solid-liquid contact interface will be increased. The increase of solid particles vibration also enhances the extraction efficiency, but can also engender a potential degradation of analytes due to the potential increase of their internal energies and temperatures. The double-extraction is also a suitable strategy that can be applied in order to increase the extraction yield. However, this strategy can increase the variability and thus risks deteriorating extraction reproducibility.

For untargeted metabolomics analyses, and particularly for the EMF, the main challenges in Solid-Liquid Extraction development is to broaden the band of polarity in order to collect wider metabolic information. This task is challenging, as the extraction of both polar and non-polar molecules in a single extraction is contradictory from a theoretical point of view. Practical and technical solutions should be thus engineered in order to overcome this problematic. In addition, a high extraction reproducibility should be assured, as the untargeted metabolomics is based on comparative analyses of relative intensities. On the other hand, the purification is not reasonable in untargeted metabolomics as it can engender a loss of information, which can cause some problematic issues related to the unavoidable influence of complex biological matrix effects on the analytical response of analytes. Moreover, the extraction development and optimization should be done using compatible reagents for LC-ESI-MS (*i.e.* the adopted analytical system in this work). Thus, surfactants, Acetone[6], hydrocarbon non-polar solvents[7], some inorganic salts and other reagents and solvents cannot be used.

[6] Acetone use in positive Electrospray Ionization is problematic, as it provokes an elevated noise due to the formation of Acetone-condensed products [4–6].
[7] They cannot be used as final solvents of the mixture that will be injected. They can be used for the earlier stages of the extraction however, and then evaporated and replaced by other compatible solvents.

1.2. Headspace-Solid Phase Microextraction

Headspace-Solid Phase Microextraction (HS-SPME) is an extraction technique that targets the volatile compounds in the gas phase. As its name indicates, it consists of two components: the HS and the SPME. The HS means the gas phase above a given sample introduced in an analysis tube. The SPME presented in Figure M. 1 is an advanced extraction technique that was introduced in 1989 by the Polish chemist Janusz Pawliszyn [7]. Its principal is based on extracting and isolating the analytes from the sample by adsorbing and concentrating them on the layer of a coated fiber. Thus, they can be eventually desorbed and introduced in the analytical instrument [8,9].

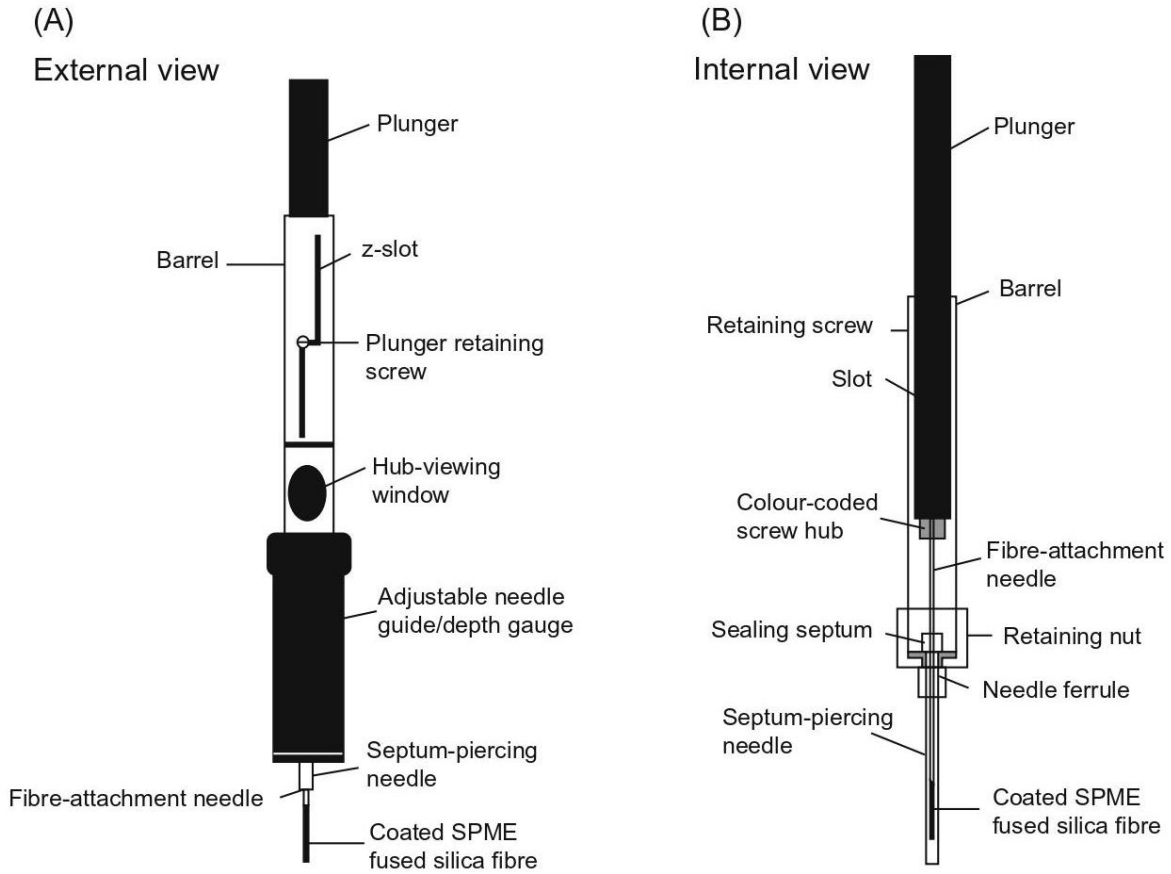


Figure M. 1: A schematic presentation of the SPME. Source: [8,10].

The fiber coating can influence the extraction following two factors: 1) its surface, and 2) its chemical composition. The coating surface is directly related to the available sites for compounds adsorption. If the surface dimensions increase, the capacity of adsorption increases, and thus the extraction yield is enhanced. The coating surface can be increased either by enlarging the geometric dimensions as the diameter or the exposable length of the fiber, or by increasing the porosity or the roughness of the surface. On the other hand, the chemical

composition of the coating influences the profile of the extracted compounds. The influence follows the similarity and affinity principles that can be determined following the non-covalent interactions. For instance, the Polydimethylsiloxane (PDMS) and the Polyethylene glycol (PEG) coatings are suitable to adsorb polar and semi-polar compounds, as they contain Oxygen atoms and small Carbon chains. Carboxen (CAR) is suitable to adsorb non-polar compounds as it consists of chains of cyclic carbons. The Divinylbenzene (DVB) is adapted to adsorb aromatic compounds and molecules containing unsaturated Carbons, as it contains Benzene and Divinyl π-bonds.

The HS-SPME is also influenced by other three main factors: **the sample incubation time, the fiber exposure time (extraction time), and the extraction temperature**.

The incubation time plays an essential role in determining the extracted molecular profile. In fact, if the incubation time is relatively short, the major components occupying the Headspace will be the most volatile compounds. However, if the incubation time is extended, the relatively less volatile compounds will be more accumulated in the Headspace and thus the possibility of their adsorption by the SPME fiber will be enhanced. In fact, they will enter in a competition with the other volatile compounds in order to access the fiber. Therefore, the incubation time should be optimized **following the targeted analytes and their volatility**.

On the other hand, **the extraction time** generally influences the efficiency of the extraction. For a given molecule, a proper exposure time should be considered in order to allow for the surpass of the "Kinetic Regime" and the reaching of the "Equilibrium Regime", as explained by Souza-Silva *et al.* [11]. Once the equilibrium is established, the adsorbed amount of the molecule can no more evolve. The SPME fiber can be considered as "saturated" at this stage.

Concerning **the extraction temperature**, the increasing of this factor generally increases the evaporation of the different compounds, which enhances their accumulation in the Headspace. However, high temperatures can negatively affect molecules adsorption on the fiber. In fact, the adsorption of most of molecules is an exothermic phenomenon. Thus, theoretically, when the temperature is increased, the equilibrium is displaced in the reverse sense, which favors desorption and leads to a decrease in extraction yield. Thus, the optimal extraction temperature is reached when the saturation of the Headspace with the evaporated molecules is dominating the equilibrium and taking over desorption.

The HS-SPME advantages for **metabolomics** are its capability for non-destructive extractions, its ability to analyze the volatile compounds in the gas phase, and its possible automation that allows for high throughput and highly reproducible extractions.

2. Analytical Technologies and Instrumentations

2.1. Chromatography

The chromatography is a chemical technique that aims to separate different molecules (analytes) present in a chemical mixture. The separation is defined by the Retention Time (RT), *i.e.* the duration taken by a molecule to pass through the chromatographic column. The name of Chromatography is composed by two Greek words: "Khroma" that means "Colors", and "Graphein" that means "Writing"[8]. The word "color" was introduced because the Chromatography was first invented in order to separate plant pigments. The inventor was the Russian-Italian botanist Mikhail Semyonovich Tsvet[9] who introduced this technique in 1903 [12,13].

Since that time, and as this invention revolutionized Analytical Chemistry (as well as Organic Chemistry and Industrial Chemistry), the Chromatography has been widely considered for developments and improvements. Its fundamental aspects and laws were in-depth studied. One of the most significant works in this frame was the explanation and the modeling of the elution, the separation and the chromatographic resolving power rules that were carried out by Jan Jozef van Deemter and his colleagues [14]. This work led to the famous "van Deemter equation" that modeled the different types of diffusions of a molecule inside the chromatographic column, and thus explained the broadening of the chromatographic peak that takes a Gaussian shape.

At the present, the Chromatography constitutes a common technique that is widely used in almost all the fields that studies organic molecules. Several types of Chromatography are currently applied in metabolomics, mainly the Liquid Chromatography (LC) and the Gas Chromatography (GC) that will be exploited in the current thesis.

2.1.1. Liquid Chromatography

As its name indicates, the Liquid Chromatography is dedicated to analyze solutions and molecules in the liquid state. It is suitable to analyze polar, nonvolatile and thermolabile molecules.

The LC handles two main components: the **stationary phase** and the **mobile phase**. The separation occurs following the different physical-chemical properties of the analyzed

[8] Khroma: "Χρῶμα", and Graphien: "Γράφειν".
[9] As by coincidence (or not), his surname "Tsvet" ("Цвет") means "Color" in Russian.

molecules. In fact, an analyzed molecule undergoes a dynamic equilibrium between the stationary phase and the mobile phase. This equilibrium is ruled by several physical-chemical interactions and factors, mainly by the polarities of the two phases and the polarity of the molecule, as well as other influencing factors such as the pH and the temperature. Therefore, for each type of molecules of interest, defined types of stationary phase and mobile phase should be selected. In **Mass Spectrometry-based metabolomics**, the most commonly used types are **the Reverse Phase Liquid Chromatography (RPLC)** and the **Hydrophilic Interaction Chromatography (HILIC)** [15]**.** Those two types and their applicable mobile phases are compatible with MS analyses and particularly with the ESI source.

The RPLC is mainly dedicated to analyze semi-polar and non-polar molecules. Its stationary phase is a non-polar phase that assures a strong retention of the similar non-polar compounds. C18 and C8 chains implanted on a base of Silica are the most used phases. The mobile phase varies between the polar H_2O that has a week elution efficiency in RPLC, and the other organic solvents as Methanol and Acetonitrile that assures a strong elution of molecules in such LC phases. Additives as Formic Acid (FA) or Ammonium Formate can be added to the elution solvents in order to adjust pH[10] and enhance the ionization of the analyzed molecules when the LC is coupled to ESI-MS. The RPLC is known for its high efficiency and resolving power for semi-polar and non-polar molecules. However, its inability to retain polar compounds present one of its major drawbacks.

On the other hand, **the HILIC is mainly dedicated for analyzing polar compounds**. The HILIC consists of a particular and sophisticated type of stationary phase. In fact, a fixed stationary phase consisting of a polar motif (*e.g.* amino-silica, amide-silica, cyano-silica or others) holds a "semi-stagnant" stationary phase based on H_2O. This latter is the phase that retains the polar compounds. The mobile phase varies between organic solvents with low elution power, and H_2O that performs the elution of polar metabolites. The HILIC application for primary and polar metabolome analyses is increasing in MS-based metabolomics [15–17]. Its use for lipids analyses is also considered over the recent past years [17], as the RPLC shows difficulties when analyzing lipids due to their high affinity to RPLC stationary phases (their elution in RPLC columns is extremely difficult). Nonetheless, the HILIC still present several drawbacks, mainly, the need for long-time conditioning and the reproducibility and robustness issues [18].

[10] Adjusting pH in RPLC is important in order to avoid the protonation or deprotonation of certain analytes. In fact, when analytes are charged, their retention in RPLC is significantly less efficient due to their high polarity.

Recently, **novel types of RPLC columns based on modified C18 stationary phases has been introduced**. Those hybrid columns are based on particular stationary phases that consist of C18 or C8 chains combined with polar or charged groups. **The objective of integrating such polar groups is to enhance the retention of polar metabolites**, without altering the analytical efficiency for semi-polar and non-polar compounds. The efficiency of this type of columns for analyzing polar molecules has been reported acceptable in the literature [18,19]. Thus, they can present a suitable choice to broaden the band of the analyzed molecules, particularly, when a wide variety of molecules and molecular families is targeted, which is the case for the meta-metabolome analysis for the EMF.

It is worth mentioning that the separation efficiency in LC can also be affected by other factors mainly related to fluid mechanics, such as the mobile phase flow, the LC column dimensions and the stationary phase particle size. The increase of mobile phase flow increases the resolution of the chromatographic peaks (sharpen the peaks) [14]. For the column dimensions, the separation is improved when the column is longer. However, the increase of column length renders the elution time longer (increases analytes RTs) and thus extends the LC run time. On the other hand, smaller column internal diameters and particle sizes (particle diameters) improve the efficiency and the resolution of the LC by limiting the longitudinal diffusion of analytes particles [14], and also shorten the elution time by increasing the column backpressure. Following those last two column properties, the LC can be defined as High Performance Liquid Chromatography (HPLC) or Ultra High Performance Liquid Chromatography (UHPLC). In fact, when the column internal diameter is 2.1 mm or less, and the particle diameter is 2 µm or less, the column is considered as UHPLC column.

It should be mentioned that the temperature could also influence LC efficiency from fluid mechanics point of view. In fact, increasing temperature decreases the viscosity of the mobile phase, and thus decreases the column backpressure. Increasing temperature enhances the resolution and decreases analytes RTs (shorten the elution time). However, from Thermodynamics point of view, the temperature also affects analyte's equilibrium of partition between the mobile phase and the stationary phase. The equilibrium is displaced according to the nature of the analyte and the thermochemistry of its partition. Thus, as the partition is affected differently for each analyte, the increase of temperature can engender nonsystematic and nonlinear shifts in elution order, and as analytes RTs are decreasing, such nonsystematic/nonlinear shifts can deteriorate the separation by leading the analytes to co-elute.

In addition, high temperature can alter analytes stability and degrade the stationary phase, which deteriorates the robustness of the column and the repeatability of the analysis.

2.1.2. Gas Chromatography

Gas Chromatography is dedicated to separate molecules in gas phase. It is a suitable technique to analyze volatile, semi-volatile and thermostable compounds. The separation in GC is mainly performed following the different boiling points of the analytes. When a temperature-ramp gradient is applied in the oven, the given analyte will be evaporated and carried through the column once its boiling point is reached. In addition, other factors can intervene in analytes separation and elution, as the polarity-based interactions between the analyte and the column, or the affinity of the analyte to the column. The GC consists of three main components: the mobile phase, the stationary phase, and the inlet.

The mobile phase consists of a carrier gas (or a vector gas) as Nitrogen, Helium, Argon, Dihydrogen, and others. Its role is to carry and push the analytes through the GC column. The stationary phase consists of a chemical group implanted on the internal layer of the GC column. Various types of stationary phases are common for GC, *e.g.*, Polydimethylsiloxane, Polydimethyldiphenylsiloxane, Polyethylene glycol, Polydicyanopropylsiloxane, and others. **Each type has specific affinity and polarity properties**. The inlet is the component where the investigated solution is injected and where the analytes are evaporated before their introduction into the column. Two modes of inlet are common for GC: the Split and the Splitless. The first implies the application of an inlet gas flow that flushes the injector in order to eliminate the solvent of the analyzed mix. The aim of this flush is to avoid column overload and the suppression of the compounds that elute at the beginning of the run by the massive peak belonging to the solvent. This injection type enhances the efficiency of the column, but can engender a loss in sensitivity for certain molecules (mainly the highly volatile molecules and the compounds with minor abundances). The Splitless mode does not imply any flush. All the injected solution (including the solvent) is injected in the column. This mode does not imply a loss of sensitivity. It is mostly used for trace analyses as well for analyzing different types of samples that are not in solution form (as the HS and the SPME).

The GC is commonly used in metabolomics for several reasons. Besides its capability to analyze volatile and semi-volatile compounds, it is known for its high performance, high efficiency, and high chromatographic resolving power [15]. These advantages provide sharp chromatographic peaks and thus assure high sensitivity and high selectivity. In addition, the GC

is a robust and reproducible separation technique [15]. The retention times are highly repeatable and can be modeled following the Kováts relation that converts RTs to "Kováts Retention Indices" (RI) [20–22]. These RIs can be calculated after injecting a mix of alkanes [23]. They can be generalized for a wide range of column types and for different GC methods. Thus, RIs can be used for compounds identification. They can be found in metabolome databases as characterizing indicators.

2.2. Mass Spectrometry

Mass Spectrometry (MS) is one of the advanced technologies that are commonly used in Analytical Chemistry. It is also considered as a scientific field *per se*, pertaining to the Physical-Chemistry research. This technology is widely used for chemical analyses, mainly for molecular characterizing, compounds detection, and quantitative analyses. It is known for its high sensitivity and advanced selectivity.

The primary role of Mass Spectrometry is to determine the monoisotopic mass of a given ion. This determination is achieved by measuring the "mass-to-charge" (*m/z*) ratio. The detection of the ion can also be used to **quantify its amount**. In fact, the MS acquires a mass spectrum consisting of two dimensions (Figure M. 2). The abscissa is the *m/z* range, and the ordinate is the intensity of the measured *m/z*. This intensity is directly related to the amount/quantity of the detected ion.

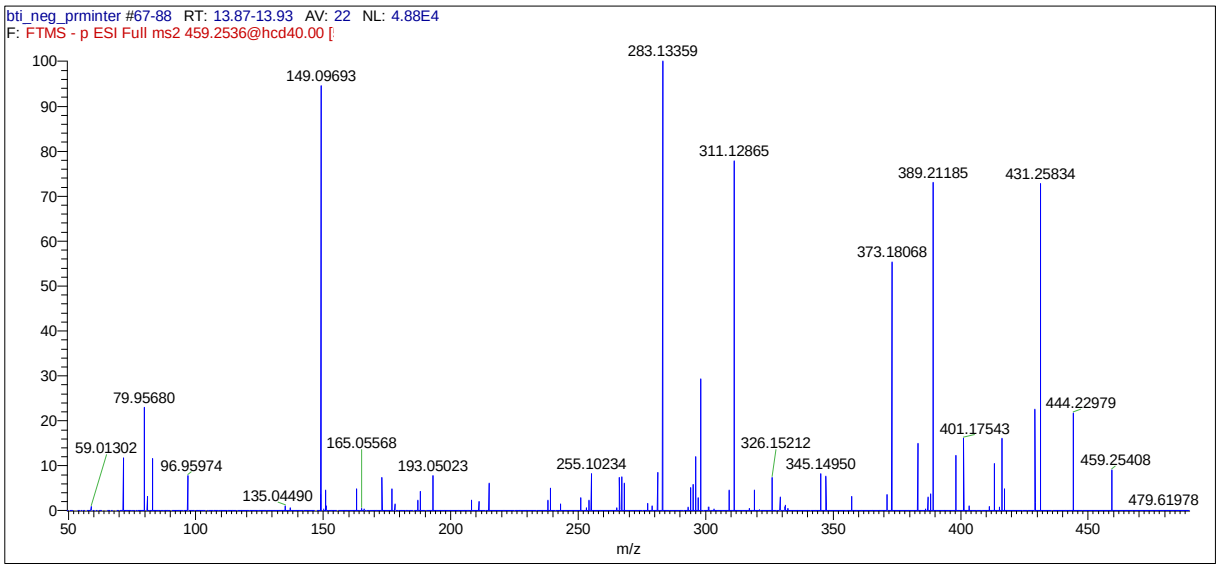


Figure M. 2: A mass spectrum (for an unknown compound with a *m/z* of 459.2536 ($[M-H]^-$)).

The MS is capable to identify a given molecule. This identification can be done by different approaches and techniques that are related to the type of the mass spectrometer. In fact, mass

spectrometers are various and different. Each type of MS is dedicated for one or several tasks. However, **the main three component of MS are mostly the same for all spectrometers: the ionization source, the mass analyzer, and the sensor/detector**. The ionization source is the component that ionize the introduced molecules (analytes) in order to charge them. Then, the mass analyzer is the component where *m/z* measures are achieved. The sensor/detector then receive these ions in order to acquire the measured *m/z* and to determine their intensities by counting the detected ions. This is done by an electronic process that detects ion signals and then transform them to a digitalized information.

2.2.1. Ionization sources

Various ionization sources are available for MS. They can be classified following their modes and the applied ionization energies. For instance, ionizations as the **Electron Impact (EI) and the Chemical Ionization (CI) are considered as "hard ionizations"**. They operate in low pressure and mostly produce fragment ions due to their high energies that lead to dissociate the analyzed ions. Both the EI and the CI are applicable for low mass molecules in gas state. On the other hand, **soft ionizations** operate with relatively low energies in order to ionize the analytes without dissociating them. The aim of avoiding their dissociation is to allow the detection of the "pseudo-molecular" ion and thus simply detect an expected trace of the analyte or determine its molecular mass when it is unknown. Several examples can be given for such modes, as **the Electrospray (ESI), the Matrix-Assisted-Laser-Desorption-Ionization (MALDI), the Atmospheric Pressure Chemical Ionization (APCI), the Atmospheric Pressure Photo-Ionization (APPI), the Atmospheric Solids Analysis Probe Ionization (ASAP)**, and others.

For the present thesis, only two modes will be exploited: **the Electron Impact and the Electrospray**. The first will be used when performing GC-MS analyses and the second will be used for LC-MS analyses.

2.2.1.1. Electron Impact Ionization

The Electron Impact Ionization was introduced in 1918 by the Canadian-American physicist Arthur Jeffrey Dempster [24]. Its principle of operation is based on an electron cannon that accelerates electrons with a kinetic energy equal to 70 eV in order to bombard the introduced analytes. Once the accelerated electron collides with the molecule, it extracts another electron belonging to the analyte and thus produce a positive-charged radical-ion with an odd number of electrons (Figure M. 3). The high energy of the accelerated electron (70 eV) allows extracting any type of molecular electrons, which leads to the production of low-stable radical-ions. Those meta-stable or unstable ions will then degrade in the gas phase and thus produce different fragment ions that will be analyzed by the mass analyzer, and then detected by the sensor/detector. This detection is the base for the production of an EI-fragments spectrum. Each spectrum constitutes a fingerprint for a given analyte. This fingerprint is reproducible and thus allow proceeding for molecule identification by structural elucidation or by spectral library search.

In metabolomics, EI is one of the common ionization modes used when coupling GC to MS. Its ability to produce reproducible fingerprints for compounds is a major advantage that allows the fast putative identification of metabolites. Nonetheless, the generation of multi-ion spectra produces highly complex GC-MS data. This issue is one of the drawbacks of this ionization mode. However, recent advances in computational data preprocessing succeeded to resolve this issue by developing algorithms as AMDIS [25], metaMS [26] and CAMERA [27] that can perform spectrum deconvolution and assemble fragments belonging to a metabolite in one molecular feature, which allows compressing datasets and facilitating data processing. Another drawback of the EI is that the excessive in-source fragmentation of the ionized analyte prohibits the detection of the “pseudo-molecular” ion. This issue renders difficult the identification of novel metabolites non-reported in the literature so far.

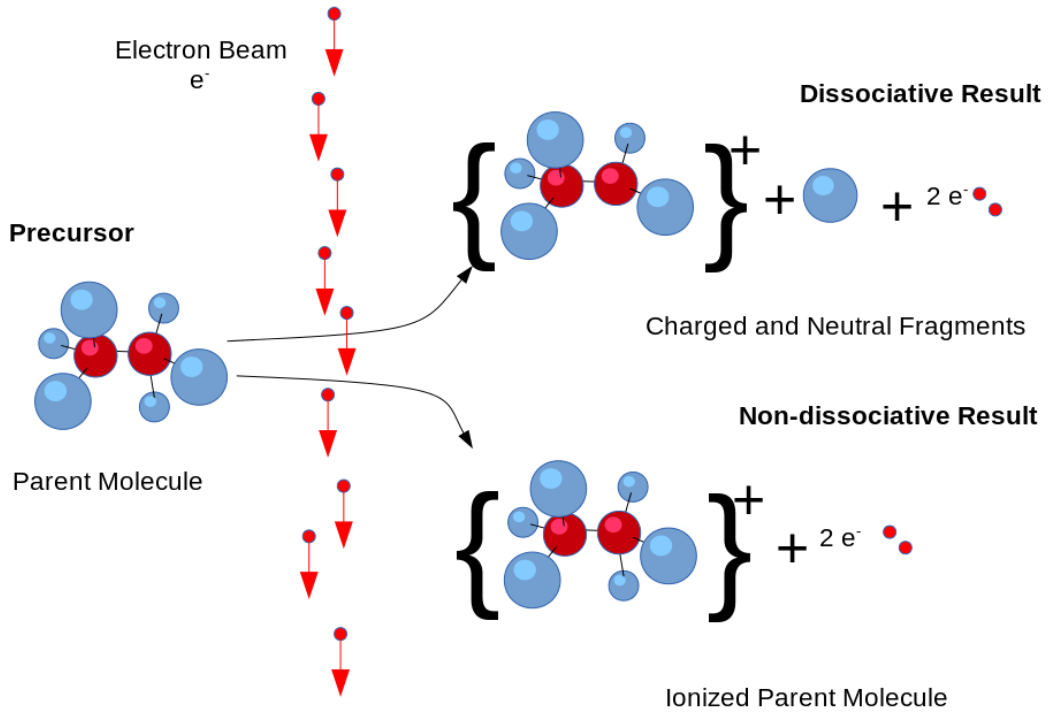


Figure M. 3: A scheme presenting the Electron Impact Ionization. Source: [28].

2.2.1.2. Electrospray Ionization

The **Electrospray Ionization (ESI)** source was introduction in 1984 simultaneously by the team of the Soviet physicist Lydia N. Gall' (Alexandrov *et al.* [29,30]), and the team of the American chemist John Bennett Fenn (Yamashita & Fenn [31]) [32,33].

The introduction of the ESI led to a revolution in Mass Spectrometry and Mass Spectrometry-based analytical methods. As one of the Atmospheric Pressure Ionization (API) techniques, it is based on the application of a high voltage-electric potential between the infusion capillary needle and the MS inlet. The introduced sample solution containing analytes is thus sprayed and distorted into a Taylor cone. The nebulization of the spray is assisted by high temperature and coaxial gas flow application. This mechanism induces the evaporation of the solvent in the droplets, and increases the electrical density on the droplet surface, until it reaches a critical point called the "Rayleigh stability limit". At this stage, a Coulombic explosion occurs. It is caused by electrostatic repulsions. The Coulombic explosion and the occurrence of oxidation-reduction reactions lead to the ionization of analytes and to their transformation into gas state (Figure M. 4) [34,35]. The ionization can be performed in positive mode (ESI+) or in negative mode (ESI–). The ESI+ favors the ionization of molecules with base properties, as they are able

to capture a positive-charged proton (*e.g.* $–NH_2$, –CO, –CN, etc. Also –OH and –COOH can be ionized in this mode). The ESI– favors the ionization of compounds with acid properties that are vulnerable to deprotonation (*e.g.* –OH, –COOH, etc.).

The ESI is considered as a soft ionization technique. It allows the detection of "molecular" ions with less occurrence of ion fragmentation, providing less complex datasets for high throughput analyses. Nonetheless, a molecule ionized in ESI can generate several types of ions, such as "adducts", "clusters", "multi-charged ions", "in-source fragments", as well as ions containing isotopes. In ESI+, the most common adducts are the $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+FA+H]^+$, $[M+FA+Na]^+$, $[M+ACN+H]^+$, and several others. In ESI–, the common types of adducts are fewer, as the $[M-H]^-$, $[M+FA-H]^-$, $[M+Cl]^-$, etc. The clusters consist of an aggregation of 2 or several "M" units with one or several charged species (*e.g.* $[2M+H]^+$, $[2M+Na]^+$, $[3M+2H]^{2+}$, $[2M-H]^-$, etc.). These ions can play an important role for determining the elemental composition of a molecule, by allowing the recognition of the ion specie, and thus narrowing the number of possible formulas. On the other hand, **they produce redundancies in metabolomics datasets, which can enrich or complicate the analyses of results according to the studied context**.
As it is dedicated for liquid phase analyses, the ESI is one of the conventional LC-MS coupling interphases. Therefore, it is **widely used for LC-MS-based metabolomics approaches** [36], mainly for analyzing liquid samples, containing polar and semi-polar, nonvolatile and thermally unstable metabolites (and large (macro)molecules as well). **Despite its advantages for metabolome analyses, the ESI presents several drawbacks. Particularly, its vulnerability to matrix effect and ion suppression phenomenon**. This phenomenon was well documented in the literature [37–39]. It leads to a decrease in metabolite signal in MS due to several causes and mechanisms related to matrix complexity (detailed in Chapter II).

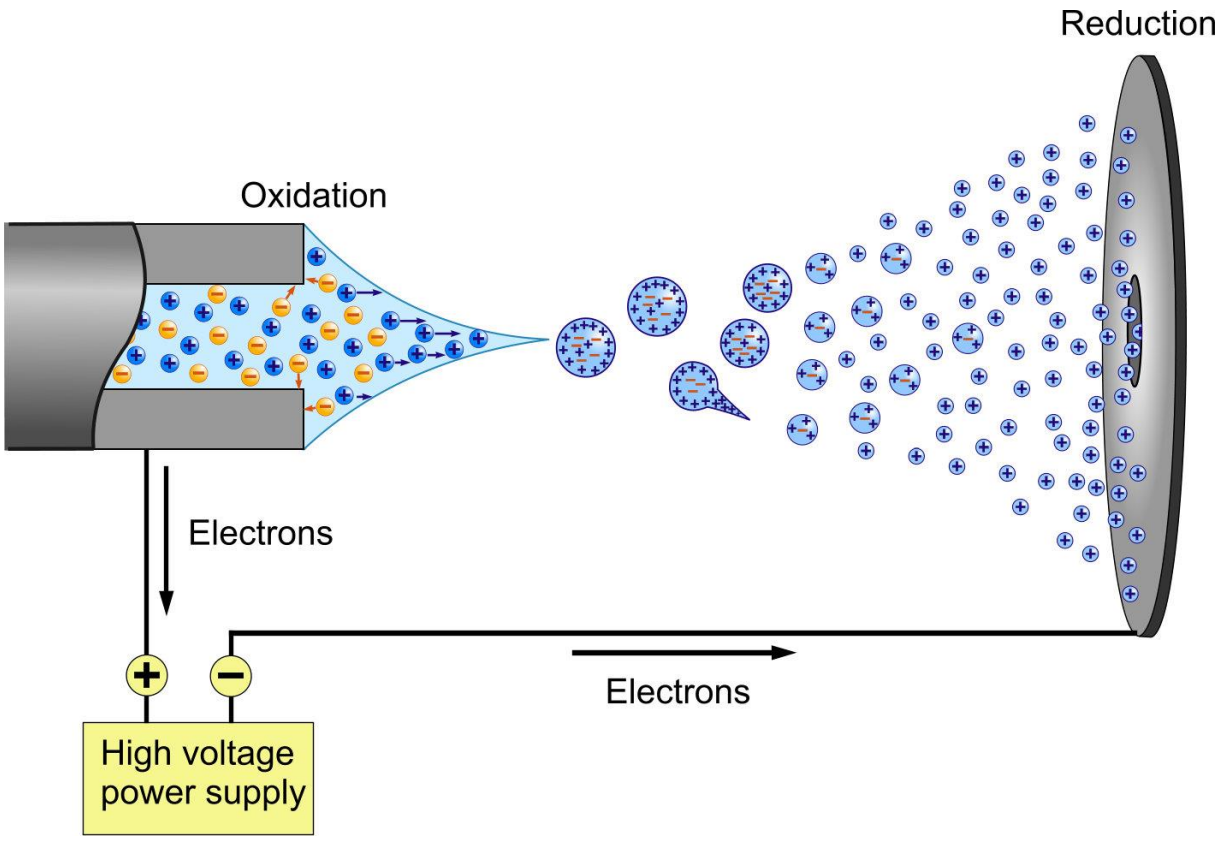


Figure M. 4: A scheme presenting the Electrospray Ionization (in positive mode). Source: [40].

2.2.2. Mass Analyzers

The mass analyzer is the component that performs *m/z* measures. These measures are based on the separation of ions according to their *m/z* ratios. Different mass analyzers are commercialized and used for metabolomics. They can be classified according to their Resolutions[11] and Resolving Powers[12] calculated following the M/ΔM formula. The commonly known and used "Low Resolution" mass analyzers are the Quadrupole (Q), the Ion Trap (IT), and the Linear Time-of-Flight (ToF). The common "High Resolution" mass analyzers are the Reflectron Time-of-Flight, the Fourier Transform Orbitrap, the Magnetic Field and/or Electric Field Sectors, and the Fourier Transform-Ion Cyclotron Resonance (FT-ICR) MS[13] known for its Ultra-High Resolution.

[11] "Resolution": an indicator that measures the sharpness of a detected *m/z* peak. It is calculated following the M/ΔM formula, where M is the measured *m/z*, and the ΔM is the width of the *m/z*, measured at the Full Width at Half Maximum (FWHM) of the peak [41].

[12] "Resolving Power": an indicator that measures the ability of a mass analyzer to separate two *m/z* peaks overlapping (the valley of the overlap is at 10 % of their intensities). It is calculated following the M/ΔM formula, where M is the *m/z* measured at the valley, and ΔM is the difference between the *m/z* ratios of the overlapping peaks [41,42].

[13] At the present, the FT-ICR is called "MRMS" by the manufacturer (Bruker Daltonics). MRMS is the abbreviation of "Magnetic Resonance Mass Spectrometry".

For the current thesis, three different mass analyzers will be used in the experiments. **The Quadrupole** will be used in two modes, 1) as a single **mass analyzer for GC-EI-MS acquisitions**, using a **DSQ II Single Quadrupole** mass spectrometer (Thermo Fisher Scientific) [43], and 2) as an ion transmitter/filter for Tandem Mass Spectrometry acquisitions (MS/MS). Those MS/MS acquisitions will be performed using hybrid mass spectrometers, where several mass analyzers are combined. For the current thesis, the used hybrid mass spectrometers are the **maXis Q/ToF (Bruker Daltonics)** [44], **associating the Quadrupole to a Reflectron ToF**, and **the Q Exactive™ Plus Q/Orbitrap™ (Thermo Fisher Scientific)** [45], **combining a Quadrupole and a C-Trap-Orbitrap complex**. Those two instruments pertain to High Resolution Mass Spectrometry (HRMS). They afford a major advantage: the precise measures of exact *m/z* ratios. This precision allows identifying the elemental compositions of the detected ions with a high confidence. The precision is assessed using the "Error" indicator expressed in Parts-per-Million (ppm). It can be calculated following the formula below:

$$Error = \left[{(m/z_{(Exp)} - m/z_{(Theo)})} \Big/ {m/z_{(Theo)}} \right] \times 10^6$$

Where $m/z_{(Exp)}$ is the experimentally measured *m/z* ratio of the compound, and $m/z_{(Theo)}$ is the theoretical *m/z* ratio calculated for the probable elemental composition.

The HRMS thus allows discriminating between isobaric compounds and isobaric isotopes. Resolving isobaric isotope peaks assures a reliable detection of intensities of isotope contributions. This reliability in isotope patterns detection provides an additional tool for determining with high confidence the elemental composition of the detected ion (the ultimate confidence for elemental composition determination by HRMS passes through the "Seven Golden Rules" that were documented by Tobias Kind and Oliver Fiehn [46]). On the other hand, discriminating the isobaric compounds allows for reliable quantitative analyses in metabolomics, where complex samples containing a wide number and variety of compounds are analyzed. Moreover, the MS/MS provides an additional tool for compounds identification and selectivity.

2.2.2.1. Quadrupole

The Quadrupole mass analyzer was invented in 1953 by the German physicists Wolfgang Paul and Helmut Steinwedel [47,48]. Its principle of operation is based on four parallel metal rods (Figure M. 5). Each pair of opposed rods is connected to the same voltage. A Radio Frequency/Direct Current (RF/DC) alternation is applied between the two pairs of rods. The applied voltages and their alternation determine different trajectories of ions following their *m/z*. At given electric conditions, only resonant ions with certain *m/z* ratios can be transmitted through the analyzer in order to reach the sensor/detector (mostly a dynode electron multiplier). Other ions with unstable trajectories will collide with the rods. They will not be transmitted at theses applied conditions and will not be detected. Thus, the ions with different *m/z* ratios are separated by the variation of the RF/DC during mass analyses. The adapted RF/DC conditions favoring the transmission of a given *m/z* ratio can be calculated following the Mathieu Equation [49]. In fact, to acquire a MS spectrum, the Quadrupole scans the defined *m/z* range by sweeping the different RF/DC conditions belonging to this *m/z* range. The sensor/detector simultaneously registers the signal at each applied RF/DC condition and calculates *m/z* following the applied RF/DC condition at the given time point.

The use of Quadrupole MS for metabolomics presents several advantages [15], mainly its large dynamic range that assures a large-amplitude detection of metabolites with various abundances, and its high scan frequency that provides a highly reliable acquisition of chromatograms, particularly when the MS is coupled to a GC.

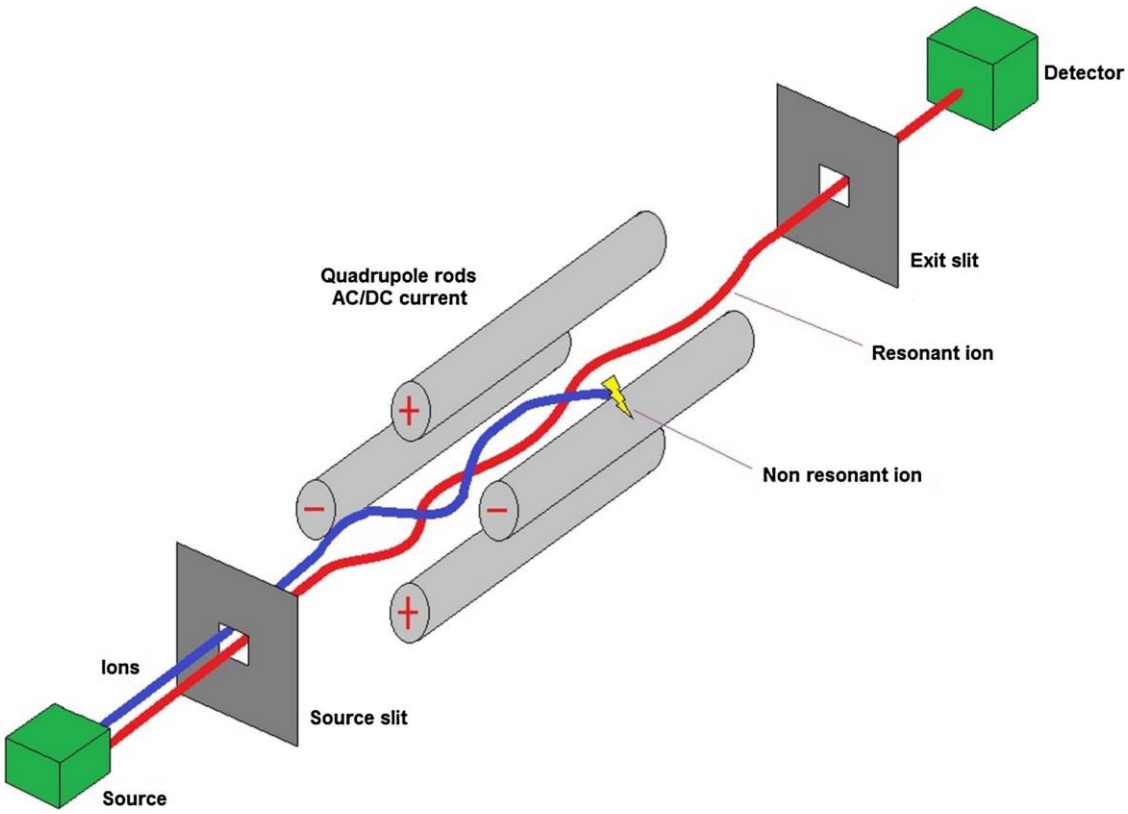


Figure M. 5: The Quadrupole mass analyzer. Source: [50].

2.2.2.2. Reflectron Time-of-Flight

The Reflectron Time-of-Flight was first introduced in 1973 by the Soviet physicist Boris Aleksandrovich Mamyrin and his colleagues [51]. It was an improved version of the linear ToF. **The introduction of the "Reflectron" to the ToF allowed the significant increase of mass analyzer's resolution**, which rendered the Reflectron ToF a high-resolution mass analyzer.

The principle of this analyzer consists in separating the ions according to their *m/z*, following the laws of classical mechanics that can be simplified by the equations below:

$$E_p = E_c$$

$$zU = \frac{1}{2}mv^2$$

$$zU = \frac{1}{2}m\frac{d^2}{t^2}$$

$$t^2 = \frac{1}{2}\frac{d^2}{U}\frac{m}{z}$$

$$t = \frac{d}{\sqrt{2U}}\sqrt{m/z}$$

Where E_p is the potential energy applied in the pulser for accelerating the ions. It depends on the ion charge "z", and the electrical potential "U" imposed by the instrument. E_p is converted into a kinetic energy "E_c". E_c of the ion is translated to velocity "v", where v = d/t. "d" is the distance of flight travelled by the ion in the analyzer, and "t" is the time of flight of the ion. Since the distance "d" travelled by the ions in the analyzer and the potential "U" applied in the pulser are fixed, the discrimination between ions is thus done according to their *m/z* ratios.

The Reflectron is an electrostatic mirror that reflects the ion beam in order to "re-assemble" the ions of the same *m/z* ratio, after they are ejected by the pulse of the pusher. The re-assembling assures the correction of the aberration in *m/z* measuring and thus leads to an exact *m/z* measuring. In fact, when the pulse (*i.e.* the potential energy that accelerates ions) is applied, ions of the same *m/z* get a slightly heterogeneous kinetics energy (due to several factors, including the repulsions between charges). Thus, they are slightly dispersed. During the reflection, ions with a higher kinetic energy will be able to penetrate more in the electrostatic field of the mirror. On the other hand, the penetration of ions with a lower kinetic energy is

lower. As the path of the Reflectron ToF is curvilinear, the two groups of ions are re-assembled and arrive simultaneously to the sensor/detector (Figure M. 6). This phenomenon is called "ion focusing". The simultaneous arrival of ions to the sensor leads to the acquisition of a sharp *m/z* peak, *i.e.*, a **highly resolved peak**. Moreover, geometrically, the Reflectron ensures a longer distance of flight without the need to increase the dimensions of the instrument (Figure M. 6). In fact, the reflection can double the travelled distance (d), which doubles the time of flight and thus serves to improve the discrimination between the different *m/z* ratios.

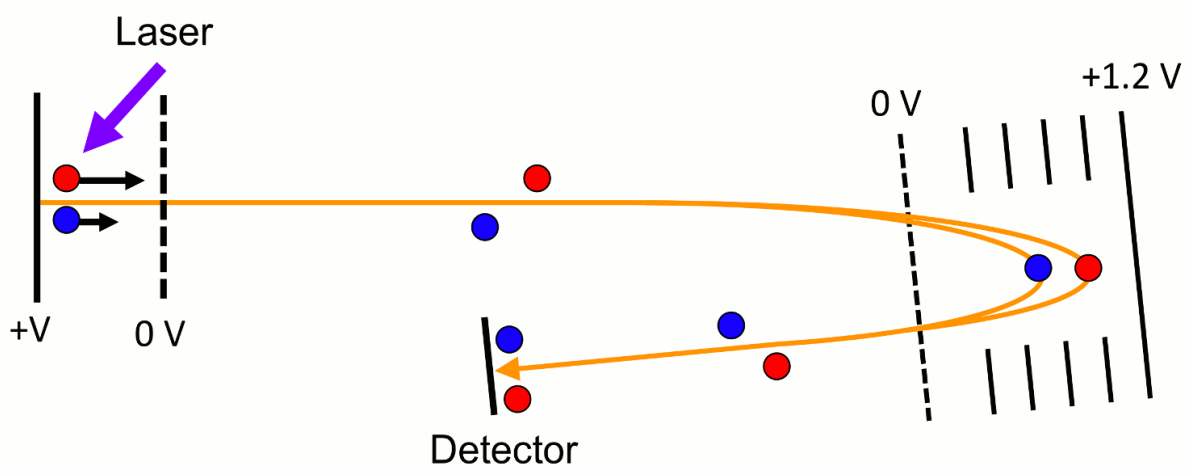


Figure M. 6: Reflectron Time-of-Flight mass analyzer. Source: [52].

The resolution of the ToF increases following the increase of the *m/z* ratio [53,54]. For the maXis Q/ToF, **the maximum resolution is estimated by 80'000 at an *m/z* equal to 1522 (measured at the FWHM)**. It is independent from scan rate as the typical sensor/detector of the ToF is the Secondary Electron Multiplier (SEM) [54]. **The measuring precision should be maintained by a mass-scale calibration**. This calibration is performed by infusing a group of ions (*e.g.* adducts of a same compound) that cover the scan range. As the theoretical *m/z* of the infused ions is known, the mass-scale can be adjusted following those *m/z* ratios.

It is worth mentioning that the precision and the resolution can be affected by the abundance of the ions. In fact, if a highly abundant compound is analyzed using ToF, its ions charge-in-space ratio increases, which leads to repulsions between charges. These repulsions then disperse the ions. By result, the detected *m/z* peak is widened. This phenomenon leads to a loss of resolution. To avoid it, ion filters can be used in order to sharpen the ion beam when transmitting ions to the ToF. However, increasing resolution through this mechanism might engender a loss of sensitivity, as a part of ions is eliminated when sharpening the ion beam by the filters.

Otherwise, the fast scan rate of the ToF is a major advantage that favors its use for GC-MS-based and LC-MS-based metabolomics. In fact, when coupling Chromatography to MS, the fast scan rate assures the acquisition of a higher number of scans/points for chromatograms, which is essential for the quantitative analyses and the detection of compounds with minor abundances. Nonetheless, one drawback is that the increase of scan rate is associated to a loss of sensitivity and a decrease in the dynamic range [54]. Thus, when performing analyses for metabolomics, a compromise should be established between the resolution, the scan rate and the sensitivity.

2.2.3.3. Fourier Transform Mass Spectrometry: C-Trap/Orbitrap complex

The Orbitrap is one of the high-resolution mass analyzers commonly used for metabolic profiling experiments. Various research works contributed to the introduction of this mass analyzer; the first known article was the original paper of Kenneth Hay Kingdon (General Electric Company) that appeared in 1923 in Physical Review [55]. Another paper that can be mentioned is that of Randall D. Knight that appeared in 1981 in Applied Physics Letters [56]. Nonetheless, the current commercialized version of the Orbitrap™ (patented by Thermo Fisher Scientific at present) was introduced in 2000 by the Russian physicist Alexander Alexeyevich Makarov (he worked for HD Technologies Ltd. at that time) [57]. This commercialized version is associated with a **C-Trap mass analyzer** that plays the role of an ion trap and an ion pusher (Figure M. 7).

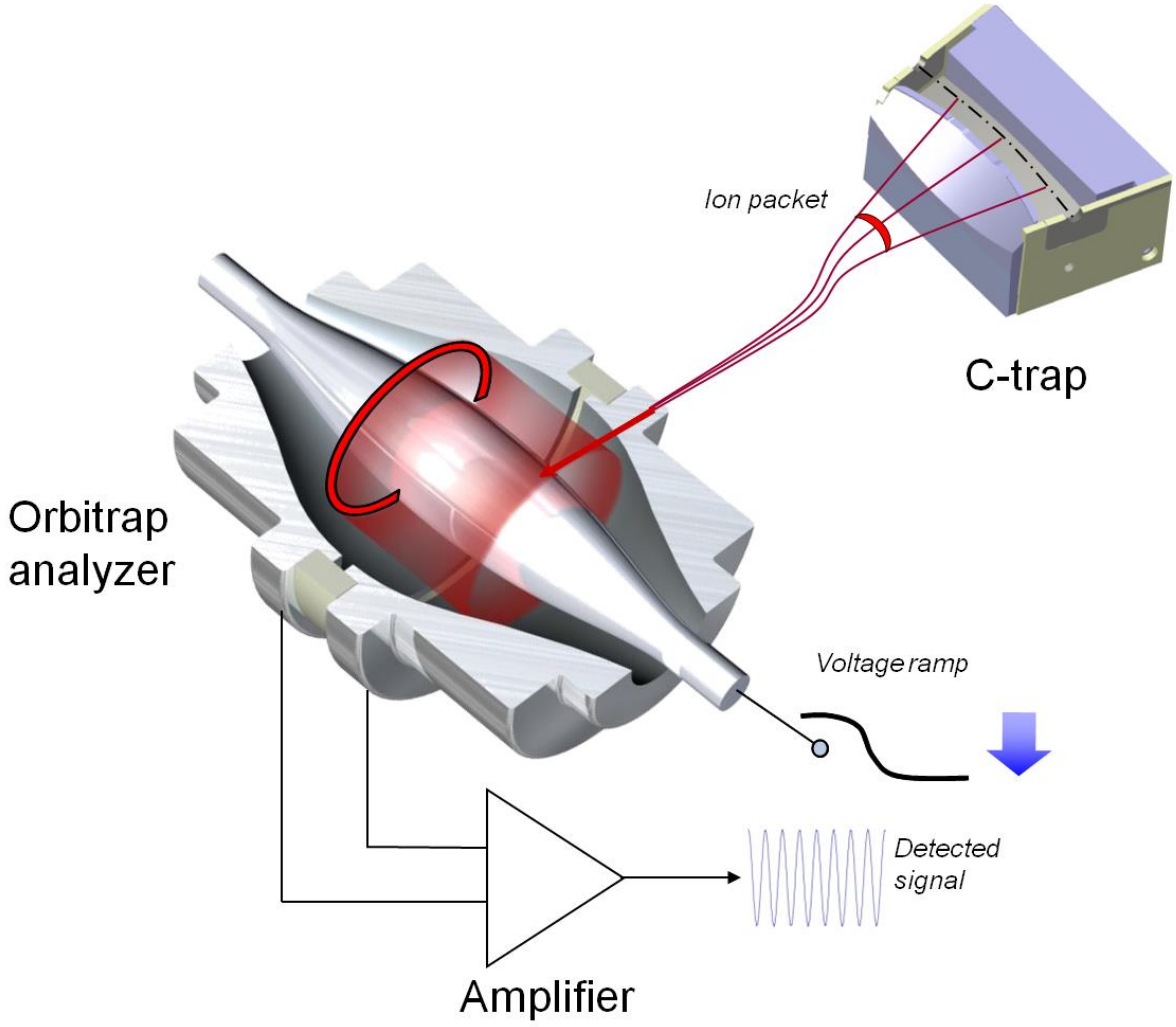


Figure M. 7: The C-Trap/Orbitrap complex. Sources: [54,58].

Ion *m/z* analysis in the C-Trap/Orbitrap complex proceeds first by the arrival of ions to the C-Trap. The latter holds and accumulates the ions and then send them by an electrostatic pulse to the Orbitrap. Ions arrive successively according to their different times of flight. At the instant of their arrival, the Orbitrap regulates its voltage ramp in order to optimize the electrostatic centripetal/centrifugal forces, to allow the introduction of the ions and to trap them into an orbital circuit inside the analyzer (along the central electrode) [59]. Each ion with a proper *m/z* ratio will thus roll with a proper velocity (or time of flight), which will be translated to a proper rolling frequency [60]. This frequency will be registered by a sensor/detector, and then the signal will be amplified. Then the registered signals will undergo a Fourier Transform to calculate the exact *m/z* ratios of each detected ion. The precision of the *m/z* measures is higher when the duration of ion trapping is higher.

Thus, the resolution of the Orbitrap depends on the scan rate (scan frequency). If the scan rate increases, the resolution and the *m/z* measuring precision decrease. For the commercial

instrument used in the current work, *i.e.*, **the Q Exactive™ Plus**, **the maximum resolution (calculated at the FWHM) is 280'000 at a *m/z* of 200 and a scan rate of 1.5 Hz** [45]. The resolution decays following the increase of *m/z* ratio, according to a $R \propto \sqrt{(m/z)^{-1}}$ relation [54,61]. Hence, as the resolution of the Orbitrap decreases with the increase of the scan rate, a compromise between those two factors (*i.e.* resolution and scan rate) should be optimized when the mass spectrometer is coupled to Chromatography. In fact, a higher resolution requires a lower scan rate, which provides a lower number of points to construct the chromatogram. As mentioned previously, a lower number of points may alter the reliability of the quantitative data, as the chromatographic peaks will be constructed with a lower number of samples (*i.e.* scan points). For the Orbitrap, the resolution-scan rate dilemma and the alteration of chromatogram shapes were experimentally demonstrated by Rajski *et al.* [62].

It should be mentioned that other factors can also influence the resolution of the Orbitrap, as the quantity of the ions that are introduced to the analyzer, and the charge-in-space ratio that can alter the resolution when it is relatively high (due to the repulsions between the ions). Here, the Q Exactive™ Plus configuration allows to optimize this factor, as well as the scan rate, by adjusting the Automatic Gain Control Target (AGC Target), and the Maximum Injection Time (Maximum IT). The AGC Target defines the maximum number of charges that can be accumulated in the C-Trap and then introduced to the Orbitrap. The Maximum IT defines the maximum time (in ms) allowed until the C-Trap pushes the ions towards the Orbitrap, regardless the number of accumulated charges. For instance, given that the Maximum IT is set to 100 ms, and the AGC Target to 3e6, if a duration of 100 ms is achieved but the number of charges accumulated in the C-Trap is < 3e6, those accumulated ions are injected in the Orbitrap. Those two parameters act synergically in order to adjust the number of charges that can get access to the Orbitrap, but also influence the scan rate (and consequently the resolution and the chromatogram shape).

2.2.4. Hybrid Mass Spectrometers for Tandem Mass Spectrometry

Tandem Mass Spectrometry (MS/MS) is a dual-stage MS technique that targets the fragmentation of an ion in order to identify its structure. It is also used for quantitative analyses, mainly using Triple Quadrupole (QqQ) mass spectrometers. MS/MS experiments require hybrid mass spectrometers, except for the Ion Trap and the ICR that alone can perform sequential multistage ion fragmentation (MS^n). In fact, the principle of these experiments is based on the selection of an ion in a first mass analyzer (Mostly Quadrupole or Ion Trap). This

ion is called “precursor” or “parent ion”. Once filtrated, it passes into a “Collision Cell”; an intermediate component of the hybrid mass spectrometer where the fragmentation of the precursor takes place. After, the generated fragments are transferred to a second mass analyzer, where their *m/z* are determined. In result, a MS/MS spectrum is acquired. This spectrum mainly consists of the detected fragments. It may contains a certain amount of the precursor, as well as certain of its clusters. This spectrum can be used for structural elucidation or spectral data search. Otherwise, intensities or chromatograms of specific fragments can also be integrated for quantitative analyses with high confidence. The fragments can be acquired either in low resolution (MS/MS) or in high resolution (HRMS/MS[14]), depending on the type of the second mass analyzer. However, as the precursor selection is commonly performed using a Quadrupole or an Ion Trap, the selection is mostly achieved in low resolution/precision, *i.e.* within a minimal *m/z* width of ± *m/z* 0.4-0.5 (which is not the case for the ICR that can select the precursor with ultra-high resolution).

Different types and mechanisms of MS/MS fragmentations are common, as the Collision-Induced Dissociation (CID), the Higher-energy C-trap Dissociation (HCD), the Electron-Transfer Dissociation (ETD), the Infrared Multiphoton Dissociation (IRMPD), the Blackbody Infrared Radiative Dissociation (BIRD), and others. In the present thesis, only two modes will be applied: **the CID and the HCD**.

The principle of operation of the CID is based on the introduction of the precursor to a multipolar collision cell or electrostatic lenses-based collision cell, where the pressure is higher than in the other components of the spectrometer, due to the presence of an inert gas (*e.g.* Argon, Nitrogen). Once the ion is in the collision cell, a particular RF and a defined collision energy are applied through the multipoles (or the lenses). A proper RF value specific for the *m/z* of the selected ion is set in order to provoke its resonance. The ion motion leads to collisions between the precursor and the atoms or molecules of the inert gas, which increases the internal energy of the ion, and finally leads to its dissociation following the applied collision energy. This fragmentation mode is available for the maXis Q/ToF. Its collision cell is a Hexapole cell supplied with Nitrogen gas. It is placed between the Quadrupole and the ToF, as shown in Figure M. 8.

[14] “HRMS/MS” is the commonly used abbreviation to describe High Resolution Tandem Mass Spectrometry. However, an alternative abbreviation; “MS/HRMS” seems to be more appropriate, as the first-stage MS (the selection of the precursor) is achieved in low resolution (except for ICR), and the second-stage MS (the detection of fragments) is achieved in high resolution.

The HCD fragmentation is the same as the CID. A higher RF voltage is however applied in this mode. This fragmentation is available for the Q Exactive™ Plus. It is performed in an Octapole HCD cell supplied with Nitrogen gas. It is placed after the C-Trap-Orbitrap complex as shown in Figure M. 9 and Figure M. 10. Once the fragment ions are produced, they are accumulated in the C-Trap and then sent to the Orbitrap. The accumulation of fragments in the C-Trap enhances the sensitivity or their detection.

It should be mentioned that the major drawback of both the CID and the HCD is that their fragmentation reactions (and consequently their generated spectra) are not easily reproducible [63–65]. In fact, their fragmentation mechanisms are dependent to diverse factors (*e.g.* the geometry of the spectrometer, the applied voltages in the source and/or in the collision cell, the collision cell pressure, the activation time, etc.) [63]. Nonetheless, attempts to investigate and standardize those fragmentations in order to construct MS/MS spectral databases have been reported in the literature and were addressed following diverse approaches [63–65].

On the other hand, for LC-MS/MS experiments, different types of MS/MS acquisitions can be performed (*e.g.* targeted MS/MS, Data Dependent Acquisitions, Data Independent Acquisitions, broadband fragmentations, etc.). In the present thesis, only two types of MS/MS acquisitions will be used: **the targeted MS/MS**, and **the Data Dependent Acquisitions (DDA)**.

In the targeted MS/MS acquisitions, the manipulator should define one or several RT ranges. Each range is specified for a defined ion (the RT range is the elution time of the analyte). The aim is to ask the Quadrupole to select this ion within this RT range, and to assure its exclusive fragmentation after filtering and eliminating all the other ions of co-eluting analytes, and preventing them from being introduced into the collision cell. Thus, all the fragment ions produced and acquired in the spectrum are assumed exclusively originating from the selected analyte.

For the DDA MS/MS, the same principle of the targeted MS/MS is applied. The only difference is that the manipulator does not define any RT ranges. However, the spectrometer will be programmed to select certain ions automatically according to their intensities and their occurrence during the Full MS acquisition. The manipulator sets the optimal parameters that allow the spectrometer to define and select the most relevant ions (as the intensity threshold, the number of precursors to be selected in one RT range, the exclusion conditions, etc.).

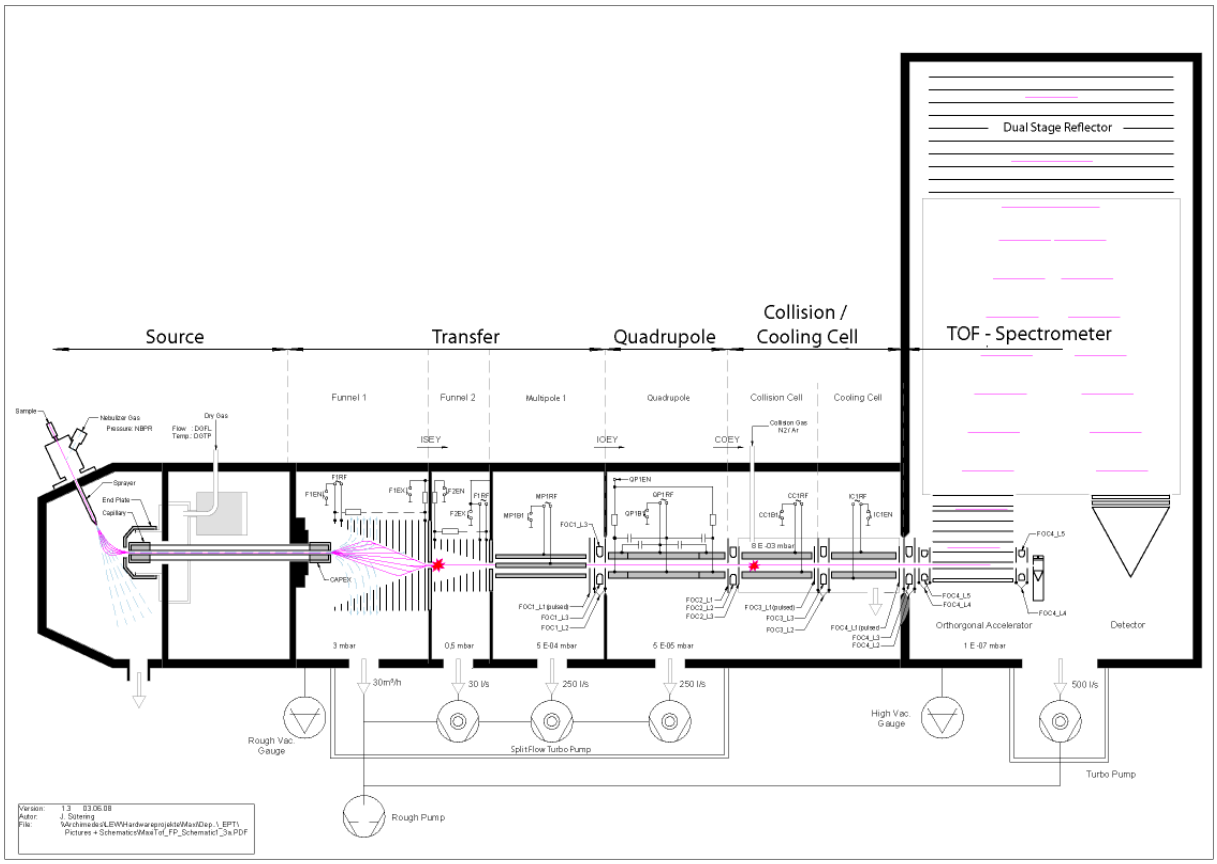


Figure M. 8: The scheme of the maXis Q/ToF inside structure. Source: [44].

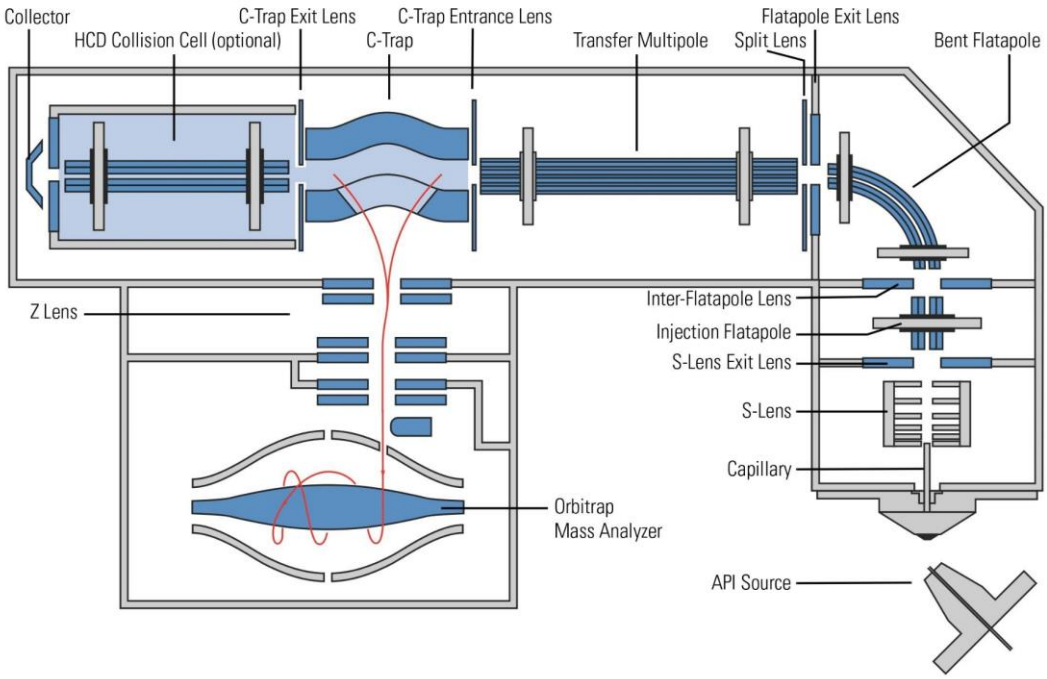


Figure M. 9: The scheme of the Q Exactive™ Plus inside structure. Source: [66].

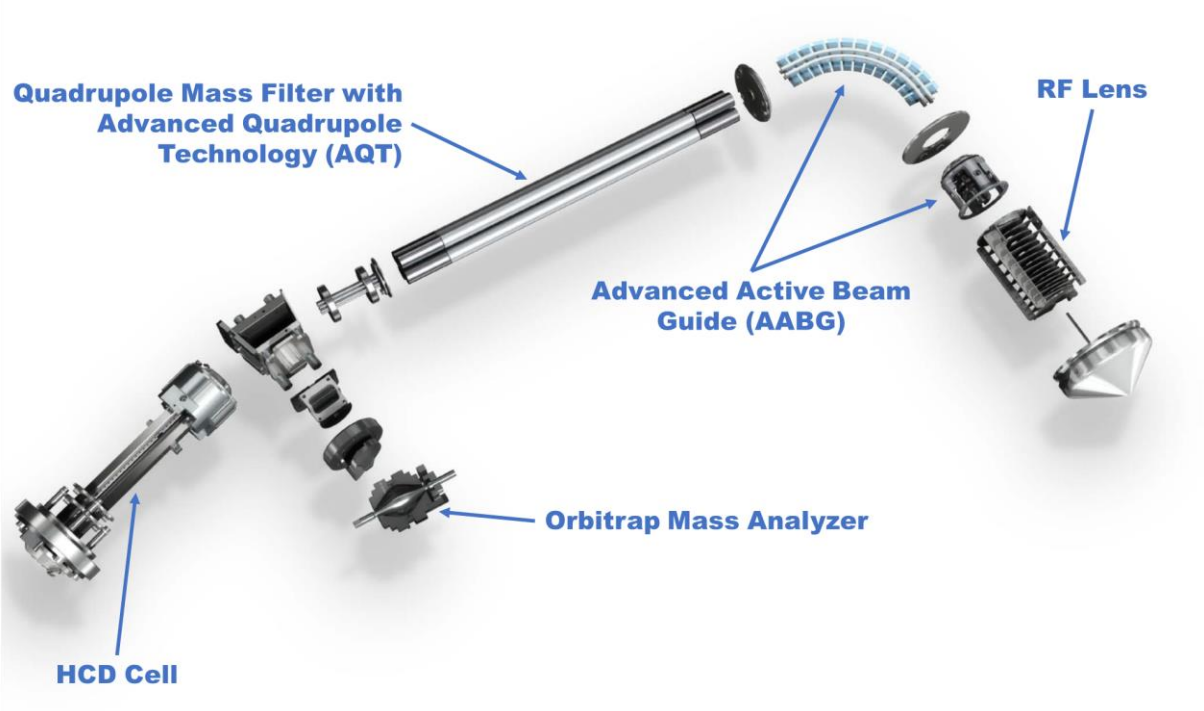


Figure M. 10: A "pseudo-realistic" presentation of the inside of Q Exactive™ Plus. Source: [67].

3. Chemometrics and Statistics

In the present thesis, diverse chemometric and statistical analyses will be used in order to explore GC-EI-MS-based and LC-ESI-HRMS-based metabolomics datasets. Those datasets are generated after raw data preprocessing, *i.e.*, after the transformation of GC-EI-MS or LC-ESI-HRMS raw data into data matrix consisting of observations, variables, and intensities. The observations represent the injected samples, the variables are the detected molecular traces (molecular features/ions), and the intensities are relative to the abundances of the molecular traces in the different samples.

Data preprocessing will be performed using the Workflow4Metabolomics platform [68–71]. The platform afford diverse tools for data curation. Most of them are based on the R software [72,73], as the XCMS algorithm [74] for peak piking, peak grouping, and RT correction, or the CAMERA algorithm for features annotation [27], or the metaMS package dedicated for GC-MS data [26]. The preprocessing pipelines used for the different axes will be described in details in each of the three chapters.

After generating the data matrix, **chemometric and statistical analyses will be applied for data processing and visualization**. They will be performed using the **R-based MetaboAnalyst platform** [75–77], and the **R software** [72,73] (Version 3.3.3). Different models will be used for data analyses in the present thesis, mainly **the Heatmap**, **the Principal Component Analysis (PCA)**, and **the Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA)**, as well as the Euclidean Distance, the *t*-Test, and the Multivariate Empirical Bayes Analysis of Variance (MEBA) [78].

3.1. Heatmap

The Heatmap [79] is a multivariate statistical data analysis that is commonly used in metabolomics, as well as in genomics and transcriptomics. The first registered appearance of this model dates back to 1873 in the "*Atlas statistique de la population de Paris*" [80], written by the French statistician Toussaint Loua.

The Heatmap consists in a three-dimensional representation of a given data matrix (Figure M. 11), where the first dimension is the observations (samples), the second is the variables (molecular features), and the third is the intensity of molecular features. This data visualization model allows observing and interpreting the presence/absence and the abundance of the

different metabolites in the different samples and groups of samples, which can be translated to qualitative and quantitative analyses.

In addition, the hierarchical clustering function is an available tool that can be exploited in the Heatmap by associating it to Dendrograms (Figure M. 11). The clustering can be based either on Euclidean Distances or on Pearson Correlations, and applied either on samples or on features (or on both). In fact, the Euclidean Distances and the Pearson Correlations aim to analyze the similarities between samples' metabolic profiles, or the similarities between metabolites' abundances through samples[15]. This analysis allows for the gathering of samples with similar metabolic profiles, and the assembling of metabolites that show similar behaviors through the different samples. Ultimately, the Dendrogram-based hierarchical clustering of samples allows sorting the groups according to their metabolic response following the applied conditions. For molecular features, the Dendrogram-based hierarchical clustering helps sorting metabolites with similar chemical/biochemical natures or behaviors.

[15] According to Withers *et al.* [81], Euclidean Distances analyze *"the differences in metabolite concentration"*, and Pearson Correlations analyze *"the shapes of metabolite expression profile"*.

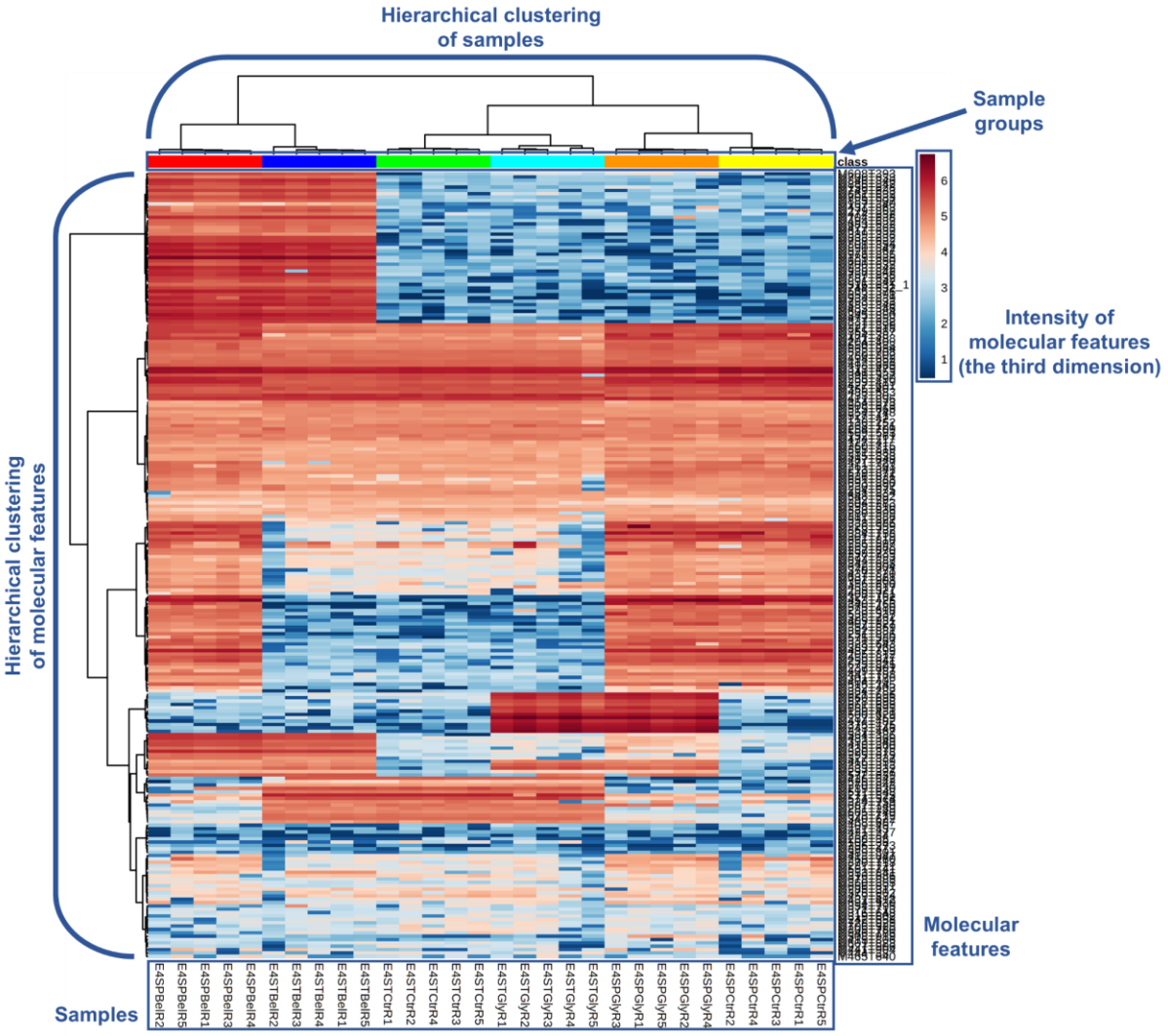


Figure M. 11: the Heatmap.

3.2. Principal Component Analysis

The Principal Component Analysis is a descriptive unsupervised multivariate statistical model. It was introduced in 1901 by the English mathematician Karl Pearson [82].

The PCA principle is based on the aggregation and regression of certain variations carried by certain variables (features). These aggregations and regressions are performed following the correlation and the anti-correlation of the concerned variations through the available observations (samples). The correlations/anti-correlations are calculated without taking into consideration the belonging of observations to the defined groups/conditions (*i.e.*, without *a priori*). Then, the different aggregates of the regressed variations are considered as "Principal Components" (PC). Each PC has a percentage of "explained variance". This percentage is related to the amplitude of the regressed variations (*i.e.* the intensities of the aggregated/regressed variables that are carrying these variations). PCs are then hierarchically

sorted in the descending order of their percentages, *i.e.*, the first PC (PC1) is the PC with the highest percentage, and so on (Figure M. 12). One variable can contribute to one or several PCs, as the PCs are not independent (controversially to the Independent Component Analysis (ICA)). The PCs are however orthogonal to each other's, *i.e.*, their regressed variations are not correlated.

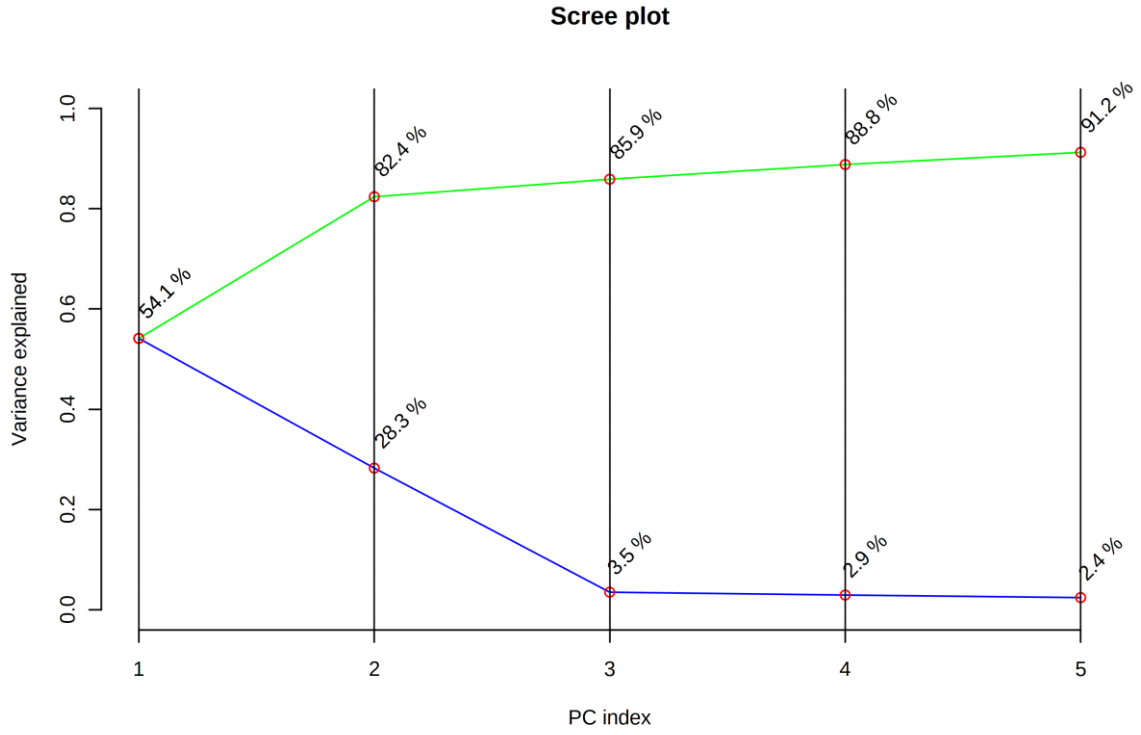


Figure M. 12: The hierarchical sorting of the first five PCs according to their percentages of "explained variance".

Hence, after defining the PCs, each of them can be translated to an axis that can be drawn in a plot. Then, the available observations (samples) are projected on this plot. They are dispersed or clustered following the amount of PCs' contribution inside each observation, *i.e.*, following the abundance of the molecular features belonging to the PCs inside the sample.

An example can be given in Figure M. 13, where four groups of samples are analyzed. The yellow and the green clusters are unpolluted samples ("Ctr") of two different soils ("SP" and "ST", respectively). The red and the blue clusters are samples belonging to the two different soils but polluted ("Bel") with the same herbicide (the yellow and the red belong to the same soil "SP", and the green and the blue are the same soil "ST").

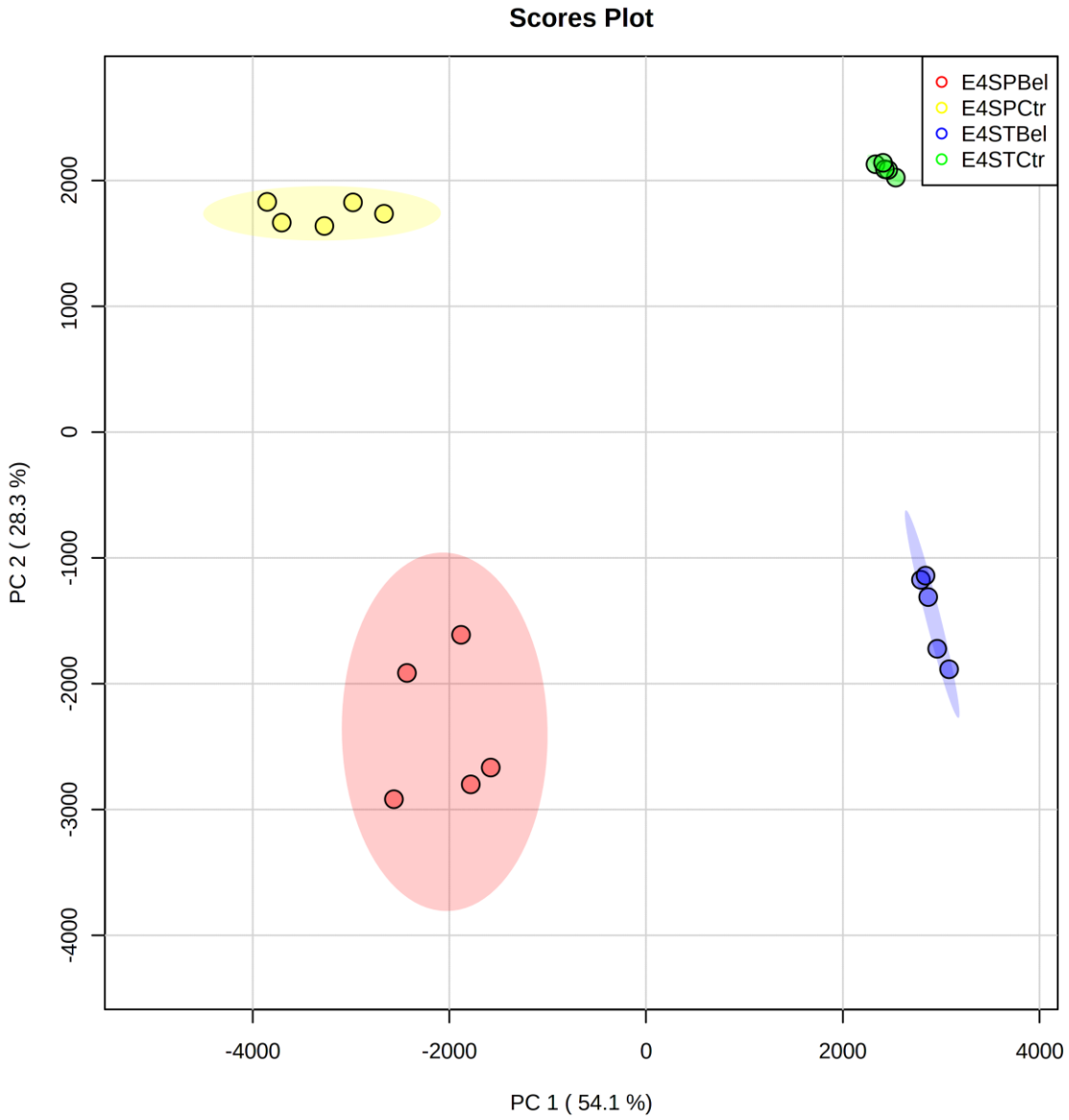


Figure M. 13: An example of the PCA: "SP" and "ST" are two different soil types. "Ctr" are the unpolluted soils, and "Bel" are the polluted soils.

The two SP and ST samples are separated following the PC1 that explains 54.10 % of variations (the horizontal axis), which means that the PC1 aggregates the variations carried by the metabolites that are specific for each soil type. On the other hand, the polluted and the unpolluted soils are separated following the PC2 that explains 28.30 % of variations (the vertical axis), which means that the PC2 aggregates the variations carried by the polluting molecules (assuming that there is no pollution impact on soils' microbial activity). In addition, the amplitudes of variations between soil-specific metabolites are higher than the intensities of the polluting molecules (and/or the detected number of soil-specific metabolites is higher than that of polluting molecules). To support those hypotheses, another type of PCA plot; "the Loadings Plot", can be exploited (Figure M. 14A). The Loadings Plot represents the features

pertaining to the PCs. They also show their contributions (their relative abundances) in the PCs. Another similar representation of data is the "Biplot" that shows both the samples (points) and the features (arrows). The advantage of the Biplot is that it reveals the correlation between the different samples and the different features (Figure M. 14B).

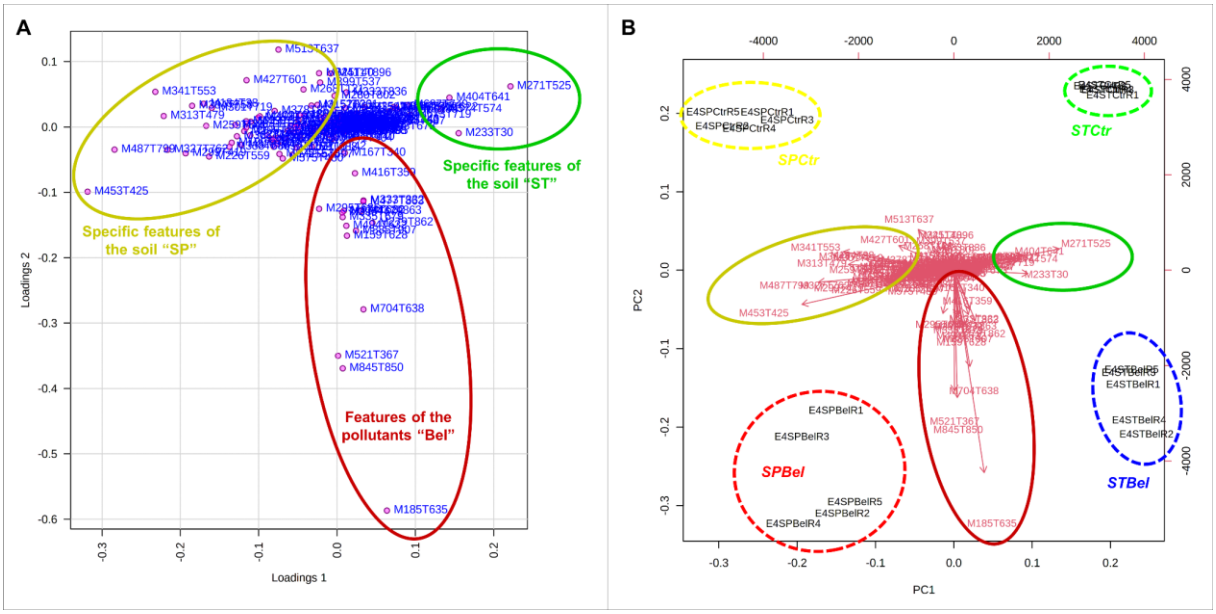


Figure M. 14: A: Loadings Plot, **B:** Biplot.
The dark yellow ellipse highlights the specific features of the soil "SP", they positively correlate between each other's and negatively correlate (anti-correlate) with the specific features of the soil "ST" (dark green ellipse). Thus, both the SP- and ST-specific features constitute the PC1.
The dark red ellipse highlights the features belonging to the pollutants issued from the herbicide. They constitute the PC2 but additionally, some of them slightly contribute to the PC1, as they are slightly polarized towards the STBel samples (towards the right side of the plot). This polarization can be explained by the fact that those pollutants are relatively more detected in the ST soil.

3.3. Orthogonal Projections to Latent Structures Discriminant Analysis

The Orthogonal Projections to Latent Structures Discriminant Analysis was introduced in 2002 by the Swedish chemometricians Johan Trygg and Svante Wold [83]. It was developed on the basis of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm developed by the Norwegian econometrician Herman Wold[16] [84]. It is an explicative supervised multivariate statistical model that aims to explore and assess the discrimination between two (or more) defined groups of samples [83,85]. The OPLS-DA consists in two types of components: the Predictive "p" and the Orthogonal "o". The Predictive is constituted by the aggregation and the regression of the variations (carried by the variables) that discriminate the two compared groups. The Orthogonal components consist in the aggregation of certain variations that (anti-)correlate between each other's, but explain unknown systematic variations occurring in the

[16] The father of Svante Wold. He was born in Norway but lived and worked in Sweden.

dataset. Those unknown systematic variations are orthogonal to (uncorrelated to/independent from) the variations explained by the Predictive "p". In the OPLS-DA, the random non-systematic variations are excluded after they are filtrated and left in the "residuals" [86].

The main advantage of the OPLS-DA is its ability to reveal confidently the variations that are exclusively correlating with the studied factor, by filtering all the orthogonal and random variations that might influence the discrimination of the compared groups. Moreover, the investigation of the systematic orthogonal variations explained by the "o" component allows assessing the confidence in the discrimination. In fact, it can reveal unknown and unexpected variations that might influence the analyses, such as instrumental drifts, sampling issues, biological variations, or other important factors that were not taken into consideration during the experimental design (*e.g.* the age, the gender, the harvesting time, etc.). One risk is that if those orthogonal variations dominate the dataset, they can hide the investigated effects or skew the interpretation of the factor-related variations.

In order to investigate the significance and the confidence of the discrimination between the metabolic profiles of the two samples, the OPLS-DA Cross-Validation test can be used (Figure M. 15). In fact, this model gives three indicators to assess the Predictive and the Orthogonal components: 1) the **R2X** that indicates the **variance** explained by the regressed variations (carried by the variables) of the component (in percentage), 2) the **R2Y** that indicates the (anti-)correlation of the **two groups of samples** to the variations explained by the component, which explains the robustness of the component, and 3) the **Q2** that indicates the **predictivity** of the component, *i.e.*, its power/ability to detect the difference and to predict the presence of two groups of samples. If R2Y > R2X, R2Y > Q2 and R2Y – Q2 ≤ 30 %, the component is considered as robust. If the Q2 > 50 %, the predictivity of the component is considered acceptable for metabolomics studies (Q2 > 90 % is ideal) [86].

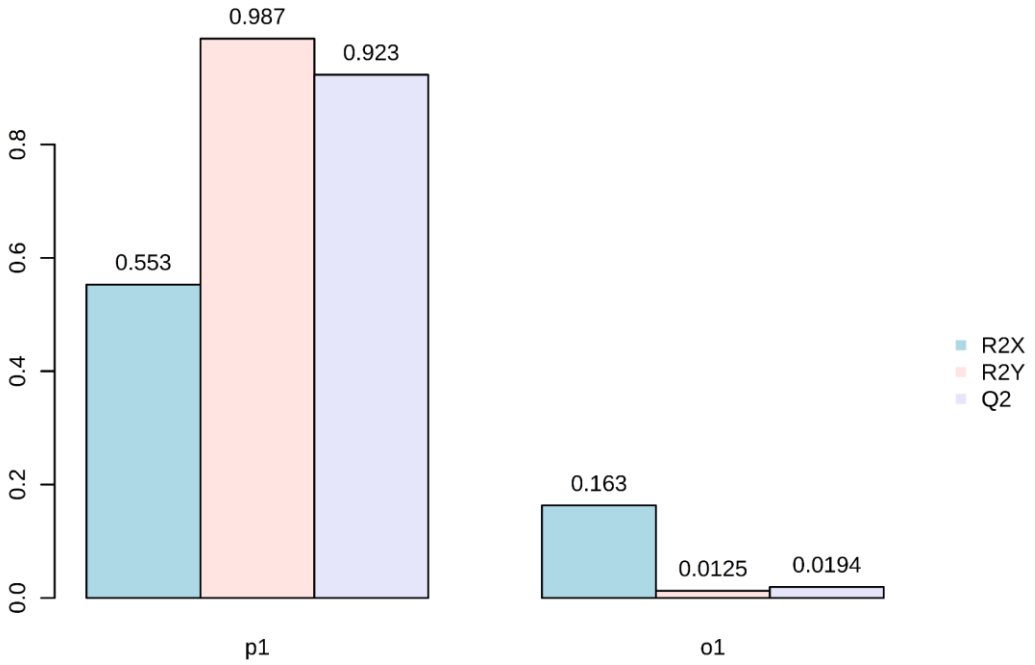


Figure M. 15: OPLS-DA Cross-Validation test.

On the other hand, the OPLS-DA is capable to search and identify the **discriminant markers** with high confidence. Markers discovery can be performed by applying the **OPLS-DA S-Plot** as shown in Figure M. 16. The S-Plot consists in two dimensions: the p[1] (horizontal axis), and the p(corr)[1] (vertical axis). On this Plot, the different variables (molecular features) are projected according to the two dimensions. The p[1] explains the contribution of the feature in the discrimination between the two groups of samples. It is directly related to the difference between feature's abundances in the two groups. The p(corr)[1] explains the significance of this difference. It is directly related to feature's "intra-group" abundance variation. If the magnitude of the abundance is significantly wide, but the "intra-group" variation is relatively high, the consideration of this feature as a marker of discrimination is less confident. In fact, as the "intra-group" variation increases the incertitude, the correlation of the abundance difference to the examined factor cannot be confidently assumed.

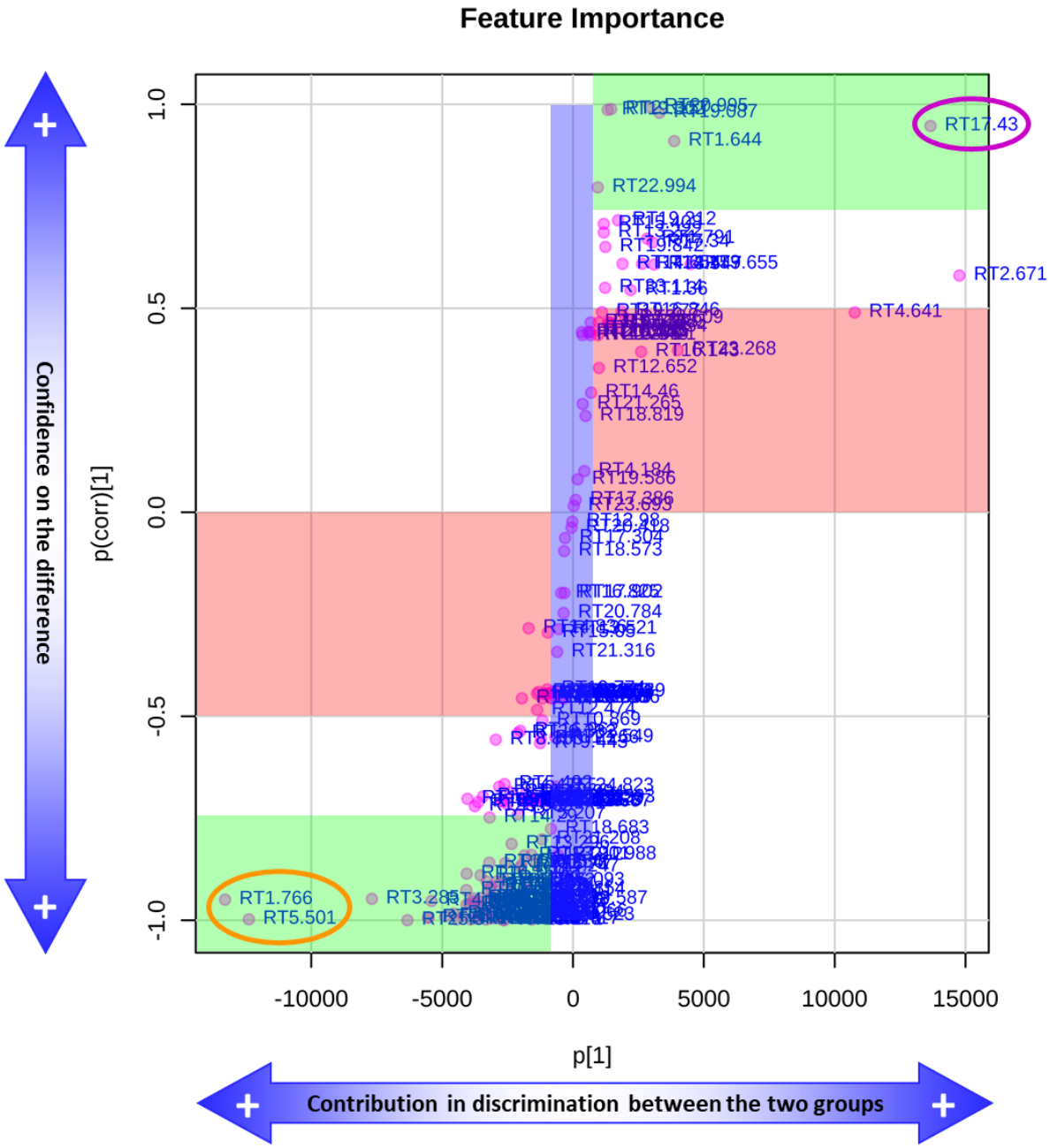


Figure M. 16: OPLS-DA S-Plot.
The S-Plot facilitates the mining of the features of interest, since it correlates the variations of features' abundances to the studied factor (factor examined by the different conditions applied to the studied groups).
The p[1] explains the magnitudes of the differences between variables' abundances in the two conditions: the further is the feature from the 0, the higher is the magnitude of its variation between the two groups.
The p(corr)[1] explains the reliability/significance of the differences between variables' abundances, by assessing their "intra-condition" variations: the further is the feature from the 0, the lower is its "intra-group" variation, and thus the reliability of its variation significance is higher.
Hence, features projected in the green zone are significant markers of discrimination, those projected in the red zone are risky, and those projected in the blue zone are non-discriminant. In this given example, the purple ellipse highlights the most confidently significant marker overexpressed in the first group, and the orange ellipse highlights the most confidently significant markers overexpressed in the second group. [86–88].

List of Abbreviations

AGC: Automatic Gain Control

AMDIS: Automated Mass spectral Deconvolution and Identification System

APCI: Atmospheric Pressure Chemical Ionization

API: Atmospheric Pressure Ionization

APPI: Atmospheric Pressure Photo-Ionization

ASAP: Atmospheric Solids Analysis Probe Ionization

BIRD: Blackbody Infrared Radiative Dissociation

CAR: Carboxen

CI: Chemical Ionization

CID: Collision-Induced Dissociation

DC: Direct Current

DDA: Data Dependent Acquisitions

DVB: Divinylbenzene

EI: Electron Impact

EMF: Environmental Metabolic Footprinting

ESI: Electrospray Ionization

ETD: Electron-Transfer Dissociation

FA: Formic Acid

FT: Fourier Transform

FWHM: Full Width at Half Maximum

GC: Gas Chromatography

HCD: Higher-energy C-trap Dissociation

HILIC: Hydrophilic Interaction Chromatography

HPLC: High Performance Liquid Chromatography

HRMS: High Resolution Mass Spectrometry

HS: Headspace

ICA: Independent Component Analysis

ICR: Ion Cyclotron Resonance

IRMPD: Infrared Multiphoton Dissociation

IT: Ion Trap

LC: Liquid Chromatography

m/z: mass-to-charge ratio

MALDI: Matrix-Assisted-Laser-Desorption-Ionization

MEBA: Multivariate Empirical Bayes Analysis of Variance

MRMS: Magnetic Resonance Mass Spectrometry

MS/MS: Tandem Mass Spectrometry

MS: Mass Spectrometry

NIPALS: Nonlinear Iterative Partial Least Squares

OPLS-DA: Orthogonal Projections to Latent Structures Discriminant Analysis

PC: Principal Component

PCA: Principal Component Analysis

PDMS: Polydimethylsiloxane

PEG: Polyethylene glycol

pH: potential of Hydrogen

ppm: Parts-per-Million

Q: Quadrupole

QqQ: Triple Quadrupole

RF: Radio Frequency

RI: Kováts Retention Indices

RPLC: Reverse Phase Liquid Chromatography

RT: Retention Time

SEM: Secondary Electron Multiplier

SPME: Solid Phase Microextraction

ToF: Time-of-Flight

UHPLC: Ultra High Performance Liquid Chromatography

References

[1] W. Henry, J. Banks, III. Experiments on the quantity of gases absorbed by water, at different temperatures, and under different pressures, Philosophical Transactions of the Royal Society of London. 93 (1803) 29–274. https://doi.org/10.1098/rstl.1803.0004.

[2] H.L. Le Chatelier, Sur un énoncé général des lois des équilibres chimiques, C.R. Hebd. Seances Acad. Sci. 99 (1884) 786–789. https://gallica.bnf.fr/ark:/12148/bpt6k3055h/f786.item.langFR (accessed October 14, 2020).

[3] S.A. Sassman, L.S. Lee, Sorption of Three Tetracyclines by Several Soils: Assessing the Role of pH and Cation Exchange, Environ. Sci. Technol. 39 (2005) 7452–7459. https://doi.org/10.1021/es0480217.

[4] T.R. Keppel, M.E. Jacques, D.D. Weis, The use of acetone as a substitute for acetonitrile in analysis of peptides by liquid chromatography/electrospray ionization mass spectrometry: Acetone vs. acetonitrile in LC/ESI-MS analysis of peptides, Rapid Commun. Mass Spectrom. 24 (2010) 6–10. https://doi.org/10.1002/rcm.4352.

[5] J. Heaton, M.D. Jones, C. Legido-Quigley, R.S. Plumb, N.W. Smith, Systematic evaluation of acetone and acetonitrile for use in hydrophilic interaction liquid chromatography coupled with electrospray ionization mass spectrometry of basic small molecules: Evaluation of acetone and acetonitrile for use in HILIC/ESI-MS, Rapid Commun. Mass Spectrom. 25 (2011) 3666–3674. https://doi.org/10.1002/rcm.5271.

[6] J. Heaton, N.W. Smith, Advantages and Disadvantages of HILIC; a Brief Overview, Chromatogr. Today. 5 (2012) 44–47. https://www.chromatographytoday.com/article/electrophoretic-separations/35/james_heaton_and_norman_w._smith/advantages_and_disadvantagesof_hilic_a_brief_overview/1204 (accessed October 15, 2020).

[7] R.P. Belardi, J.B. Pawliszyn, The Application of Chemically Modified Fused Silica Fibers in the Extraction of Organics from Water Matrix Samples and their Rapid Transfer to Capillary Columns, Water Quality Research Journal. 24 (1989) 179–191. https://doi.org/10.2166/wqrj.1989.010.

[8] J. Pawliszyn, Handbook of Solid Phase Microextraction, First Edition, Elsevier, Oxford, U.K., 2012. https://doi.org/10.1016/C2011-0-04297-7.

[9] J. Pawliszyn, Theory of Solid-Phase Microextraction, Journal of Chromatographic Science. 38 (2000) 270–278. https://doi.org/10.1093/chromsci/38.7.270.

[10] R.E. Shirey, 4 - SPME Commercial Devices and Fibre Coatings, in: J. Pawliszyn (Ed.), Handbook of Solid Phase Microextraction, First Edition, Elsevier, Oxford, U.K., 2012: pp. 99–133. https://doi.org/10.1016/B978-0-12-416017-0.00004-8 (accessed October 15, 2020).

[11] É.A. Souza-Silva, R. Jiang, A. Rodríguez-Lafuente, E. Gionfriddo, J. Pawliszyn, A critical review of the state of the art of solid-phase microextraction of complex matrices I. Environmental analysis, TrAC Trends in Analytical Chemistry. 71 (2015) 224–235. https://doi.org/10.1016/j.trac.2015.04.016.

[12] М.С. Цвет, О новой категории адсорбционных явлений и о применении их к биохимическому анализу, Труды Варшавского общества естествоиспытателей, отд. Биологии. 14 (1903) 20–39. http://books.e-heritage.ru/book/10078100 (accessed October 18, 2020).

[13] M.S. Tsvet, Michael Tswett's First Paper on Chromatography, M. Woelm, Eschwege, Germany, 1954. https://books.google.fr/books/about/Michael_Tswett_s_First_Paper_on_Chromato.html?id=XvvfwAEACAAJ&redir_esc=y (accessed October 9, 2020).

[14] J.J. van Deemter, F.J. Zuiderweg, A. Klinkenberg, Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography, Chemical Engineering Science. 50 (1956) 3869–3882. https://doi.org/10.1016/0009-2509(96)81813-6.

[15] M. Bedair, L.W. Sumner, Current and emerging mass-spectrometry technologies for metabolomics, TrAC Trends in Analytical Chemistry. 27 (2008) 238–250. https://doi.org/10.1016/j.trac.2008.01.006.

[16] Z. Liu, S. Rochfort, Recent progress in polar metabolite quantification in plants using liquid chromatography-mass spectrometry: Recent progress in polar metabolite quantification in plants, J. Integr. Plant Biol. 56 (2014) 816–825. https://doi.org/10.1111/jipb.12181.

[17] D.-Q. Tang, L. Zou, X.-X. Yin, C.N. Ong, HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS: HILIC-MS FOR METABOLOMICS, Mass Spec Rev. 35 (2016) 574–600. https://doi.org/10.1002/mas.21445.

[18] Z. Liu, S. Rochfort, A fast liquid chromatography–mass spectrometry (LC–MS) method for quantification of major polar metabolites in plants, Journal of Chromatography B. 912 (2013) 8–15. https://doi.org/10.1016/j.jchromb.2012.10.040.

[19] H. Ghosson, A. Schwarzenberg, F. Jamois, J.-C. Yvin, Simultaneous untargeted and targeted metabolomics profiling of underivatized primary metabolites in sulfur-deficient barley by ultra-high performance liquid chromatography-quadrupole/time-of-flight mass spectrometry, Plant Methods. 14 (2018) 62. https://doi.org/10.1186/s13007-018-0329-0.

[20] E. Kováts, Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone, Helvetica Chimica Acta. 41 (1958) 1915–1932. https://doi.org/10.1002/hlca.19580410703.

[21] E.Sz. Kováts, P.B. Weisz, Über den Retentionsindex und seine Verwendung zur Aufstellung einer Polaritätsskala für Lösungsmittel, Berichte Der Bunsengesellschaft Für Physikalische Chemie. 69 (1965) 812–820. https://doi.org/10.1002/bbpc.19650690911.

[22] L.S. Ettre, The Kováts Retention Index System, Anal. Chem. 36 (1964) 31A-41A. https://doi.org/10.1021/ac60214a727.

[23] M. Lucero, R. Estell, M. Tellez, E. Fredrickson, A retention index calculator simplifies identification of plant volatile organic compounds, Phytochemical Analysis. 20 (2009) 378–384. https://doi.org/10.1002/pca.1137.

[24] A.J. Dempster, A new Method of Positive Ray Analysis, Phys. Rev. 11 (1918) 316–325. https://doi.org/10.1103/PhysRev.11.316.

[25] AMDIS, Mass Spectrometry Data Center. (n.d.). https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:amdis (accessed October 18, 2020).

[26] R. Wehrens, G. Weingart, F. Mattivi, metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics, Journal of Chromatography B. 966 (2014) 109–116. https://doi.org/10.1016/j.jchromb.2014.02.051.

[27] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets, Anal. Chem. 84 (2012) 283–289. https://doi.org/10.1021/ac202450g.

[28] E. Mason, Diagram of Electron Ionization, 2015. https://commons.wikimedia.org/wiki/File:Electron_Ionization.svg (accessed October 13, 2020). Rights: Evan Mason / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0).

[29] М. Александров, Л. Галль, Н. Краснов, В. Николаев, В. Павленко, В. Шкуров, Экстракция ионов из растворов при атмосферном давлении-новый метод масс-спектрометрического анализа биоорганических веществ, Дан СССР. 277 (1984) 379–383. http://mass-spektrometria.ru/download/2005(2)/T2N1.pdf (accessed October 18, 2020).

[30] M.L. Alexandrov, L.N. Gall, N.V. Krasnov, V.I. Nikolaev, V.A. Pavlenko, V.A. Shkurov, Extraction of ions from solutions under atmospheric pressure as a method for mass spectrometric analysis of bioorganic compounds, Rapid Communications in Mass Spectrometry. 22 (2008) 267–270. https://doi.org/10.1002/rcm.3113.

[31] M. Yamashita, J.B. Fenn, Electrospray ion source. Another variation on the free-jet theme, J. Phys. Chem. 88 (1984) 4451–4459. https://doi.org/10.1021/j150664a002.

[32] V.G. Zaikin, A.A. Sysoev, Mass Spectrometry in Russia, Eur J Mass Spectrom (Chichester). 19 (2013) 399–452. https://doi.org/10.1255/ejms.1248.

[33] A. Bruins, A General Perspective on the Development of Liquid Chromatography Mass Spectrometry (LC/MS), in: M.L. Gross, R.M. Caprioli (Eds.), The Encyclopedia of Mass Spectrometry, Elsevier, Boston, 2016: pp. 159–171. https://doi.org/10.1016/B978-0-08-043848-1.00022-5.

[34] L. Konermann, E. Ahadi, A.D. Rodriguez, S. Vahidi, Unraveling the Mechanism of Electrospray Ionization, Anal. Chem. 85 (2013) 2–9. https://doi.org/10.1021/ac302789c.

[35] G.J. Van Berkel, V. Kertesz, Using the Electrochemistry of the Electrospray Ion Source, Anal. Chem. 79 (2007) 5510–5520. https://doi.org/10.1021/ac071944a.

[36] A. Damont, M.-F. Olivier, A. Warnet, B. Lyan, E. Pujos-Guillot, E.L. Jamin, L. Debrauwer, S. Bernillon, C. Junot, J.-C. Tabet, F. Fenaille, Proposal for a chemically consistent way to annotate ions arising from the analysis of reference compounds under ESI conditions: A prerequisite to proper mass spectral database constitution in metabolomics, J Mass Spectrom. 54 (2019) 567–582. https://doi.org/10.1002/jms.4372.

[37] J.-P. Antignac, K. de Wasch, F. Monteau, H. De Brabander, F. Andre, B. Le Bizec, The ion suppression phenomenon in liquid chromatography–mass spectrometry and its consequences in the field of residue analysis, Analytica Chimica Acta. 529 (2005) 129–136. https://doi.org/10.1016/j.aca.2004.08.055.

[38] F. Gosetti, E. Mazzucco, D. Zampieri, M.C. Gennaro, Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry, Journal of Chromatography A. 1217 (2010) 3929–3937. https://doi.org/10.1016/j.chroma.2009.11.060.

[39] A. Furey, M. Moriarty, V. Bane, B. Kinsella, M. Lehane, Ion suppression; A critical review on causes, evaluation, prevention and applications, Talanta. 115 (2013) 104–122. https://doi.org/10.1016/j.talanta.2013.03.048.

[40] A. Dahlin, ESI positive mode, 2015. https://www.flickr.com/photos/visualize_your_science/21589986840/in/photostream/ (accessed October 13, 2020). Rights: Andreas Dahlin / CC BY (https://creativecommons.org/licenses/by/2.0).

[41] K.K. Murray, R.K. Boyd, M.N. Eberlin, G.J. Langley, L. Li, Y. Naito, Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013), Pure and Applied Chemistry. 85 (2013) 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06.

[42] A.D. McNaught, A. Wilkinson, Compendium of Chemical Terminology, 2.3.3, International Union of Pure and Applied Chemistry (IUPAC), 2014. http://goldbook.iupac.org/ (accessed October 9, 2020).

[43] DSQ II Single Quadrupole GC/MS, (2006). http://www.thermo.com.cn/Resources/200802/productPDF_26943.pdf (accessed October 13, 2020).

[44] maXis User Manual, (2008). http://202.202.232.211/attachments/193/maXis_User_Manual.pdf (accessed October 10, 2020).

[45] Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ Mass Spectrometer, Thermo Fisher Scientific. (n.d.). https://www.thermofisher.com/order/catalog/product/IQLAAEGAAPFALGMBDK (accessed October 9, 2020).

[46] 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.

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

[48] W. Paul, H. Steinwedel, Apparatus for separating charged particles of different specific charges, US2939952A, 1960. https://patents.google.com/patent/US2939952A/ (accessed October 10, 2020).

[49] É.L. Mathieu, Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique., Journal de Mathématiques Pures et Appliquées. 13 (1868) 137–203. http://sites.mathdoc.fr/JMPA/PDF/JMPA_1868_2_13_A8_0.pdf (accessed October 10, 2020).

[50] G. Santoiemma, Recent methodologies for studying the soil organic matter, Applied Soil Ecology. 123 (2018) 546–550. https://doi.org/10.1016/j.apsoil.2017.09.011.

[51] B.A. Mamyrin, V.I. Karataev, D.V. Shmikk, V.A. Zagulin, The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution, Zh. Eksp. Teor. Fiz. 64 (1973) 82–89. http://www.jetp.ac.ru/cgi-bin/dn/e_037_01_0045.pdf (accessed October 9, 2020).

[52] K.K. Murray, In the reflection, the higher energy ion (red) takes a longer path but arrives at the detector at the same time as the lower energy ion (blue) of the same mass., 2017. https://commons.wikimedia.org/wiki/File:Reflectron_schematic.gif (accessed October 13, 2020). Rights: K. K. Murray / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0).

[53] S. Beck, A. Michalski, O. Raether, M. Lubeck, S. Kaspar, N. Goedecke, C. Baessmann, D. Hornburg, F. Meier, I. Paron, N.A. Kulak, J. Cox, M. Mann, The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics, Mol Cell Proteomics. 14 (2015) 2014–2029. https://doi.org/10.1074/mcp.M114.047407.

[54] R.A. Zubarev, A. Makarov, Orbitrap Mass Spectrometry, Anal. Chem. 85 (2013) 5288–5296. https://doi.org/10.1021/ac4001223.

[55] 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.

[56] R.D. Knight, Storage of ions from laser-produced plasmas, Appl. Phys. Lett. 38 (1981) 221–223. https://doi.org/10.1063/1.92315.

[57] A. Makarov, Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis, Anal. Chem. 72 (2000) 1156–1162. https://doi.org/10.1021/ac991131p.

[58] Thermo Fisher Scientific, Cross-section of the C-trap and Orbitrap analyzer (ion optics and differential pumping not shown). Ion packet enters the analyzer during the voltage ramp and form rings that induce current detected by the amplifier, 2012. https://commons.wikimedia.org/wiki/File:OrbitrapMA%26Injector.png (accessed October 9, 2020). Rights: Thermo Fisher Scientific (Bremen) / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0).

[59] A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat, S. Horning, Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer, Anal. Chem. 78 (2006) 2113–2120. https://doi.org/10.1021/ac0518811.

[60] T. Sakurai, H. Nakabushi, T. Hiasa, K. Okanishi, A new multi-passage time-of-flight mass spectrometer at JAIST, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 427 (1999) 182–186. https://doi.org/10.1016/S0168-9002(98)01565-4.

[61] K. Strupat, O. Scheibner, M. Bromirski, High-Resolution, Accurate-Mass Orbitrap Mass Spectrometry – Definitions, Opportunities, and Advantages (Technical Note 64287), (2016). https://assets.thermofisher.com/TFS-Assets/CMD/Technical-Notes/tn-64287-hram-orbitrap-ms-tn64287-en.pdf (accessed October 9, 2020).

[62] Ł. Rajski, M. del M. Gómez-Ramos, A.R. Fernández-Alba, Large pesticide multiresidue screening method by liquid chromatography-Orbitrap mass spectrometry in full scan mode applied to fruit and vegetables, Journal of Chromatography A. 1360 (2014) 119–127. https://doi.org/10.1016/j.chroma.2014.07.061.

[63] F. Ichou, Mise en place d'une méthode de calibration pour construire une base de données MS/MS et développement d'un outil pour l'identification en ESI-HR-MS/MS de composés organophosphorés, Thèse de Doctorat, Université Pierre et Marie Curie - Paris VI, 2013. https://www.theses.fr/2013PA066115 (accessed October 18, 2020).

[64] F. Ichou, D. Lesage, X. Machuron-Mandard, C. Junot, R.B. Cole, J.-C. Tabet, Collision cell pressure effect on CID spectra pattern using triple quadrupole instruments: a RRKM modeling: Pressure effect on CID spectra pattern using TQ, J. Mass Spectrom. 48 (2013) 179–186. https://doi.org/10.1002/jms.3143.

[65] S. Herrera-Lopez, M.D. Hernando, E. García-Calvo, A.R. Fernández-Alba, M.M. Ulaszewska, Simultaneous screening of targeted and non-targeted contaminants using an LC-QTOF-MS system and automated MS/MS library searching, Journal of Mass Spectrometry. 49 (2014) 878–893. https://doi.org/10.1002/jms.3428.

[66] Thermo Scientific Q Exactive Plus Orbitrap LC-MS/MS System, (2016). https://assets.thermofisher.com/TFS-Assets/CMD/Specification-Sheets/PS-63912-LC-MS-Q-Exactive-Plus-Orbitrap-PS63912-EN.pdf (accessed October 9, 2020).

[67] Quanfirmation Plus, (2016). https://assets.thermofisher.com/TFS-Assets/CMD/brochures/BR-63890-LC-MS-Q-Exactive-Plus-Orbitrap-BR63890-EN.pdf (accessed October 9, 2020).

[68] F. Giacomoni, G. Le Corguille, M. Monsoor, M. Landi, P. Pericard, M. Petera, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E.A. Thevenot, C. Caron, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics. 31 (2015) 1493–1495. https://doi.org/10.1093/bioinformatics/btu813.

[69] Y. Guitton, M. Tremblay-Franco, G. Le Corguillé, J.-F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E.A. Thévenot, Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics, The International Journal of Biochemistry & Cell Biology. 93 (2017) 89–101. https://doi.org/10.1016/j.biocel.2017.07.002.

[70] Galaxy Workflow4Metabolomics, Galaxy Workflow4Metabolomics. (n.d.). https://galaxy.workflow4metabolomics.org/ (accessed February 28, 2020).

[71] Galaxy Workflow4Metabolomics, Galaxy Workflow4Metabolomics. (n.d.). https://workflow4metabolomics.usegalaxy.fr/ (accessed October 14, 2020).

[72] The R Project for Statistical Computing, R Project. (n.d.). https://www.r-project.org/ (accessed October 14, 2020).

[73] R. Ihaka, R. Gentleman, A Free Software Project, R Project. (n.d.). https://cran.r-project.org/doc/html/interface98-paper/paper_2.html (accessed October 14, 2020).

[74] 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.

[75] J. Chong, D.S. Wishart, J. Xia, Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis, Current Protocols in Bioinformatics. 68 (2019) e86. https://doi.org/10.1002/cpbi.86.

[76] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Research. 46 (2018) W486–W494. https://doi.org/10.1093/nar/gky310.

[77] MetaboAnalyst, MetaboAnalyst. (n.d.). https://www.metaboanalyst.ca/ (accessed February 28, 2020).

[78] Y.C. Tai, T.P. Speed, A multivariate empirical Bayes statistic for replicated microarray time course data, Ann. Statist. 34 (2006) 2387–2412. https://doi.org/10.1214/009053606000000759.

[79] L. Wilkinson, M. Friendly, The History of the Cluster Heat Map, The American Statistician. 63 (2009) 179–184. https://doi.org/10.1198/tas.2009.0033.

[80] T. Loua, Atlas statistique de la population de Paris, J. Dejey & cie, Paris, France, 1873. https://books.google.fr/books?id=yVDkyoNPZEwC (accessed June 16, 2020).

[81] E. Withers, P.W. Hill, D.R. Chadwick, D.L. Jones, Use of untargeted metabolomics for assessing soil quality and microbial function, Soil Biology and Biochemistry. 143 (2020) 107758. https://doi.org/10.1016/j.soilbio.2020.107758.

[82] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2 (1901) 559–572. https://doi.org/10.1080/14786440109462720.

[83] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics. 16 (2002) 119–128. https://doi.org/10.1002/cem.695.

[84] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, NY, 1966: pp. 391–420. https://ci.nii.ac.jp/naid/20001378860/en/ (accessed October 18, 2020).

[85] J. Boccard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Analytica Chimica Acta. 769 (2013) 30–39. https://doi.org/10.1016/j.aca.2013.01.022.

[86] S. Wiklund, Multivariate data analysis for Omics, (2008). https://metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf (accessed October 18, 2020).

[87] J. Cohen, Things I have learned (so far)., American Psychologist. 45 (1990) 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304.

[88] A. Roux, Analysis of human urinary metabolome by liquid chromatography coupled to high resolution mass spectrometry, Thèse de Doctorat, Université Pierre et Marie Curie - Paris VI, 2011. https://tel.archives-ouvertes.fr/tel-00641529 (accessed October 18, 2020).

Chapter I

"Extractions and Chemometrics: a dialectic relation"

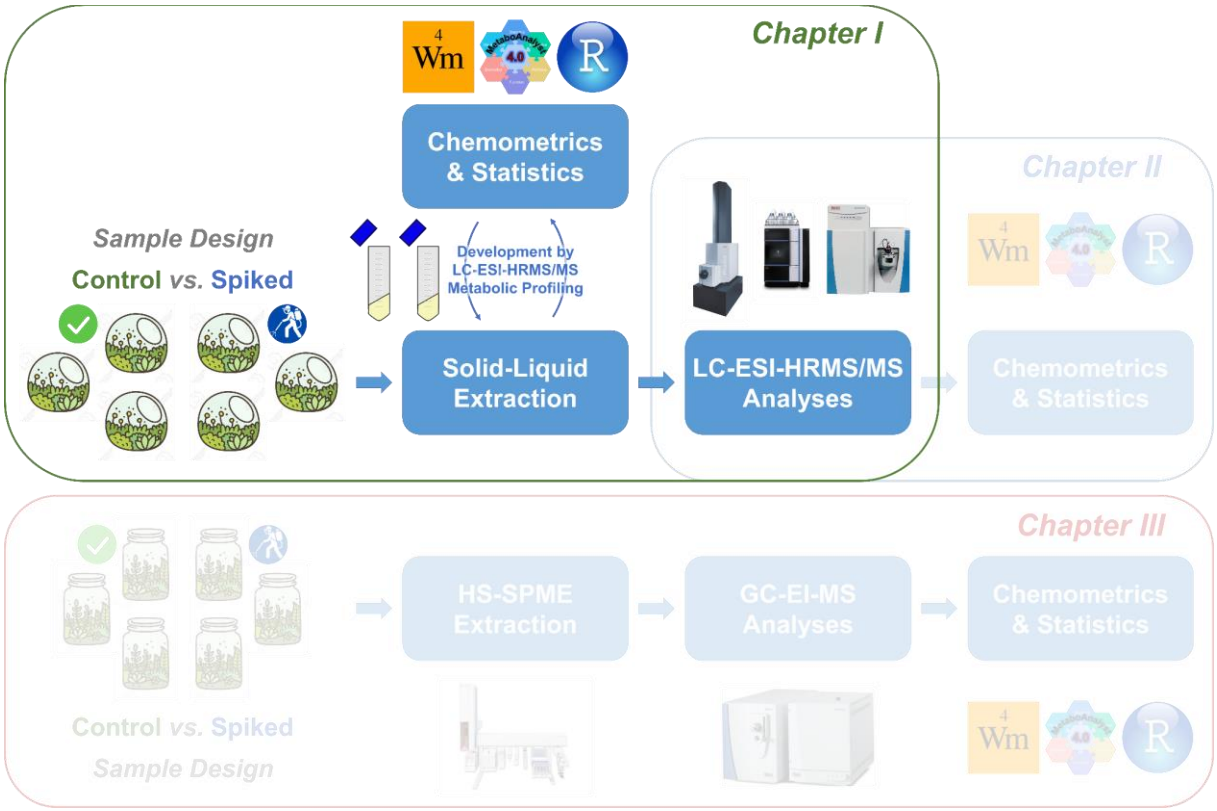
Preamble

In the present Chapter, the **set-up and development of the meta-metabolome extraction** will be addressed. In fact, the meta-metabolome extraction is the key-step and the bottleneck of the EMF approach, where several required criteria should be satisfied in order to collect sufficient and reliable metabolic information. First, this information will be essential for the determination of **pesticides' fate in the environmental matrix**, which can be done by detecting its active substance, its formulation ingredients, and their transformation by-products. On another hand, a "sufficient" metabolic information requires the extraction of the maximum number of endometabolites. In fact, there is no previous knowledge about the possible impact of the emerging pesticides, and consequently, the affected microbial metabolic pathways cannot be *a priori* expected. Thus, statistically, detecting a higher number of endometabolites helps increasing the possibility of detecting microbial biomarkers. These biomarkers are the indicators of **pesticide impact on matrix microbial community**. Hence, the wider is the collected meta-metabolic information, the deeper is the understanding of **pesticide's environmental fate and impact**. Among others, the optimization of an efficient meta-metabolome extraction is essential for a reliable determination of the "resilience time".

Nevertheless, to achieve the mentioned requirements, several challenging tasks should be addressed. Indeed, **broadening the band of the extractible metabolites is problematic from a physical-chemical point of view**. The single extraction of **diverse types and families of metabolites** including polar, semi-polar and non-polar molecules is highly problematic. In addition, **high reproducibility** is required for the extraction in order to allow for reliable comparative statistical analyses based on the relative quantification of molecular ions. Moreover, an efficient extraction may lead to **complex metabolic profiles**, which can engender several **analytical drifts** and **difficulties in data processing**. On the other hand, problematics regarding **data interpretation** can be identified when developing **an analytical method for untargeted metabolomics**. In fact, the definition of the "optimal" conditions is ambiguous and hard to be determined, as the generated metabolic datasets are complex and the experimental designs are large and multi-factorial.

In order to address these challenging problematics, **novel extraction protocols** will be set-up and applied. They will be compared to previously published reference protocols in order to assess their performance. Moreover, as these comparisons are complex, **a novel LC-HRMS-**

based untargeted metabolic profiling strategy will be set-up in order to help **assessing the "optimal" protocol**, by using **the computational chemometric tools and the advanced statistical models**. They will be used to investigate certain criteria that are defined according to the **analytical requirements** and the **environmental objectives**.



Liquid Chromatography-High Resolution Mass Spectrometry-based untargeted profiling as a tool for analytical development: Assessment of novel extraction protocols for herbicide-polluted soil meta-metabolomics

Hikmat Ghosson[1,2,*], Yannick Brunato[1,2], Delphine Raviglione[1,2],
Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,3,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: *Equal Contribution (Last Co-authors)*

Publication

In-preparation.

Keywords

Untargeted Metabolic Profiling; Solid-Liquid Extraction; Liquid Chromatography-High Resolution Mass Spectrometry; Soil Metabolomics; Herbicides

List of Abbreviations related to the experimental design

E1: Simple extraction based on EtOAc + HCl

E2: Simple extraction based on MeOH

E3: Simple extraction based on MeOH + FA

E4: Double extraction-based protocol (ACN/iPA; H2O/MeOH)

E5: Double extraction-based protocol (ACN/iPA + FA; H2O/MeOH + FA)

SP: Soil of Perpignan

ST: Soil of Torreilles

Bel: Soil polluted with formulated Nonanoic acid (Beloukha®) herbicide

Gly: Soil polluted with formulated Glyphosate Tartan Super 360™ herbicide

Ctr: Unpolluted control soil

BLKX: Blank Extraction

E1QC: pool Quality Control sample containing similar volume aliquots from all the 30 soil extracts issued from extraction protocol “E1”

E2QC: pool Quality Control sample containing similar volume aliquots from all the 30 soil extracts issued from extraction protocol “E2”

E3QC: pool Quality Control sample containing similar volume aliquots from all the 30 soil extracts issued from extraction protocol “E3”

E4QC: pool Quality Control sample containing similar volume aliquots from all the 30 soil extracts issued from extraction protocol “E4”

E5QC: pool Quality Control sample containing similar volume aliquots from all the 30 soil extracts issued from extraction protocol “E5”

QC: pool Quality Control sample containing similar volume aliquots from all the 150 soil extracts

LoadQC: pool Quality Control sample injected to load the LC-HRMS system

BLKInj: Blank Injection

1. Introduction

Herbicides use is essential for agricultural activities as crop protection and crop yield enhancement. Nevertheless, the use of these chemical products risks polluting the soil [1] and affecting the activity of its organisms [2]. This activity is basic for soil health and productivity [3,4]. Hence, assessing environmental fate and impact of herbicides is essential for studying their risks on soil quality. However, the assessment of both the fate and the impact of herbicides in such complex biological matrix faces various issues. First, both the classic synthetic herbicides and the emerging nature-originating herbicides [5] are complex chemical agents. They consist of diverse known and unknown (bio)chemical components, particularly when their formulated products are used. On the other hand, soil's endogenous molecules are key indicators of organisms' activities and mediation. Their detection, quantification and characterization is thus important to assess herbicides impact on soil's biosystems. The extraction and analyses of these molecules are however known to be challenging [4,6].

Analytical Chemistry and its advanced approaches provide a powerful tool to address this problematic. Meta-metabolomics, targeting small molecules (< 1000-1500 Da) originating from a whole community [7,8] can be a promising tool for assessing pollution impact on soil as suggested by Jones *et al.* [9]. Moreover, Patil *et al.* [10] and Salvia *et al.* [11] suggested a "universal" meta-metabolomics-based approach called Environmental Metabolic Footprinting (EMF) to study both the fate and the impact of pesticides in soils and sediments. The basis of the EMF approach is presented in Figure C.I 1. It targets both soil endometabolome and herbicide's xenometabolome (*i.e.* its active compounds, its transformation products, and its formulation agents). The larger is the analyzed meta-metabolome, the deeper is the understanding of herbicide's fate and impact. Thus, an optimal extraction allowing analyzing diverse types of metabolites is in fact the key point toward performant and robust EMF analyses. Developing a performant extraction protocol in this framework is however a challenging task that should deal with several requirements. First, a "broadband" extraction is needed in order to deal with the wide chemical diversity imposed by the complex nature of the studied samples. Indeed, soil contains a large biochemical diversity [3]. A large part of its molecules is still undiscovered so far. On the other hand, formulated herbicides consist of complex mixtures containing active substances as well as polymer- and tensioactif-based emulsifiers. In addition, herbicides' transformation by-products are diverse and their in-depth investigation is needed. In fact, degradation pathways are specific to the biotic and abiotic conditions of each eco-biosystem and thus can differ between the various environmental matrices. Furthermore, and

particularly for the emerging nature-originating herbicides, their activities mostly lay on multiple compounds acting in synergic and/or pleotropic modes of action [12]. Their complexities are thus higher and may contain unknown molecules [5,12]. In conclusion, such a large molecular diversity requires a "wide" extraction in term of polarity. Developing a "broadband" extraction covering a wide polarity scale and different types of metabolites/families of metabolites allows collecting a wide biochemical information. This increases the probability of detecting xenometabolites and also determining soil pollution biomarkers that might be general for several soil organisms (primary metabolites), or specific to particular organisms (secondary metabolites) [3]. On the other hand, the developed extraction protocol requires an optimal compromise between its wide molecular coverage and its yield, in order to increase the quality and the quantity of the collected metabolic information. Moreover, extraction reproducibility should be considered as an important requirement, as the untargeted meta-metabolomics lays on comparative experimental designs and relative quantitative analyses.

On the other hand, another problematic arises when developing extraction protocols and analytical methods for untargeted (meta-)metabolomics. The bases to define the "optimum" for such approaches are complex and relative to the studied context. They require reliable indicators and suitable sophisticated tools dedicated to reveal, explain and assess these indicators. For instance, the extraction yield is constrained to the need of the "broadband" coverage of meta-metabolome. Determining the "optimal" compromise between those two factors is hard to be assessed. The reproducibility cannot be determined by examining quantitative data of each of the detected compounds, as the approach is untargeted and deals with large datasets. Matrix effect assessment is not applicable as the absolute quantification with stable isotope-labelled standards is not reasonable at this stage of the study (*i.e.* method development).

The aim of the current work is first to introduce two versions of a novel solid-liquid extraction protocol dedicated for meta-metabolomics-based approaches and particularly the EMF. The protocol is based on a 2-steps extraction with two different miscible binary mixes of solvents or acidified solvents. It will be tested on two different types of herbicides applied on two different soils and will be compared to previously published extraction protocols [13,14]. The second main objective of the current work is to suggest a novel concept aiming to assess the optimal extraction protocol for untargeted (meta-)metabolomics, particularly for the EMF. The concept lays on a Liquid Chromatography-High Resolution Mass Spectrometry-based untargeted metabolic profiling, developed to examine the different extraction protocols applied

to soil samples. It will provide indicators and computational/statistical tools dedicated to assess the optimal extraction. Indicators will be based on 4 criteria: 1) The width of meta-metabolome coverage in term of molecular diversity and polarity, 2) the compromise between the width of extraction coverage and extraction yield, 3) the extraction reproducibility, and finally 4) the ability of the extraction to discriminate between polluted and unpolluted soil samples, *i.e.* the key objective of the EMF approach [10,11].

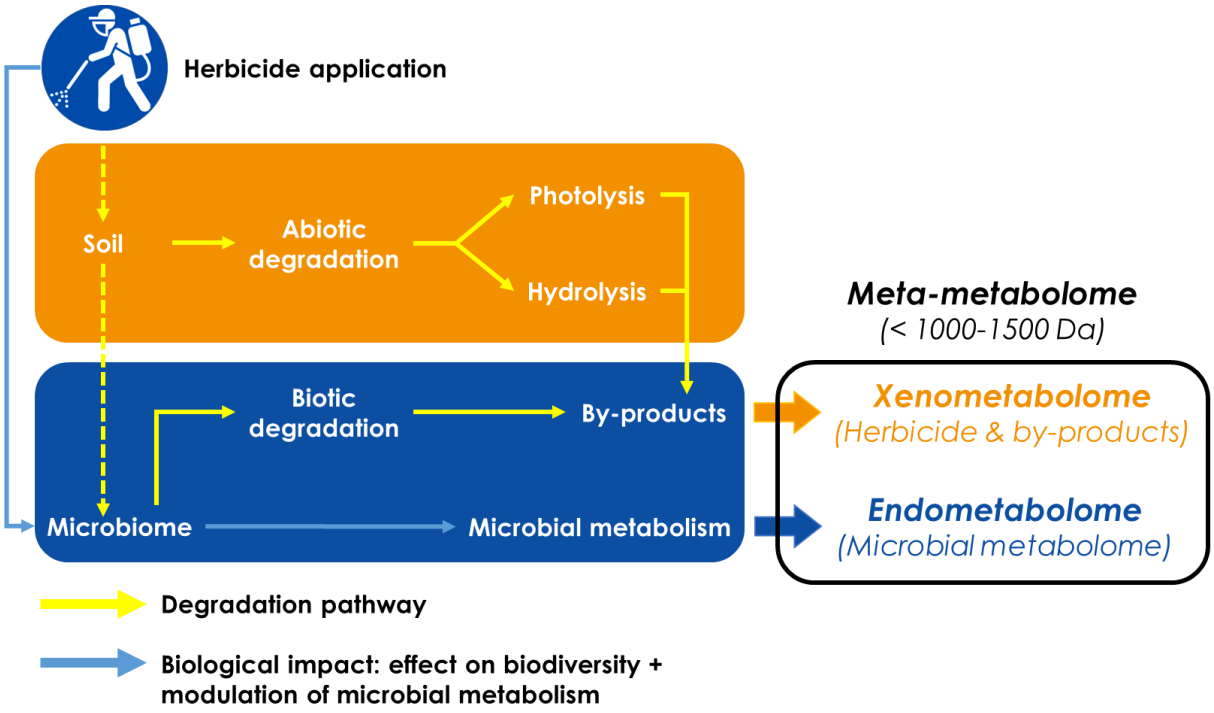


Figure C.I 1: The Environmental Metabolic Footprinting (EMF) concept.
Diagram adapted with modifications from Patil *et al.* [10], with permission from authors.

2. Material and Methods

2.1. Chemicals and materials

For the preparation of soil microcosms: Polystyrene Multiroir™ boxes were purchased from Fisher Scientific (Illkirch, France). For extractions: Water (H_2O) HPLC LC-MS grade, Acetonitrile (ACN) HPLC LC-MS grade, and 15 mL Soda glass test tubes (100 × 16.00 × 0.8-1.0 mm) were purchased from VWR International (Fontenay-sous-Bois, France). Methanol (MeOH) LC/MS and Formic acid (FA) LC-MS were purchased from CARLO ERBA (Val de Reuil, France). 2-Propanol (iPA) ULC/MS – CC/SFC was purchased form Biosolve Chimie SARL (Dieuze, France). Ethyl acetate (EtOAc) HPLC, ≥ 99.7 % was purchased from Honeywell (Seelze, Germany). Hydrochloric acid (HCl) 1.18 (~ 37 %) Analytical reagent grade was purchased from Fisher Scientific (Loughborough, U.K.). 50 mL fisherbrand centrifuge tubes with Polypropylene plug seal caps were purchased from Fisher Scientific (Illkirch, France). 0.22 µm Polytetrafluoroethylene (PTFE) filters and 2 mL vials were purchased *via* Analytic Lab (Castelnau-le-Lez, France). For LC-HRMS analyses: Water HPLC LC-MS grade was purchased from VWR International (Fontenay-sous-Bois, France). Acetonitrile LC/MS, 2-Propanol for LC/MS, and Formic Acid LC-MS were purchased from CARLO ERBA (Val de Reuil, France). Fluka® Sodium hydroxide solution (~1.0 M NaOH, for HPCE) was purchased from Honeywell (Seelze, Germany).

Tartan Super 360™ formulated solution (Barclay Chemicals Manufacturing Ltd., Mulhuddart, Ireland) containing 360 g L^{-1} of Glyphosate (synthetic herbicide), and Beloukha® formulated solution (JADE, Mérignac, France) containing 680 g L^{-1} of Nonanoic acid (natural-originating herbicide) were provided by the Sica CENTREX (Torreilles, France). Herbicides spiking solutions were prepared by 714.3-times and 312.5-times dilution of formulated Glyphosate and Nonanoic acid in H_2O (HPLC LC-MS grade), respectively.

Reference standards Glyphosate PESTANAL™ (analytical standard), Diquat dibromide monohydrate PESTANAL™ (analytical standard), Nonanoic acid 96 %, L-Methionine, L-Isoleucine, L-Leucine (all reagent grade, ≥ 98 % (HPLC)), Diclofenac sodium salt (analytical standard), and Chloramphenicol ≥ 98 % (HPLC) were purchased from Sigma-Aldrich (Saint-Quentin-Fallavier, France).

2.2. Soil material

Two different soils were selected for the study: “Soil of Perpignan” (SP), and “Soil of Torreilles” (ST). SP was collected from an arable field at the agricultural domain of the “Institut Universitaire de Technologie” (IUT) of Perpignan, France (42°40'53.7"N 2°53'54.8"E). ST was collected from an arable field at the agricultural domain of the “Centre Expérimental des Fruits et Légumes du Roussillon” (Sica CENTREX) of Torreilles, France (42°45'15.1"N 2°59'03.4"E). For both soils, the surface layer (15 cm) was collected on 3 different points separated by 1.5 meter. Soils composition analyses and characterization were performed by Galys/Arterris Laboratory (Toulouges, France) accredited by the French Accreditation Committee – Cofrac (Accreditation N° 1-6798). Results of soils analyses are detailed in Table A.I 1 (Appendix I). In brief, properties of SP are the following: 54.70 % of sand, 29.10 % of silt, 16.20 % of clay, 27.50 g Kg^{-1} of organic matter, 15.99 g Kg^{-1} of organic Carbon, 1.25 g Kg^{-1} of total Nitrogen, 99.00 meq Kg^{-1} of Cation Exchange Capacity (CEC), 2.60 mg Kg^{-1} of exchangeable Mn, 2.80 mg Kg^{-1} of exchangeable Cu, 260.30 % Ca/CEC, pH of 8.04 in water. According to the Soil Textural Triangle of the United States Department of Agriculture [15,16], this soil is classified as a Sandy Loam soil. For ST, properties are the following: 20.50 % of sand, 58.70 % of silt, 20.8 % of clay, 20.83 g Kg^{-1} of organic matter, 11.78 g Kg^{-1} of organic Carbon, 0.96 g Kg^{-1} of total Nitrogen, 91.60 meq Kg^{-1} of CEC, 15.22 mg Kg^{-1} of exchangeable Mn, 75.14 mg Kg^{-1} of exchangeable Cu, 67.20 % Ca/CEC, pH of 8.40 in water. This soil is classified as a Silt Loam soil according to the Soil Textural Triangle of the United States Department of Agriculture [15,16]. After collection, soils were homogenized and dried at ambient temperature (≈ 20 °C). After they were completely dried, they were manually grinded and then passed through a DIN-ISO 3310/1 2.00 mm sieve (Retsch, Haan, Germany). Then, they were humidified at 20 % of moisture, and stocked at 4 °C in the dark until the experiment.

2.3. Experimental design and the set-up of soil microcosms

The laboratory experimental design was based on 150 soil microcosms (75 of SP and 75 of ST). For each soil type, three different environmental conditions were applied: unpolluted control soils (Ctr), soils polluted with the Tartan Super 360™ formulated Glyphosate (Gly), and soils polluted with the Beloukha® formulated Nonanoic acid (Bel). Each environmental condition consisted of 50 microcosms (25 for SP and 25 for ST). Then, each environmental condition of a given soil type was divided into 5 batches, each consisting of 5 microcosms (5 environmental/biological replicates). These 5 batches were dedicated to test 5 different

extraction protocols: E1, E2, E3, E4 and E5 (detailed in the subsequent sections). The experimental design describing the different conditions and giving codes for the samples is summarized in Table A.I 2 (Appendix I).

Soil microcosms were prepared by weighing 40.00 g (± 0.50 g) of moist soil in Polystyrene Multiroir™ boxes (internal length: 55 mm, internal width: 40 mm, internal height: 43 mm). They were all incubated in a GC 401 growth chamber (Nüve, Saracalar, Turkey) for 10 days in order to re-establish the biological and microbial activity. Aerobic incubation conditions were 24 hours day/night cycle with alternation of light/dark, 28 °C/18 °C of temperature, and 40 % RH/65 % RH of humidity (Figure A.I 1 – Appendix I). The soil moisture was maintained at 20 % during the incubation and throughout the experiment, following a standardized environmental protocol implemented and published in previous works [10,17], aiming to assure conditions that are comparable to real environmental cases.

Polluted soil samples were spiked with herbicide spiking solutions (described in Section 2.1.) 4 hours before extractions. Spiking was done by applying 1 mL of spiking solutions on soil microcosms. This corresponds to one-time the agronomic field dose of each herbicide (12.60 µg of Glyphosate, and 54.40 µg of Nonanoic acid per gram of soil). 1 mL of H_2O for HPLC LC-MS grade (solvent of herbicide spiking solutions) was applied on unpolluted control soils. For all samples, the 1 mL was applied slowly on the upper layer of soil and then microcosms were smoothly shaken. This in order to assure an optimal dispersion of the applied solutions on the soil surface.

2.4. Extraction protocols and samples preparation

5 different extraction protocols: E1, E2, E3, E4 and E5 were applied for meta-metabolome analyses. For all extractions, 15.00 g (± 0.10 g) of soil material were taken from the microcosm after homogenization, and then transferred into a 50 mL fisherbrand centrifuge tube.

E1 corresponds to an adapted version of the protocol published by Romdhane *et al.* [13]. First, 3 mL of cold HCl 0.10 M solution (5 °C) are added in the 50 mL tube containing soil material, and then samples are shaken for 30 sec using Vortex hand shaker. Then, 15 mL of cold EtOAc (5 °C) are added and a 15 sec shaking is applied using the Vortex hand shaker. After, samples are swirled for 60 min using a KM-2 AKKU rotating shaker (Edmund Bühler GmbH, Bodelshausen, Germany), with a rotating speed of 420 RPM. Next, samples are centrifuged for 10 min at a rotation speed of 4700 RPM and a temperature of 10 °C using an Allegra X-30R Centrifuge (Beckman Coulter, Brea, CA, U.S.). After centrifugation, 10 mL of the supernatant

are transferred into a new 50 mL fisherbrand centrifuge tube. The remaining supernatant is eliminated and then new 15 mL of the cold EtOAc (5 °C) are introduced in the tube containing soil material. The same 15 sec Vortex shaking, 60 min swirling and 10 min centrifugation procedures are repeated. Then 10 mL of the new supernatant are taken and added to the first 10 mL taken at the first stage. A total volume of 20 mL of extract is thus collected. It is homogenized by shaking on Vortex hand shaker. Then, 10 mL of the recovered extract are transferred into a 15 mL Soda glass test tube.

E2 and E3 correspond to two adapted versions of the protocol published by Anastassiades *et al.* [14]. For E2: 20 mL of cold MeOH (5 °C) are added in the 50 mL tube containing soil material, and then samples are shaken for 30 sec using Vortex hand shaker. Next, samples are swirled for 10 min using BenchMixer™ Multi-Tube Vortexer (Benchmark Scientific, Sayreville, NJ, U.S.) at a rotation speed of 2500 RPM. After, samples are centrifuged for 10 min at a rotation speed of 4700 RPM and a temperature of 10 °C using the Allegra X-30R Centrifuge. After centrifugation, 10 mL of the supernatant are transferred into a 15 mL Soda glass test tube. For E3, the same procedure of extraction is applied by using cold MeOH + 1 % FA (v/v) (5 °C) instead of MeOH as extraction solvent.

E4 and E5 are the novel extraction protocols proposed by the present work. For E4: 10 mL of cold ACN/iPA 70:30 (v/v) (5 °C) are added in the 50 mL tube containing soil material. After, samples are shaken for 30 sec using Vortex hand shaker and then swirled for 5 min using BenchMixer™ Multi-Tube Vortexer at a rotation speed of 2500 RPM. Then, samples are centrifuged for 10 min at a rotation speed of 4700 RPM and a temperature of 10 °C using the Allegra X-30R Centrifuge. After centrifugation, 7 mL of the supernatant are transferred into a new 50 mL fisherbrand centrifuge tube. Then, 10 mL of cold H_2O/MeOH 90:10 (v/v) (5 °C) are introduced in the tube containing the soil material. The tube is then shaken, swirled and centrifuged under the same conditions as for the first step. After centrifugation, 7 mL of the supernatant are taken and added to the first 7 mL taken at the first stage. A total volume of 14 mL of extract is thus collected. It is homogenized by shaking on Vortex hand shaker and then 10 mL of the recovered extract are transferred into a 15 mL Soda glass test tube. For E5, the same procedure of extraction is applied by using cold (5 °C) acidified solvents (ACN/iPA 70:30 (v/v) + 1 % FA (v/v) and H_2O/MeOH 90:10 (v/v) + 1 % FA (v/v)).

For all extraction protocols, the 10 mL recovered extracts are evaporated under vacuum at 30 °C until dry, using an EZ-2^{plus} evaporator (Genevac, Ipswich, U.K.). The dry residue is then re-

dissolved in 1 mL of cold solvents (5 °C): MeOH for E1 and E2, MeOH + 1 % FA (v/v) for E3, H_2O/ACN/iPA/MeOH 45:35:15:5 (v/v/v/v) for E4, and H_2O/ACN/iPA/MeOH 45:35:15:5 (v/v/v/v) + 1 % FA (v/v) for E5. Then samples are spiked with Chloramphenicol and Diclofenac as Internal Standards (IS) (final concentration of 2 µg mL^{-1} for each). After, samples are dissolved and homogenized by mixing for 30 sec using the Vortex hand shaker, and then transferred into 2 mL vials after filtration through 0.22 µm PTFE filters.

For each extraction protocol, 3 replicates of blank extraction (BLKX) were performed by applying the same extraction procedures on empty 50 mL fisherbrand centrifuge tubes. This is in order to apply blank subtraction and to eliminate extraction contaminations from the dataset when data are processed. On the other hand, two types of pool QC samples were prepared for LC-HRMS and MS/MS analyses. First, "E"-pools (E1QC, E2QC, E3QC, E4QC and E5QC) were prepared by mixing similar volume aliquots (30 µL) from all soil extracts issued form the same extraction protocol, *i.e.* the 30 samples issued from an extraction protocol "E", consisting of the three environmental conditions (Ctr, Gly and Bel), applied on the two soil types (SP and ST) (Table A.I 2 – Appendix I). Then, 200 µL were taken from each of the "E"-pools and mixed in order to prepare the "pool QC" containing similar volume aliquots from all the 150 soil extracts.

2.5. LC-HRMS and MS/MS methods

LC-HRMS method developments and analyses were performed using a Vanquish UHPLC$^+$ Focused LC system equipped with an online degasser, a binary pump system, a temperature-controlled autosampler, and a column compartment (Thermo Fisher Scientific, Waltham, MA, U.S.), coupled with a maXis Electrospray-Quadrupole/Time-of-Flight (ESI-Q/ToF) mass spectrometer (Bruker Daltonics, Bremen, Germany; Billerica, MA, U.S.).

2.5.1. LC conditions

UHPLC column Luna® Omega Polar C18 (particle size: 1.6 µm, pore size: 100 Å, length: 100 mm, internal diameter: 2.1 mm, solid support: fully porous Silica) from Phenomenex (Torrance, CA, U.S.) was selected to achieve compounds separation. It was chosen in order to widen the band of the analyzable metabolites (in term of polarity), as the modified C18 columns has proven their performance in enhancing the retention efficiency for polar and semi-polar metabolites [18,19]. The column was equipped with a Phenomenex SecurityGuard™ ULTRA UHPLC Polar C18 2.1 mm column guard cartridges for protection during analyses. The mobile phase consisted of two phases of elution solvents: phase A is H_2O + 0.1 % FA (v/v), and phase

B is ACN + 0.1 % FA (v/v). Chromatographic separation was conducted by applying the following elution gradient: 0 % (B) (100 % aqueous) during 2.50 min, from 0 % to 100 % (B) in 12.50 min, 100 % (B) during 2.00 min, from 100 % to 0 % (B) in 1.00 min, and 0 % (B) during 2.00 min (a total of 20.00 min per run). The flow rate was set to 450 µL min^{-1}, column oven temperature was maintained at 30 °C. The injection volume was 5 µL. Only the elution from 0.20 min to 18.00 min was introduced into the mass spectrometer. The rest was diverted to waste in order to reduce ESI source and MS system fouling.

2.5.2. HRMS conditions

For the ESI-Q/ToF conditions, three main MS segments were established: two Full HRMS acquisition segments, and a calibration segment. The first Full HRMS acquisition segment was set between 0.20-15.00 min of Retention Time (RT), and the second was between 15.00-18.00 min of RT. The calibration segment was between 0.00-0.20 min of RT, which corresponds to LC void time (dead time). For the first Full HRMS acquisition segment: scan range was set between 90 *m/z* and 1000 *m/z* with 1.20 Hz of spectra rate (0.83 sec for 1 scan), corresponding to 7638 spectra summation (no rolling average was applied). Spectra acquisition was in both Profile (Continuum)[17] and Line (Centroid)[18]. Maximum Intensity was used for Line Spectra Calculation. "Focus" mode was active. For the ESI Source, a positive ion polarity mode was applied with a capillary voltage of 3500 V and an end plate offset of -500 V. Nitrogen nebulizer pressure was set to 2.4 Bar, the dry gas flow to 10.0 L min^{-1}, and the dry heater temperature to 200 °C. Tune parameters for Funnels, Quadrupole and Collision Cell were optimized in order to favor the transfer of ions with *m/z* between 90 and 1000 to the ToF analyzer. All these parameters are detailed in Section A.I.2.1. (Appendix I). The second Full HRMS acquisition segment was particularly dedicated to analyze the LC elution range with 100 % organic phase (B). All parameters are the same as for the first acquisition segment, except for the dry gas flow that was set to 8.0 L min^{-1} in order to enhance the detection of compounds eluted at this stage of the LC gradient. In fact, when the mobile phase is 100 % ACN, the volatility of the nebulized solvent is higher and thus a lower dry gas flow allows better compounds detection. For the

[17] Profile (Continuum) is the basic mode for *m/z* signal acquisition. It registers the distributed signal across *m/z* values (*i.e.* the spectrum) continuously when scanning the *m/z* range. Thus, the *m/z* profile will be drawn as a continuous function and the acquired *m/z* distribution of signal (*m/z* peaks) will take a Gaussian shape.

[18] Line (Centroid) mode transforms the *m/z* profile acquired in Continuum to a discrete function. In fact, the centroidization is performed by a mathematical transformation that regresses the Gaussian *m/z* peaks into *m/z* sticks (lines). Following the applied transformation algorithm, stick *m/z* value can be the "center of mass" of the Gaussian *m/z* peak (calculated through the weighted arithmetic mean), or the *m/z* value of the local maximum (the apex) of the Gaussian peak (as for the current case). The centroidization process compresses and reduces the data size. However, it engenders a loss of information [20].

calibration segment, a pre-run internal mass-scale calibration was performed in High Precision Calibration (HPC) mode using Sodium Formate (NaF) calibration solution (0.05 % FA + 0.50 mL NaOH 1.0 M in 50 mL of H_2O/iPA 50:50 (v/v)) that was automatically infused at the beginning of each injection. All parameters were the same as for the first Full HRMS acquisition segment, except for the nebulizer pressure that was set to 0.7 Bar (the suitable pressure for the syringe infusion flow rate equal to 3 µL min^{-1}), and for the dry gas flow that was adjusted to 4.0 L min^{-1} in order to reduce the de-clustering. For the ToF analyzer, voltage settings are described in details in Section A.I.2.2. (Appendix I). The resolving power was experimentally assessed at the Full Width at Half Maximum (FWHM) of *m/z* peaks along the scan range. All values are shown in Table A.I 3 (Appendix I).

2.5.3. MS/MS acquisitions

Two different Data Dependent Acquisition-based (DDA) MS/MS methods were set for compounds fragmentation by Collision-Induced Dissociation (CID) in Q/ToF's Nitrogen-supplied Hexapole collision cell. The three segments (two for acquisition and one for calibration) previously described for the Full HRMS method were applied for both MS/MS methods. The same ESI Source, Funnels, Quadrupole and ion transfer parameters were set. Acquisition settings were however adapted for the fragmentation. The scan range was expanded to detect small fragments. It was set between *m/z* 50 and 1000 with a spectra rate equal to 2.00 Hz (0.50 sec for 1 scan, corresponding to 4545 spectra summation). The spectra rate was increased in order to acquire a higher number of MS/MS spectra. Spectra acquisition was in both Profile and Line. Maximum Intensity was used for Line Spectra Calculation. "Focus" mode was active. Auto MS/MS mode (DDA) with an absolute intensity threshold equal to 3500 cts was applied for the two acquisition segments.

For method N°1: the selection was based on the number of precursors. It was set to 3, *i.e.*, a first Full HRMS scan selects the most 3 intense ions (precursors), and then 3 different MS/MS spectra are generated, each corresponding to one of the three selected precursors (after they were isolated in the Quadrupole and fragmented in the Hexapole). The active exclusion was set to 5 spectra (*i.e.* when a precursor is fragmented for 5 times, it will be excluded to allow the selection and the fragmentation of a new different precursor).

For method N°2: the number of precursors selected for an MS/MS scan was set to 1. The active exclusion was set to 4 spectra.

For both methods, the exclusion will be released after 1 min. The precursor will be reconsidered for fragmentation if its intensity increases by a factor of 2 in the next scan. An additional "Smart

Exclusion" was also activated (set to 5 ×) in order to enhance background exclusion. For CID, an increasing collision energy ramp following the increase of *m/z* is set. It is summarized in Table A.I 4 (Appendix I).

2.6. Analytical sequence for LC-HRMS/MS-based untargeted profiling

The untargeted metabolic profiling consisted of two LC-HRMS/MS analytical injection batches. For both, sequences were initiated with two MeOH blank injections (BLKInj), followed by injecting all blank extractions. Then, 4 "pool QC" were injected in order to load the LC-ESI-Q/ToF system (named "LoadQC"), followed by 75 soil extract samples randomly selected and randomly injected in order to minimize the effect of instrumental drifts. A "pool QC" was injected every 5 sample injections in order to control and correct potential "inter-batch" and "intra-batch" drifts. After injecting the 75 samples, the 5 "E"-pool samples were injected for MS/MS acquisitions (MS/MS method N°1 applied in the first sequence and MS/MS method N°2 applied in the second sequence). At the end of each sequence, a mix of reference standards containing Glyphosate (20 µg mL^{-1}), Diquat dibromide monohydrate (20 µg mL^{-1}), Nonanoic acid (10 µg mL^{-1}), L-Methionine (10 µg mL^{-1}), L-Isoleucine (10 µg mL^{-1}), L-Leucine (10 µg mL^{-1}), Diclofenac sodium salt (2 µg mL^{-1}) and Chloramphenicol (2 µg mL^{-1}) dissolved in MeOH, and another identical standards mix dissolved in H_2O/ACN/iPA/MeOH 45:35:15:5 (v/v/v/v) were injected, followed by a final H_2O/ACN/iPA/MeOH 45:35:15:5 (v/v/v/v) blank injection. Samples were maintained at 10 °C in the LC temperature-controlled autosampler along the analytical sequence. For further details, sequences documents are published on the European Bioinformatics Institute (Hinxton, U.K.) MetaboLights platform [21,22].

2.7. Data processing and software

LC system, LC-ESI-Q/ToF hyphenation and analytical sequence piloting were performed using HyStar 3.2.49.4 (Bruker Daltonics). ESI-Q/ToF piloting, LC-HRMS and MS/MS data acquisition were performed using otofControl 4.0.97.4560 (Bruker Daltonics). Raw LC-HRMS data were acquired in ".d" folder format. A new post-acquisition mass-scale internal calibration (HPC mode) was performed using Compass DataAnalysis 4.3 (Bruker Daltonics), and then ".d" files were converted to NetCDF using the same software, in order to process them using Galaxy Workflow4Metabolomics platform [23–25]. All the NetCDF files are published on the MetaboLights platform (European Bioinformatics Institute) [21]. The pre-processing workflow and all its detailed parameters are published on the Galaxy Workflow4Metabolomics platform [26]. In brief, the following XCMS-based [27] data pre-processing pipeline was applied: a

"centWave" peak piking [28] is performed (ROI considered when detecting 5 consecutive scans with minimum intensity equal to 2000 and maximum *m/z* deviation of 5 ppm – S/N cutoff: 10), followed by a first "PeakDensity" peak grouping (bandwidth: 15 sec). Then, a loess/non-linear "PeakGroups" retention time adjustment is applied (degree of smoothing: 0.8), followed by a second "PeakDensity" peak grouping (same parameters as the first), a peak filling step, and a CAMERA-based peak annotation [29] (correlation threshold: 0.75). The considered signal value for ion features was the chromatographic peak area. After, the blank subtraction was applied by eliminating all features detected more than one time in blank extraction injections. Next, an "intra-batch" signal correction was applied using the "batch correction" function with a "loess" regression model [30] (0.8 of span), followed by a matrix clean-up according to feature's CV in pool QC injections (all features with area RSD upper than 30 % through pool QC injections were eliminated from the dataset) [31]. After generating the data matrix, statistical analyses were performed using the R-based MetaboAnalyst platform[19] [32–34]. Euclidean Distances and Euclidean Distances SDs were calculated using R 3.3.3 software. The command lines are shown in Section A.I.4. (Appendix I). Compass DataAnalysis 4.3 was used for counting molecular features directly from raw data using "Find Molecular Features" (FMF) algorithm. The following FMF detection parameters were applied: an S/N threshold equal to 10, a correlation coefficient threshold equal to 0.75, a minimum compound length equal to 5 spectra, a smoothing width equal to 10 followed by an additional smoothing, and a mass spectrum calculation applied to profile spectra only, after subtracting constant MS background. FMF detection parameters were selected to be concordant with the parameters of "centWave" peak piking and CAMERA-based annotations. Manual raw data processing and MS/MS data exploring were also performed using Compass DataAnalysis 4.3.

[19] Command lines cannot be retrieved due to a version update on platform's website.

3. Results and Discussions

The LC-HRMS/MS-based untargeted metabolic profiling generated 238 LC-HRMS(/MS) data files. Data of the 30 blank extractions, the 150 samples and the 30 pool QC that were injected between samples were pre-processed by the automated pipeline. They will be considered for the statistical analyses and the manual raw data processing. The 10 "E"-pool MS/MS data and the 4 external standards injections data will be considered for manual raw data processing. Data of the 6 blank injections and the 8 pool QC that were injected at the beginning of the sequences (LoadQC) will be however excluded from the rest of the study.

Automated data pre-processing provided an "original" data matrix consisting of 411 variables (molecular traces/features) and 180 observations (samples/injections) divided into 31 factors (the different extraction-soil-environmental groups + the QC). A "light" data matrix was generated after the elimination of ion redundancies (*e.g.*, ion adducts, ion clusters, isotopes). The reduction was done by only keeping the most intense feature of a given "compound group" (*i.e.* a group of features that are considered originating from the same compound according to CAMERA grouping [29]). Features filtration led to consider 234 molecular features. The two generated data matrices (original and light) will be used for the statistical analyses described in the following parts.

3.1. Observational investigations of LC-HRMS raw data

LC-HRMS raw data are the foundation of the generated datasets. Their investigation provides general understanding for the analytical information and its quality. Thus, LC Chromatograms and MS data were investigated by observing Base Peaks Chromatograms (BPC), and by counting molecular features (described in Section 2.7.). BPCs are presented in Figure A.I 2-Figure A.I 7 (Appendix I). Each BPC is originating from a sample injection belonging to an extraction-soil-environmental group. The blank extraction BPC is also presented in the background (grey chromatogram) to show if peaks correspond to soil meta-metabolome or contaminations issued from the extraction.

Several remarks are revealed by the observational investigation of BPCs. First, BPCs belonging to E1 protocol (purple chromatograms) show relatively poor meta-metabolic profiles through all soil-environmental groups. Major peaks observed are also present in the blank extraction. The poor profile can be hypothetically explained by the fact that when the non-polar EtOAc solvent was applied on soil, soil particles (relatively polar material containing salts, metal complexes, ions and water) were repelled and aggregated, which led to reduce the contact

interface between the solvent and soil's material. The phenomenon is shown in Video S 1 (accessible online *via* the reference [35]).

Otherwise, the comparison between BPCs belonging to ST and those belonging to SP shows that ST's metabolic profiles are significantly poorer in term of the detected endometabolites (Figure A.I 2 *vs*. Figure A.I 3). This observation can be explained by the results of soil analysis (Section 2.2., and Section A.I.1. of Appendix I). In fact, Organic Matter and Organic Carbon amounts are lower in ST (Table A.I 1 – Appendix I). Moreover, exchangeable Copper is significantly higher for ST (75.14 mg Kg^{-1}, *vs.* 2.80 mg Kg^{-1} for SP). The Cu is known for its anti-microbial activity [36]. This fact can be an additional explanation for the poor endometabolic profile of ST, as the microbial activity is one of the most important sources of endometabolites in soil.

On the other hand, BPCs of soils polluted with formulated Nonanoic acid herbicide show that the product presents a heavy/complex xenometabolic profile (orange and blue semi-boxes) when compared to the formulated Glyphosate herbicide profile (purple semi-boxes). BPCs of E2, E3, E4 and E5 (Figure A.I 6 and Figure A.I 7) show a massive profile of peaks eluting between 13.50 min and 16.00 min of RT (blue semi-boxes). BPCs of E4 and E5 reveal additional massive profiles eluting between 5.00 min and 7.50 min of RT (orange semi-boxes). Such massive peaks profiles mostly consist of polymers or emulsifiers originating from the herbicide (as they are detected in Bel samples only). In addition, an enhanced detection of the massive profiles of xenometabolites (particularly the part eluting between 14.50 min and 16.00 min of RT) can be noticed in ST when compared to SP. This phenomenon can be explained by the matrix effect and the ion suppression [37] that must be higher in SP, as its metabolic profile is more complex. Thus, xenometabolome detection may be dependent of the soil type and its properties.

Besides, MS data were explored by molecular features (MF) counting (using the FMF algorithm described in Section 2.7.). Results are shown in Table C.I 1.

Table C.I 1: Numbers of molecular features detected in each group of samples, through the different extraction protocols.
Data shown below represent the means of the numbers of molecular features calculated through the 5 replicates of each group. The means of molecular features counted in blank extractions are subtracted from the results shown in the table.

Group	E1	E2	E3	E4	E5
SPCtr	38	1093	1274	1349	1181
STCtr	19	846	1045	1276	1195
SPGly	174	1151	1461	1799	1722
STGly	178	995	1323	1728	1624
SPBel	624	1669	1965	2886	2868
STBel	714	1961	2027	2991	3187

MF-counting results are concordant with observations noticed by BPCs exploring. For instance, numbers of MF counted in E1 through sample groups are the lowest. In addition, MF for SPCtr are higher than for STCtr, which confirms that SP is richer in terms of endometabolites. Regarding the pollution conditions, Bel groups show the highest MF. The effect of soil type on xenometabolome detection can also be noticed through MF-counting. In fact, if MF numbers of Ctr (that corresponds to the endometabolome) are subtracted from MF belonging to Bel, higher numbers of MF (corresponding to the xenometabolome) can be observed in ST (Table A.I 5 – Appendix I), particularly for E2 and E3. For Gly, the same phenomenon is significantly observed only in E2 and E3.

On the other hand, endometabolome extraction efficiency was investigated by exploring SPCtr and STCtr data. First, regarding SP, the extraction showing the highest number of molecular traces is E4, followed by E3, E5, E2 and then E1. For ST, E4 showed the higher number of molecular traces, followed by E5, E3, E2 and then E1.

Nevertheless, both observational results of BPCs and MF are limited to general conclusions that risk to be biased by several analytical and data interpretation issues. For instance, BPCs are not suitable to reveal minor compounds, as they only show the chromatogram belonging to a "Base Peak", *i.e.* the most intense peak at a given MS scan (or RT). This principle risks hiding chromatographic peaks belonging to ions co-eluting with the Base Peak but presenting a lower intensity. In addition, observational exploring of chromatograms is unable to reveal detailed information in such complex datasets and multi-factorial experimental designs. For MF-counting approach, as the calculation of the detected MF numbers does include ion redundancies, the method risks biasing the real numbers of the detected molecules, and thus the conclusions about extraction performances. In addition, MF-counting approach suffers from a

higher vulnerability to noise and artefacts integration when compared to the automated data pre-processing approach. In fact, artefacts elimination using "features' CVs in QC"-based technique is not applicable for the MF-counting approach. This matrix clean-up step, performed after the blank subtraction, led to the elimination of 83.94 % of features detected by the XCMS-based automated pre-processing. It partially explains the difference between the numbers of the detected features when comparing the automated pre-processing approach to the MF-counting approach. The difference can also be the result of the application of two different peak piking algorithms [38]. Furthermore, extractions reproducibility and the ability to discriminate between polluted and unpolluted soils cannot be assessed through those two approaches. Therefore, in-depth data investigations require sophisticated computational and statistical analyses with higher performance and reliability. Untargeted metabolic profiling will thus be exploited in order to assess the optimal extraction protocol.

3.2. Assessment of molecular diversity coverage and extraction yield

The first criterion defined to assess extraction performance is the width of meta-metabolome coverage (in term of molecular diversity/polarity). Another related criterion that should be considered is the compromise between the width of extraction coverage and the extraction yield. To investigate those criteria, complex metabolic profiles should be decomposed and visualized as the following: 1) traces of detected metabolites should be defined by RT and *m/z*, and restrained to 1 feature per metabolite, 2) features intensities (related to their concentration) should be compared according to the different samples and factors, *i.e.* extraction-soil-environmental groups.

To address this problematic, a Heatmap analysis [39] was performed. The Heatmap is a three-dimension data analysis that visualizes the intensities of features in the different samples. Thus, these samples can be compared according to the different conditions (*i.e.* extraction protocols, soil types and environmental conditions). In addition, hierarchical clustering performed using an associated Dendrogram is an available option that can be exploited in the Heatmap. It allows for the aggregation of features/compounds that are correlated according to their intensity profiles through the different samples, and/or the aggregation of samples correlated according to the similarity of their metabolic profiles.

To perform the Heatmap analysis, QC samples were excluded from the "light" data matrix that was selected in order to exclude the redundancies of metabolites (234 features). A generalized logarithm transformation (glog) was performed on the dataset in order to reduce the "size

effect" and thus avoid the hiding of low-intensity compounds. Features intensity standardization was not applied in order to visualize the null intensities that correspond to compounds that are not detected in a given sub-batch. Hierarchical clustering was applied to features in order to discriminate between xenometabolites of the different herbicides, and endometabolites of the different soils. The applied clustering algorithm was "Ward", and the distance measure was "Euclidean". Results are shown in Figure C.I 2.

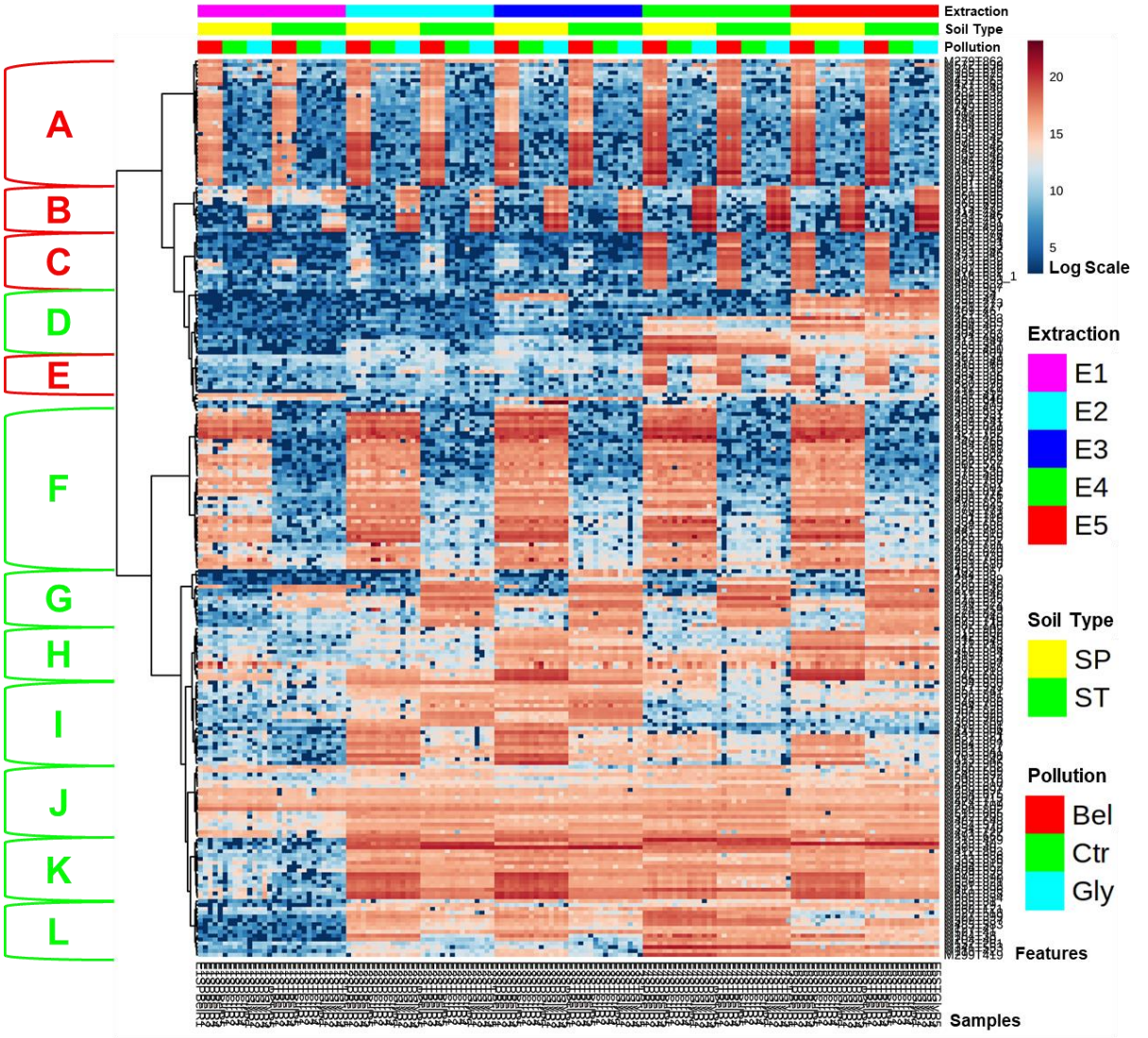


Figure C.I 2: Dendrogram-Heatmap analysis.
The numbers of features per cluster are as the following: A: 33; B: 12; C: 15; D: 17; E: 10; F: 43; G: 15; H: 14; I: 22; J: 19; K: 17; L: 14. Clusters in red are corresponding to xenometabolites, and those in green are for endometabolites.
3 features between E and F were excluded from the clustering and determined as the following: a compound only detected by the E1 protocol, a compound presenting the highest abundance in the E3 protocol, and a compound only detected (with a significantly-high abundance) in the E3SPGly sub-batch.
Plot generated using MetaboAnalyst.

Dataset representation by the Heatmap and the Dendrogram clustering of features led to identify 12 clusters of metabolites (A to L – Figure C.I 2).

First, concerning xenometabolites; the cluster A consists of 33 metabolites only detected in Bel groups, they correspond to xenometabolites originating from the formulated Nonanoic acid herbicide. The intensity scale shows that the highest abundances of the majority of these xenometabolites is in the extracts of E4 and E5, followed by the extracts of E3 and E2. The lowest intensities for these compounds are in the E1 extracts. The cluster C (15 metabolites) is also showing xenometabolites of the Nonanoic acid herbicide. These xenometabolites are however significantly intense only in E4 and E5 extracts (slightly more intense in those of E4). In addition, The E cluster shows 10 compounds also detected in E4 and E5 extracts and significantly present in the Bel groups. They are also detected in the Gly groups but with lower abundances. One hypothesis is that these compounds could be identified as common formulation agents that are present in both the Nonanoic acid and the Glyphosate formulated herbicides. Their highest intensities are in the E4 extracts, and particularly in the SP soil.
The B cluster consists of 12 xenometabolites originating from the formulated Glyphosate herbicide, as they are only detected in the Gly groups. The highest abundances of these xenometabolites is in the E4 extracts, followed by the E5, E3, E2, and finally the E1 extracts.

Concerning endometabolome, the cluster F containing the highest number of metabolites (43) shows SP-specific endometabolites. They are detected in all extraction protocols, with slightly higher abundances in E4 and E5 extracts, and significantly lower abundances in the E1 extracts.

On the other hand, the cluster G with 15 metabolites shows ST-specific endometabolites. They are detected in E5 (highest abundances), E4, E3 and E2 extracts. Only 6 of those metabolites are detected in the E1 extracts with relatively low intensities.

Finally, clusters D, H, I, J, K, and L show the endometabolites that are common for both the SP and the ST soils (86 metabolites in total). The D cluster (17 metabolites) mainly consists of endometabolites with the highest abundances in the extracts of the 2-steps binary solvents-based protocols (E4 and E5). It can be divided into two sub-clusters differentiating between compounds with higher abundances in one of the two extractions (12 for E5 *vs.* 5 for E4). The H cluster (14 metabolites) mainly consists of endometabolites with the highest abundances in the acidified extracts (E3 and E5). The abundances are comparable between the two extracts, with slightly higher intensities in E5. The I cluster (22 metabolites) mainly shows metabolites with highest abundances in the MeOH-based protocols extracts (E2 and E3). The intensities of these metabolites are higher in the E3 extracts. 8 of these metabolites can be significantly detected in the E5 extract but with lower intensities if compared to the intensities in E3. The J

cluster show 19 metabolites that are detected in all extracts with comparable intensities. The K cluster show 17 metabolites that are also detected in all extracts with comparable intensities, except for the E1 extracts, where they show almost a null intensity. 14 metabolites are shown in the L cluster that groups the metabolites that are detected in the E2, E3, E4 and E5 extracts, but showing significantly higher abundances in E4.

In conclusion, the results presented above clearly show that E4 and E5 extraction protocols are the most performant protocols for xenometabolome extraction and/or detection (as the intensities can also be influenced by the injection solvent that affects the elution and the ionization). This superiority for xenometabolome extraction/detection can be hypothetically explained by the fact that the presence of H_2O as a part of the extraction solvent mix (≈ 45 %) can enhance the extraction of formulation agents, particularly surfactants and emulsifiers. In fact, these compounds are known to be soluble in water. In addition, their role is to enhance the solubility of herbicide's active compound in water in order to facilitate its in-field application. However, regarding endometabolome extraction efficiency, further investigations are needed in order to draw more clear conclusions. Other analyses were thus performed in order to study the width of meta-metabolome coverage in term of polarity. Therefore, a different Dendrogram-Heatmap-based data visualization was performed after splitting the dataset into three different sub-datasets. The splitting was performed according to RT, which is, in theory, directly related to the polarity scale (LogP). The three sub-datasets were generated after splitting the light dataset by RT ranges: the first "polar" sub-dataset contains 19 compounds eluted between 0.30 min and 4.20 min, where 4.20 min corresponds to the RT of the putatively MS/MS-identified 2'-Deoxyadenosine (LogP: -0.50). The second "semi-polar" sub-dataset contains 93 compounds eluted between 4.20 min and 10.75 min (RT of the IS Diclofenac – LogP: +4.40). The third "non-polar" sub-dataset consists of 122 compounds eluted between 10.75 and 18 min. The results of the polarity-scale Heatmap analysis are shown in Figure C.I 3.

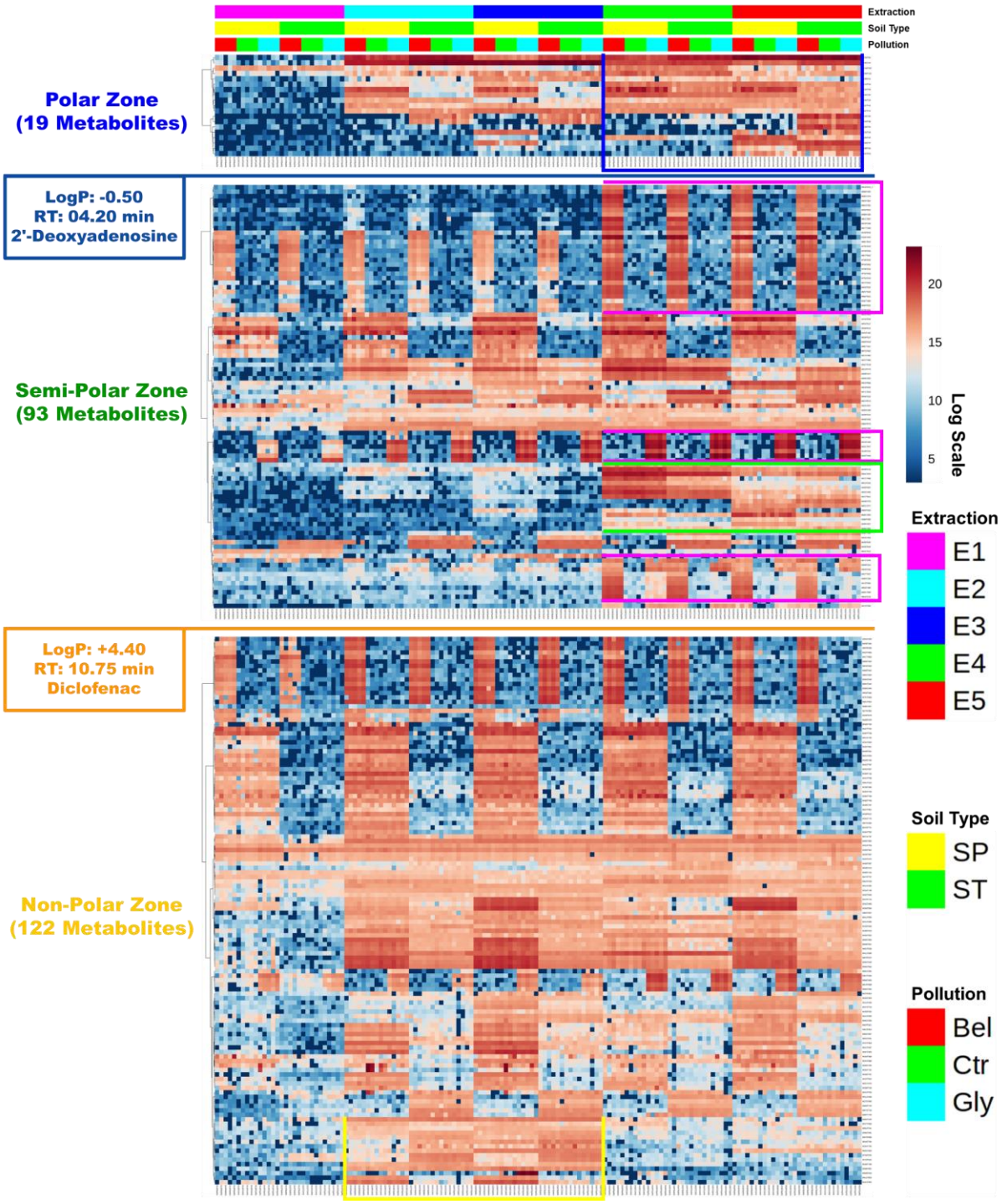


Figure C.I 3: Polarity-scale Dendrogram-Heatmap analyses.
Plots generated using MetaboAnalyst.

For polar metabolites, the protocol E5, followed by the E4, shows the best performance when compared to other protocols (the blue semi-box). This can be explained by the presence of H_2O as a part of the extraction solvent mix (~ 45 %). For semi-polar metabolites, the E4, followed by the E5, shows the best performance for the detection of xenometabolites eluting at this range (purple semi-boxes) and for the endometabolome (the 15 compounds in the green semi-box).

For non-polar metabolites, the performances of E5 and E3 seem to be comparable. The only superiority noticed for E3 (and then for E2) is the ability to detect the 15 metabolites highlighted in the yellow semi-box. This superiority can be hypothetically explained by the fact that despite the presence of ≈ 35 % of ACN and 15 % of iPA in the solvent mixes of E4 and E5, 100 % of MeOH is still relatively more efficient for extracting non-polar compounds. The H_2O seems to be the dominant component in protocols E4 and E5.

In conclusion, the E4 and the E5 protocols has proven their superiority for the extraction and/or the detection of xenometabolites (particularly the formulation agents) and for polar and semi-polar endometabolites (*e.g.* primary metabolites). The E3 protocol is more suitable for the extraction of non-polar xenometabolites as lipids and secondary metabolites.

It is worth mentioning that results provided by the Dendrogram-Heatmap analyses are supporting the observations discussed when assessing the raw data (BPCs and MF counting – Section 3.1.). In fact, Heatmap plots showed 1) the poor metabolic profiles extracted by the E1 protocol, 2) the rich endometabolome of the SP when compared to ST, and 3) the complex xenometabolome of the formulated Nonanoic acid herbicide when compared to the xenometabolome of the formulated Glyphosate herbicide. Nonetheless, the Heatmap was unable to reveal the xenometabolome that was exclusively detected in ST.

3.3. Assessment of extractions reproducibility

To assess extractions' reproducibilities, the calculation of Euclidean Distances was selected as a statistical tool that allows assessing the similarities between the analyzed metabolic profiles of several injections. Therefore, Euclidean Distances measures were applied to the 5 environmental/biological replicates belonging to each group. An RSD is calculated as a score allowing the assessment of the reproducibility of the extraction protocol in a given environmental condition. The original dataset containing ion redundancies was used for these calculations in order to include all possible variations. Results are summarized in Table C.I 2.

Table C.I 2: RSDs of Euclidean Distances.
For the 30 QC injections, the RSD is 26.59 %. It is defined as the score of the analytical variability. An RSD of 30 % is considered as the threshold of optimum for extraction reproducibility.
*: RSD below 30 % (optimal reproducibility).

Group	**RSD of Euclidean Distances (%)**				
	E1	***E2***	***E3***	***E4***	***E5***
SPCtr	* 26.67	66.79	72.99	* 22.07	* 26.39
STCtr	* 23.03	35.19	* 25.35	* 26.29	* 29.64
SPGly	40.24	* 25.04	36.77	* 27.45	38.42
STGly	34.67	* 23.99	41.65	55.08	77.14
SPBel	89.86	* 22.05	* 20.49	* 27.44	* 17.66
STBel	* 26.38	38.41	40.08	38.47	* 27.51

The results show that for Ctr and Bel, E5 proved a good reproducibility. For Gly, however, the reproducibility was altered. The most reproducible extraction for the Gly in both soils is the E2. The E4 extraction also shows an acceptable reproducibility. One hypothesis could be drawn in order to explain this result: Glyphosate xenometabolites elutes in the semi-polar zone, and soils semi-polar endometabolites present higher abundances in E4 and E5. Thus, co-elutions are occurring between these metabolites, which can alter their ionization and detection. This phenomenon is apparently not occurring for the xenometabolites of the Nonanoic acid herbicide. To prove these hypotheses, manual integration of peak areas of EICs belonging to the co-eluted compounds should be performed and compared by simple RSD calculations. This data processing step will be considered as a perspective for the current work.

It should be noted that the Euclidean Distances tool should be assessed in-depth in order to evaluate its vulnerability to be influenced by an extreme variation occurring in only one variable (molecular feature). This issue can cause the increasing of the RSD due to one non-significant random contamination or artefact, meanwhile the total metabolic profile is well reproduced between the replicates. Hence, this issue can alter the interpretation of the assessment of extraction reproducibility.

Another point to be evaluated is the influence of the "size effect" on the RSD values. In fact, if the intensities of features are high, the RSD values tend to decrease. Thus, the quantitative performance can dominate and hide the variations, and the assessment of the reproducibility can be altered. To resolve this "size effect" issue, one potential solution is to apply the glog transformation on the data before applying the Euclidean Distance measures. The assessment of this potential solution will be considered as a perspective of the current work.

3.4. Assessment of the discriminant powers of the extraction protocols

The ability of an extraction protocol to discriminate between a polluted soil and an unpolluted control soil is a key point for the EMF approach. The Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) [40] and its Cross-Validation were selected as a suitable tool in order to assess this important factor. In fact, the OPLS-DA is an explicative supervised multivariate analysis that prioritize the differences between the two defined conditions/groups to be compared (polluted *vs.* unpolluted). It filtrates and eliminates all random variations, and reveal the systematic variations that are orthogonal to the main investigated factor (*i.e.* the pollution and its impact). The investigation of such systematic orthogonal variations allows for the assessing of the reliability and the confidence of the discrimination power [41,42]. All these factors are explained by the different scores provided by the Cross Validation; the p dimension is the predictive component that examines the discrimination power between the compared factors. Its R2X represents the percentage of the explained variation through the discriminative component (the explained variation by a regressed number of features that are considered systematically discriminative between the two factors). The R2Y allows assessing the model performance by representing the coefficient of correlation of the two groups of samples to the explained discrimination. The Q2 represents the predictivity of the model, which is directly related to the discrimination power and its significance or its overfitting. The o dimensions are the orthogonal components. Each of them represents the explained variation by a regressed number of features that are undergoing a systematic variation that is orthogonal to the main component "p". Its R2X represents the percentage of the variability explained by the component. Its R2Y allows assessing the correlation of the two groups of samples to the orthogonal systematic variation, and Q2 shows the significance of this orthogonal systematic variation. If R2Y and Q2 are above 50 %, the orthogonal variation is thus dominating and can put in question the reliability of the results and/or the analytical method.

Results of OPLS-DA Cross-Validations are shown in Table C.I 3. For each extraction protocol, 4 different analyses were performed as the following: SPCtr *vs.* SPGly, STCtr *vs.* STGly, SPCtr *vs.* SPBel, and STCtr *vs.* STBel. Analyses were performed on the original dataset after glog transformation and Pareto scaling were applied (glog-transformed intensities are mean-centered and divided by the square root of the standard deviation of each variable [43]).

Table C.I 3: OPLS-DA Cross-Validation results. Performed on the original dataset after glog transformation and Pareto scaling.
All data are in %.

	SP						**ST**					
	p1			***o1***			***p1***			***o1***		
	R2X	*R2Y*	*Q2*	*R2X*	*R2Y*	*Q2*	*R2X*	*R2Y*	*Q2*	*R2X*	*R2Y*	*Q2*
Gly												
E1	17.30	97.30	60.60	11.50	02.61	04.11	17.40	97.40	62.10	13.00	02.48	03.56
E2	19.80	97.20	70.10	16.00	02.61	04.93	18.10	98.50	64.70	13.60	01.35	04.35
E3	19.80	99.00	70.30	12.10	00.89	05.74	18.70	96.70	62.60	14.90	03.19	04.38
E4	**25.50**	98.00	**82.60**	12.40	01.83	02.81	**20.60**	93.30	**69.80**	18.10	06.47	08.08
E5	**20.30**	97.20	**70.80**	12.00	02.52	09.92	**19.20**	98.90	**67.20**	14.40	01.03	05.06
Bel												
E1	22.30	99.20	78.80	10.80	00.70	04.37	20.30	99.00	73.10	10.70	01.02	02.77
E2	29.90	99.40	89.30	12.80	00.57	02.03	29.30	98.80	87.40	10.40	01.13	01.81
E3	26.40	98.50	82.40	14.30	01.40	03.20	25.60	97.50	81.70	13.00	02.43	02.79
E4	**36.20**	99.70	**94.20**	12.00	00.30	01.06	**32.00**	99.00	**91.70**	17.30	00.91	01.89
E5	**34.10**	98.50	**90.30**	11.10	01.43	01.26	**31.20**	99.70	**91.10**	11.90	00.25	01.99

The results show that for all extraction protocols, the OPLS-DA was able to reveal significant and reliable discriminations between the polluted soils and the control soils ($R2Y_{(p1)} > R2X_{(p1)}$, $R2Y_{(p1)} > Q2_{(p1)}$, $R2Y_{(p1)} - Q2_{(p1)} < 30$ %[20], and $Q2_{(p1)} > 50$ % [42]). In addition, no significant systematic orthogonal variations were found (o1). The most powerful discriminations were observed in E4, where the $R2X_{(p1)}$ and the $Q2_{(p1)}$ showed the highest scores in all the examined comparisons (Table C.I 3 – in green). The contributions in variations ($R2X_{(p1)}$) prove that the E4 is the most performant for xenometabolome extraction and detection. The high significance revealed by the $Q2_{(p1)}$ shows that this extraction is still able to discriminate the polluted soil despite the presence of a rich and complex metabolic profile originating from soils' endometabolomes (particularly the endometabolome of SP). On the other hand, $R2X_{(p1)}$ and $Q2_{(p1)}$ scores show that E5 is the second most performant protocol for xenometabolome extraction and for the discrimination between unpolluted and polluted soils (Table C.I 3 – in red).

It is worth mentioning that the results of the OPLS-DA Cross-Validations that were performed on the light dataset (Table A.I 6 – Appendix I) showed similar results that can lead to identical conclusions, which demonstrates that in the current case, the elimination of ion redundancies did not significantly affect the metabolic information held by the datasets and explained by the statistical data analyses. This conclusion is supported by the results of the Principal Component

[20] Except Between STGly and STCtr in E5 (= 31.70 %).

Analysis (PCA) [44] that was performed on both the original and the light data matrices (Section A.I.3.1. – Appendix I).

On the other hand, one limitation of the use of the OPLS-DA should be noted. This limitation can affect the assessment of the discrimination between the polluted and the unpolluted groups. In fact, a biased interpretation of the OPLS-DA results can occur when an extraction protocol is exclusively more selective to one of the two components of the meta-metabolome (*i.e.* the xenometabolome or the endometabolome). Such selectivity could affect the significance of the discrimination power. For instance, if the extraction protocol is moderately selective for the xenometabolites, but extremely less efficient for the endometabolome extraction, the discrimination between the two compared groups will appear to be highly significant. Therefore, in order to avoid such biased results interpretations, the coupling of the OPLS-DA to the Dendrogram-Heatmap analyses is required.

4. Perspectives

The current work still need further in-depth investigations. The relative quantification of herbicides' active compounds through the different protocols and the different soil samples should be performed and discussed. The MS/MS annotation is also a perspective for the current study. It should be performed in order to identify the extracted metabolites and to help explaining the different results highlighted by the statistical analyses.

5. Conclusions

In the present work, a novel analytical development strategy based on untargeted metabolic profiling was introduced to assess the development of novel extraction protocols dedicated for the EMF proxy. The strategy investigated the performance of two novel 2-steps-based mixed solvents extractions that aim to broaden the detection of polluted soils xenometabolites and endometabolites. These extractions were compared to previously published protocols.

The automated data pre-processing combined with divers multivariate statistical tools were performed in order to assess the different analytical criteria that define the "optimum". Heatmap analyses allowed investigating the covered metabolic diversity and the associated extraction yields. For the two examined soil types and the applied herbicides (*e.g.* Glyphosate and Nonanoic acid), results showed that the novel protocols E4 and E5 were the most performant for xenometabolome and polar and semi-polar endometabolome extraction. These two protocols have been shown capable for extracting non-polar endometabolome, but with a lower performance if compared to 100 % MeOH-based extractions. The OPLS-DA and its Cross-Validation showed that E4, followed by E5, were the most powerful extractions to discriminate between the polluted and the unpolluted soils. The Euclidean Distances measures showed that the extractions of the meta-metabolomes of unpolluted soils and soils polluted with the Nonanoic acid herbicide were the most reproducible with E4 and E5. However, for the soils polluted with the Glyphosate herbicide, the reproducibility was degraded for the E5 (nut acceptable for E4). Further data investigations are needed in order to understand and explain the alteration of the reproducibility in these conditions.

Hence, for the current study, E4 followed by E5 were considered as the optimal extractions. Nevertheless, it should be mentioned that the definition of the "optimum" still depends on the environmental conditions, as soil properties and the applied herbicides. The assessment of other extraction protocols is required when the environmental conditions, the factors or the aims are

different. For instance, the present study demonstrated that the protocol E3 showed a relevant performance for non-polar metabolites extraction. The novel untargeted metabolic profiling-based strategy can thus be adopted as a suitable tool that helps assessing the optimal protocols and methods dedicated for defined conditions and objectives. This novel strategy has been shown as a powerful tool for the analytical development of untargeted (meta-)metabolomics approaches that consist of large datasets and complex multifactorial experiments. It succeeded to surpass the limitations of the observational examination of BPCs and was able to examine the defined analytical criteria successfully.

Acknowledgments

Authors would like to acknowledge the collaborators from the Sica CENTREX – Torreilles, particularly Mrs. Aude Lusetti, for supplying herbicides solutions and helping in soil sampling. Mélina Ramos, Engr. (CRIOBE USR3278 EPHE-CNRS-UPVD – Perpignan) and Dr. Ludivine Garcia (IUT – Perpignan) are also acknowledged for helping in soil sampling. Colleagues from the Centre de Formation et de Recherche sur les Environnements Méditerranéens (CEFREM UMR5110 CNRS-UPVD – Perpignan), particularly Dr. Thierry Courp and Dr. Bruno Charrière are acknowledged for supplying soil sieving material. Acknowledgments to Dr. Nicolas Le Yondre and Philippe Jéhan, Engr. (CRMPO – Rennes) for their valuable advices regarding Q/ToF tunings. Authors also acknowledge Dr. Yann Guitton (LABERCA – Nantes), Dr. Anne-Emmanuelle Hay (Ecologie Microbienne – Lyon), and Pr. Thierry Noguer (BAE-LBBM – Perpignan) for their scientific advices that helped improving the current work.

This work was supported by the European Regional Development Fund (ERDF) under the Interreg POCTEFA PALVIP project (POCTEFA 2014-2020). The funding institution had no role in the experimental design, the data processing, or in writing and reviewing the manuscript.

Ph.D. fellowship grant was awarded to HG by the French Ministry of Higher Education, Research and Innovation (MESRI), *via* the Doctoral School ED 305 “Energie et Environnement” (Université de Perpignan Via Domitia).

The LC-ESI-Q/ToF method developments and analyses had been performed using the Biodiversité et Biotechnologies Marines (Bio2Mar) facilities – Métabolites Secondaires Xénobiotiques Métabolomique Environnementale (MSXM) platform at the Université de Perpignan Via Domitia (http://bio2mar.obs-banyuls.fr/).

List of Abbreviations

ACN: Acetonitrile

AWC: Available Water Capacity

BPC: Base Peaks Chromatogram

C/N: Carbon-to-Nitrogen ratio

CC: Convergence Chromatography

CEC: Cation Exchange Capacity

CID: Collision-Induced Dissociation

CV: Coefficient of Variation

DDA: Data Dependent Acquisition

EIC: Extracted Ion Chromatogram

EMF: Environmental Metabolic Footprinting

ESI: Electrospray Ionization

EtOAc: Ethyl Acetate

FA: Formic Acid

FMF: Find Molecular Features

FWHM: Full Width at Half Maximum

glog: generalized logarithm transformation

HPC: High Precision Calibration

HPCE: High Performance Capillary Electrophoresis

HPLC: High Performance Liquid Chromatography

HRMS: High Resolution Mass Spectrometry

iPA: 2-Propanol

IS: Internal Standard

LC: Liquid Chromatography

MeOH: Methanol

MF: Molecular Features

MS/MS: Tandem Mass Spectrometry

MS: Mass Spectrometry

NaF: Sodium Formate

NetCDF: Network Common Data Form

OPLS-DA: Orthogonal Projections to Latent Structures Discriminant Analysis

PCA: Principal Component Analysis

ppm: Parts-per-Million

PTFE: Polytetrafluoroethylene

Q/ToF: Quadrupole/Time-of-Flight

QC: Quality Control

RF: Radio Frequency

ROI: Region of Interest

RPM: Revolutions per Minute

RSD: Relative Standard Deviation

RT: Retention Time

S/N: Signal-to-Noise ratio

SD: Standard Deviation

SFC: Supercritical Fluid Chromatography

UHPLC: Ultra High Performance Liquid Chromatography

Vpp: Peak-to-Peak Voltage

References

[1] M. Gavrilescu, Fate of Pesticides in the Environment and its Bioremediation, Eng. Life Sci. 5 (2005) 497–526. https://doi.org/10.1002/elsc.200520098.

[2] H.M.G. van der Werf, Assessing the impact of pesticides on the environment, Agriculture, Ecosystems & Environment. 60 (1996) 81–96. https://doi.org/10.1016/S0167-8809(96)01096-1.

[3] E. Withers, P.W. Hill, D.R. Chadwick, D.L. Jones, Use of untargeted metabolomics for assessing soil quality and microbial function, Soil Biology and Biochemistry. 143 (2020) 107758. https://doi.org/10.1016/j.soilbio.2020.107758.

[4] J.S. Buyer, B. Vinyard, J. Maul, K. Selmer, R. Lupitskyy, C. Rice, D.P. Roberts, Combined extraction method for metabolomic and PLFA analysis of soil, Applied Soil Ecology. 135 (2019) 129–136. https://doi.org/10.1016/j.apsoil.2018.11.012.

[5] C. Bertrand, C. Prigent-Combaret, A. Gonzales-Coloma, Chemistry, activity, and impact of plant biocontrol products, Environ Sci Pollut Res. 25 (2018) 29773–29774. https://doi.org/10.1007/s11356-018-3209-2.

[6] T.L. Swenson, S. Jenkins, B.P. Bowen, T.R. Northen, Untargeted soil metabolomics methods for analysis of extractable organic matter, Soil Biology and Biochemistry. 80 (2015) 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007.

[7] M.A. O'Malley, Metametabolomics, in: W. Dubitzky, O. Wolkenhauer, H. Yokota, K.-H. Cho (Eds.), Encyclopedia of Systems Biology, 1st ed., Springer-Verlag New York, New York, NY, 2013: pp. 1296–1297. https://doi.org/10.1007/978-1-4419-9863-7_903.

[8] W. Dubitzky, O. Wolkenhauer, H. Yokota, K.-H. Cho, eds., Encyclopedia of Systems Biology, 1st ed., Springer-Verlag New York, New York, NY, 2013. https://www.springer.com/gp/book/9781441998620 (accessed August 15, 2020).

[9] O.A.H. Jones, S. Sdepanian, S. Lofts, C. Svendsen, D.J. Spurgeon, M.L. Maguire, J.L. Griffin, Metabolomic analysis of soil communities can be used for pollution assessment: Metabolomics of soil communities, Environ Toxicol Chem. 33 (2014) 61–64. https://doi.org/10.1002/etc.2418.

[10] C. Patil, C. Calvayrac, Y. Zhou, S. Romdhane, M.-V. Salvia, J.-F. Cooper, F.E. Dayan, C. Bertrand, Environmental Metabolic Footprinting: A novel application to study the impact of a natural and a synthetic β-triketone herbicide in soil, Science of The Total Environment. 566–567 (2016) 552–558. https://doi.org/10.1016/j.scitotenv.2016.05.071.

[11] M.-V. Salvia, A. Ben Jrad, D. Raviglione, Y. Zhou, C. Bertrand, Environmental Metabolic Footprinting (EMF) vs. half-life: a new and integrative proxy for the discrimination between control and pesticides exposed sediments in order to further characterise pesticides' environmental impact, Environ Sci Pollut Res. 25 (2018) 29841–29847. https://doi.org/10.1007/s11356-017-9600-6.

[12] H. Ghosson, D. Raviglione, M.-V. Salvia, C. Bertrand, Online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted volatile metabolomics for studying emerging complex biopesticides: a proof of concept, Analytica Chimica Acta. 1134 (2020) 58–74. https://doi.org/10.1016/j.aca.2020.08.016.

[13] S. Romdhane, M. Devers-Lamrani, L. Barthelmebs, C. Calvayrac, C. Bertrand, J.-F. Cooper, F.E. Dayan, F. Martin-Laurent, Ecotoxicological Impact of the Bioherbicide Leptospermone on the Microbial Community of Two Arable Soils, Front. Microbiol. 7 (2016). https://doi.org/10.3389/fmicb.2016.00775.

[14] M. Anastassiades, D.I. Kolberg, A. Benkenstein, E. Eichhorn, S. Zechmann, D. Mack, C. Wildgrube, I. Sigalov, D. Dörk, A. Barth, Quick Method for the Analysis of numerous Highly Polar Pesticides in Foods of Plant Origin via LC-MS/MS involving Simultaneous Extraction with Methanol (QuPPe-Method), EU Reference Laboratories for Residues of Pesticides, Fellbach, Germany, 2016.
https://www.eurl-pesticides.eu/userfiles/file/EurlSRM/meth_QuPPe_Version9_1.pdf
(accessed June 15, 2020).

[15] United States Department of Agriculture - Natural Resources Conservation Service, Soil Texture Calculator, Natural Resources Conservation Service. (n.d.).
https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_054167
(accessed September 1, 2020).

[16] United States Natural Resources Conservation Service - Soil Science Division, Soil Survey Manual, revised, United States Department of Agriculture, 2017. https://books.google.fr/books?id=dieUtAEACAAJ.

[17] S. Romdhane, M. Devers-Lamrani, J. Beguet, C. Bertrand, C. Calvayrac, M.-V. Salvia, A.B. Jrad, F.E. Dayan, A. Spor, L. Barthelmebs, F. Martin-Laurent, Assessment of the ecotoxicological impact of natural and synthetic β-triketone herbicides on the diversity and activity of the soil bacterial community using omic approaches, Science of The Total Environment. 651 (2019) 241–249. https://doi.org/10.1016/j.scitotenv.2018.09.159.

[18] Z. Liu, S. Rochfort, A fast liquid chromatography–mass spectrometry (LC–MS) method for quantification of major polar metabolites in plants, Journal of Chromatography B. 912 (2013) 8–15. https://doi.org/10.1016/j.jchromb.2012.10.040.

[19] H. Ghosson, A. Schwarzenberg, F. Jamois, J.-C. Yvin, Simultaneous untargeted and targeted metabolomics profiling of underivatized primary metabolites in sulfur-deficient barley by ultra-high performance liquid chromatography-quadrupole/time-of-flight mass spectrometry, Plant Methods. 14 (2018) 62. https://doi.org/10.1186/s13007-018-0329-0.

[20] R. Smith, A.D. Mathis, D. Ventura, J.T. Prince, Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view, BMC Bioinformatics. 15 (2014) S9. https://doi.org/10.1186/1471-2105-15-S7-S9.

[21] H. Ghosson, Y. Brunato, D. Raviglione, M.-V. Salvia, C. Bertrand, MTBLS2044: Liquid Chromatography-High Resolution Mass Spectrometry-based untargeted profiling as a tool for analytical development: assessment of novel extraction protocols for herbicide-polluted soil meta-metabolomics, MetaboLights. (2020). https://www.ebi.ac.uk/metabolights/MTBLS2044 (accessed September 6, 2020 - Status: Submitted).

[22] K. Haug, K. Cochrane, V.C. Nainala, M. Williams, J. Chang, K.V. Jayaseelan, C. O'Donovan, MetaboLights: a resource evolving in response to the needs of its scientific community, Nucleic Acids Research. 48 (2019) D440–D444. https://doi.org/10.1093/nar/gkz1019.

[23] Galaxy Workflow4Metabolomics, Galaxy Workflow4Metabolomics. (n.d.). https://galaxy.workflow4metabolomics.org/ (accessed February 28, 2020).

[24] Y. Guitton, M. Tremblay-Franco, G. Le Corguillé, J.-F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E.A. Thévenot, Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics, The International Journal of Biochemistry & Cell Biology. 93 (2017) 89–101. https://doi.org/10.1016/j.biocel.2017.07.002.

[25] F. Giacomoni, G. Le Corguille, M. Monsoor, M. Landi, P. Pericard, M. Petera, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E.A. Thevenot, C. Caron, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics. 31 (2015) 1493–1495. https://doi.org/10.1093/bioinformatics/btu813.

[26] H. Ghosson, Y. Brunato, D. Raviglione, M.-V. Salvia, C. Bertrand, Liquid Chromatography-High Resolution Mass Spectrometry-based untargeted profiling as a tool for analytical development: assessment of novel extraction protocols for herbicide-polluted soil meta-metabolomics - data preprocessing pipeline, Galaxy Workflow4Metabolomics. (2020). https://workflow4metabolomics.usegalaxy.fr/u/hikmatghosson/w/w4m2044 (accessed September 6, 2020).

[27] 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.

[28] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics. 9 (2008) 504. https://doi.org/10.1186/1471-2105-9-504.

[29] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets, Anal. Chem. 84 (2012) 283–289. https://doi.org/10.1021/ac202450g.

[30] F.M. van der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema, Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping, J. Proteome Res. 8 (2009) 5132–5141. https://doi.org/10.1021/pr900499r.

[31] A. Roux, D. Lison, C. Junot, J.-F. Heilier, Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review, Clinical Biochemistry. 44 (2011) 119–135. https://doi.org/10.1016/j.clinbiochem.2010.08.016.

[32] MetaboAnalyst, MetaboAnalyst. (n.d.). https://www.metaboanalyst.ca/ (accessed February 28, 2020).

[33] J. Chong, D.S. Wishart, J. Xia, Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis, Current Protocols in Bioinformatics. 68 (2019) e86. https://doi.org/10.1002/cpbi.86.

[34] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Research. 46 (2018) W486–W494. https://doi.org/10.1093/nar/gky310.

[35] H. Ghosson, The aggregation of soil material following the application of Ethyl Acetate solvent, CRIOBE USR3278 - EPHE-CNRS-UPVD, 2019. https://hal.archives-ouvertes.fr/hal-02966421 (accessed October 16, 2020).

[36] G. Grass, C. Rensing, M. Solioz, Metallic Copper as an Antimicrobial Surface, Appl. Environ. Microbiol. 77 (2011) 1541–1547. https://doi.org/10.1128/AEM.02766-10.

[37] A. Furey, M. Moriarty, V. Bane, B. Kinsella, M. Lehane, Ion suppression; A critical review on causes, evaluation, prevention and applications, Talanta. 115 (2013) 104–122. https://doi.org/10.1016/j.talanta.2013.03.048.

[38] E. Kenar, H. Franken, S. Forcisi, K. Wörmann, H.-U. Häring, R. Lehmann, P. Schmitt-Kopplin, A. Zell, O. Kohlbacher, Automated Label-free Quantification of Metabolites from Liquid Chromatography–Mass Spectrometry Data, Mol Cell Proteomics. 13 (2014) 348–359. https://doi.org/10.1074/mcp.M113.031278.

[39] L. Wilkinson, M. Friendly, The History of the Cluster Heat Map, The American Statistician. 63 (2009) 179–184. https://doi.org/10.1198/tas.2009.0033.

[40] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics. 16 (2002) 119–128. https://doi.org/10.1002/cem.695.

[41] J. Boccard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Analytica Chimica Acta. 769 (2013) 30–39. https://doi.org/10.1016/j.aca.2013.01.022.

[42] S. Wiklund, Multivariate data analysis for Omics, (2008). https://metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf (accessed October 18, 2020).

[43] R.A. van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics. 7 (2006) 142. https://doi.org/10.1186/1471-2164-7-142.

[44] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2 (1901) 559–572. https://doi.org/10.1080/14786440109462720.

Chapter II

"A biomarker or a suppressed ion?"

Preamble

In the EMF approach, the determination and the characterization of **environmental biomarkers** are essential steps towards the explanation of pesticide's impact on the microbial communities of a given matrix. These biomarkers can be detected in the endometabolome that is produced and influenced by the microbial activity.

As the EMF is based on untargeted metabolomics-based approaches, the first step to determine biomarkers passes through **the comparison of the "relative intensities" of the detected endometabolites between the control samples and the polluted samples**. The relative intensities represent semi-quantitative indicators that are directly related to the concentration of the metabolite in the analyzed sample.

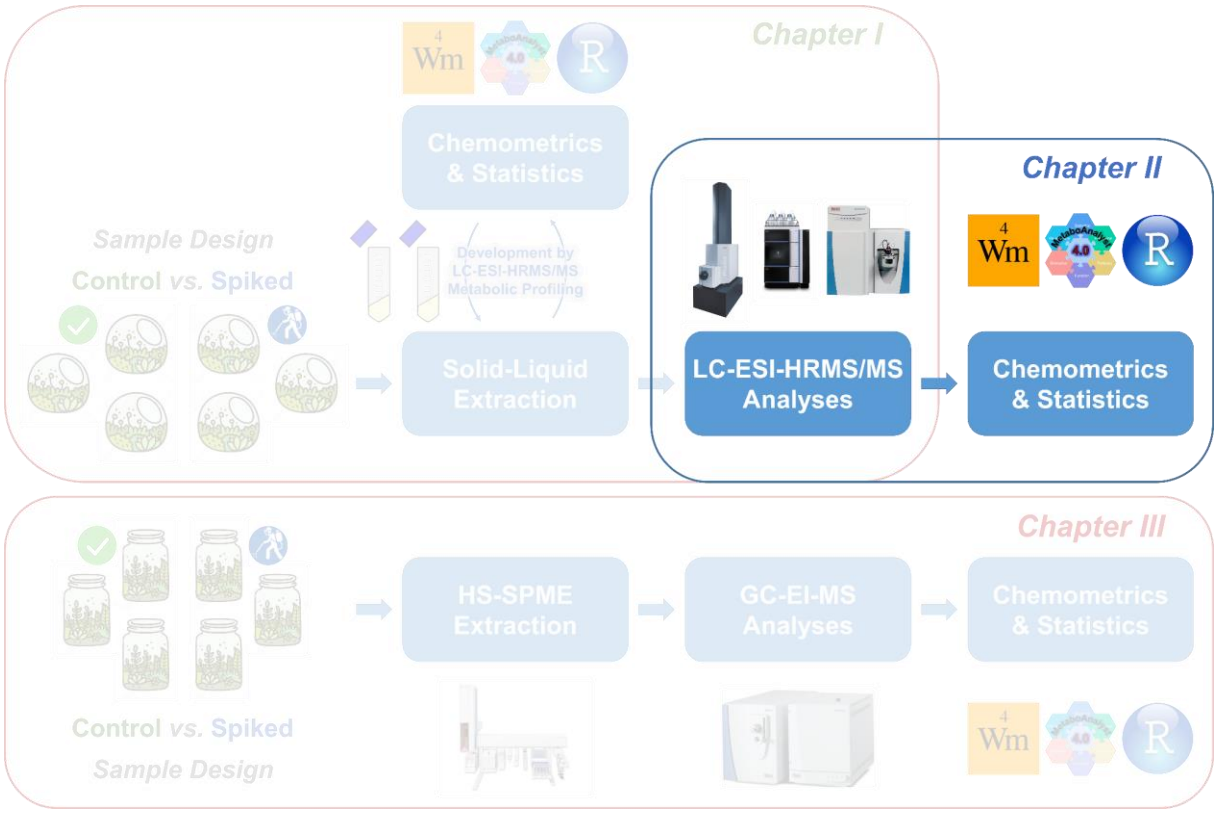
In Mass Spectrometry, the relative intensities are generated following the ionization and the ion detection process. However, when it comes to Electrospray Ionization, the process can be influenced by several factors and mechanisms. One of the most common phenomenon is **the "matrix effect" that directly affects the relative intensities** of the metabolites by influencing their ionization and/or detection. The matrix effect can be a function of sample complexity. The higher is the complexity of the metabolic profile, the higher is the occurrence of matrix effect. Thus, samples with significantly heterogeneous metabolic profiles can present different matrix effects, and consequently different analytical responses. **This problematic renders uncertain the study based on the comparison of relative intensities between samples**. In fact, intensity variation can be issued from **the difference of matrix effect**, and **not from a biological response**.

In the EMF case, the compared samples theoretically exhibit different complexities in their metabolic profiles. The control-unpolluted sample contains endometabolites only. However, the polluted sample contains endometabolites and xenometabolites originating from the applied pesticides. The xenometabolome of a pesticide is known to be highly complex, as it consists of active substances, formulation ingredients and transformation by-products. Thus, a significant difference in term of complexity can be expected between the compared meta-metabolic profiles. A heterogeneous matrix effect between samples can consequently occur.

In the present Chapter, the coincident observation and the awareness of Mass Spectrometry fundamentals will lead to prove these concerns. The **Ion Suppression** phenomenon will be

detected when searching for biomarker candidates. Therefore, the occurrence, the causes and the impact of this matrix effect will be explained. **A pragmatic analytical solution will be then implemented in order to overcome the influence of this analytical drift** and its consequences on the subsequent experiments and biological/environmental conclusions.

It should be noted that the work that was achieved in the current Chapter is the result of the resumption of a research work that started before the beginning of the thesis. The initial work consisting of a research internship (done by A. Ben Jard) had already performed the extractions and prepared the environmental samples that were then analyzed during the current thesis. This chronological fact explains the reason behind the non-application of the extraction protocols that were developed and described in Chapter I.



Electrospray Ionization and heterogeneous matrix effects in Liquid Chromatography-Mass Spectrometry-based meta-metabolomics: A biomarker or a suppressed ion?

Hikmat Ghosson[1,2,*], Yann Guitton[3], Amani Ben Jrad[1,2], Chandrashekhar Patil[1], Delphine Raviglione[1,2], Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,4,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), Oniris, INRAE, Nantes, F-44307, France

4: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: *Equal Contribution (Last Co-authors)*

Publication

Ghosson, H. *et al. Rapid Commun. Mass Spectrom.* **(2021)**, 35(2):e8977. doi:10.1002/rcm.8977

Keywords

Biomarkers; Ion Suppression; Matrix Effect; Untargeted Metabolomics; Liquid Chromatography-Mass Spectrometry; Electrospray Ionization

Abstract

Rationale

Correct biomarkers determination in metabolomics is crucial for unbiased conclusions and reliable applications. However, this determination is subjected to several drifts, *e.g.* matrix effects and ion suppression in Liquid Chromatography-Mass Spectrometry-based approaches. This phenomenon provokes critical issues for biomarkers determination, particularly during comparative studies dealing with samples exhibiting heterogeneous complexities.

Methods

Occurrence of the issue was coincidentally noticed when studying the environmental impact of a complex bioinsecticide: *Bacillus thuringiensis israelensis*. Studied samples consisted of insecticide-spiked sediments and untreated control sediments. QuEChERS extractions followed by LC-ESI-Q/ToF analyses were performed on sediments after 15 days of incubation. Meta-metabolomes containing pesticide xenometabolites and sediments' endometabolites were in-depth analyzed using XCMS-based computational data preprocessing. Multivariate statistical analyses (PCA, OPLS-DA) and raw data crosschecks were performed to search for environmental biomarkers.

Results

Multivariate analyses and raw data crosschecks led to the selection of 9 metabolites as biomarker candidates. However, when exploring mass spectra, co-elutions were noticed between 7 of these metabolites and multi-charged macromolecules originating from the pesticide. Provoked false positives were thus suspected due to a potential ion suppression exclusively occurring in the spiked samples. A dilution-based approach was then applied. It confirmed 5 metabolites as suppressed ions.

Conclusions

Ion suppression should be considered as a critical issue for biomarkers determination when comparing heterogeneous metabolic profiles. Raw chromatograms and mass spectra crosscheck is mandatory to reveal potential ion suppressions in such cases. The dilution is a suitable approach to filtrate reliable biomarker candidates before their identification and absolute quantification.

1. Introduction

Mass Spectrometry-based metabolomics is a growing domain of analytical chemistry that targets small organic molecules (> 1000-1500 Da). It is widely applied for different research fields [1,2]. One of its major objectives is to discover new biomarkers [3–7]. Biomarkers discovery helps to understand and explain the impact of biotic and/or abiotic factors that affect biological systems. Therefore, a correct determination of biomarkers is essential for unbiased scientific conclusions and reliable applications. This process requires rigorous protocols that are decreasingly considered in untargeted metabolomics studies over the recent past years [8–11].

The reliable determination of a biomarker is assessed by i) the significance of its variation in abundance following the application of a defined condition, and ii) the correlation of this variation to the applied condition. Nonetheless, the quantification of metabolites is vulnerable to several variations unrelated to the applied condition, such as the variations related to pre-analytical or analytical issues, *e.g.* the biological variation between samples, the low repeatability of the extraction protocol, and the analytical/instrumental drifts. In addition, other aberrations related to data handling and statistical analyses can occur [8,12–15]. Studies that explore the pre-analytical and analytical issues have been extensively reported in the literature [8,9,16–18]. For instance, biological variations between samples and the impact of these variations on determining biomarkers was reviewed by Wu & Li [16]. This work assessed different sample normalization approaches. For Liquid Chromatography-Mass Spectrometry (LC-MS) -based metabolomics, the work highlighted the importance of adjusting all samples to a standard concentration before the analysis, as the difference in concentration levels can lead to different matrix effects and thus a heterogeneous analytical response. Analytical and instrumental drifts were addressed by Broadhurst *et al.* [17]. This work proposed guidelines for Quality Control (QC) and sustainability assurance of MS-based metabolomics studies. The work highlighted the risks of metabolite response variability through matrix-specific and sample-specific ionization suppression, particularly when the strategy of "biologically-identical QC samples" is adopted. Thus, matrix effect is often reported as a major issue in metabolomics. This phenomenon must be considered in order to avoid data misinterpretation. In this context, the present work focuses on a widely-known type of matrix effect: the "Ion Suppression" phenomenon; an analytical/instrumental drift mainly observed during LC-MS-based metabolomics experiments, and in particular, when the applied ionization mode is the

commonly-used [19] Electrospray Ionization (ESI) source (as well for the Atmospheric Pressure Chemical Ionization (APCI) source).

Since its introduction in 1984 by Alexandrov *et al.* [20–22], and Yamashita & Fenn [23], ESI source revolutionized the Mass Spectrometry-based analytical methods. It is widely used for LC-MS-based metabolomics approaches, mainly for analyzing liquid samples, containing polar and semi-polar, nonvolatile and thermally unstable metabolites. Despite its advantages for metabolome analyses, the ESI presents several drawbacks including its vulnerability to matrix effect and to the ion suppression phenomenon that is well described and discussed in the literature [24–26]. It leads to a decrease in the metabolite signal in MS due to several causes and mechanisms related to matrix complexity. The main ion suppression mechanisms were documented by Antignac *et al.* [24] and Furey *et al.* [26] and can be summarized as follows:

- A competition between co-eluting molecules to access the available charge.
- The precipitation of the analyte by co-eluting macromolecules.
- The change in surface viscosity of spray droplets, due to the presence of non-volatile (macro)molecules. This presence can decrease the evaporation of the droplets solvent and thus prevents analytes from being emitted to the gas phase [27].
- The neutralization of the analyte's charge by a co-eluting compound after it is emitted to the gas phase (due to the relative basicity in the gas phase).

Hence, ion suppression is a function of matrix concentration and complexity. This phenomenon can behave differently between samples, leading to "sample-to-sample" variations [24–26]. Further, this different behavior between samples can risk the reliability in determining biomarkers, such as when the suppression of a metabolite signal occurs exclusively in a specific group of samples due to the higher complexity of their metabolic profiles, or if other components of their matrix are different (*e.g.* pH, degree of salinity, or concentration of inorganic compounds). Clearly, this phenomenon can be problematic in untargeted metabolomics, where diverse conditions are examined. This can lead to compare samples exhibiting heterogeneous complexities and different levels of matrix effect.

The present paper addresses the expected "selective" ion suppression and its risks on environmental biomarkers determination, after it was coincidentally observed and then studied in-depth during the development of an LC-HRMS-based meta-metabolomics approach called Environmental Metabolic Footprinting (EMF) [28,29]. The EMF concept is summarized in Figure A.II-A 1 (Appendix II-A). The broad aim of the EMF is to assess fate and impact of

Biocontrol Agents (BA) in the environment. Therefore, the studied complex BA – the *Bacillus thuringiensis israelensis* (Bti) – is applied to a group of sediment samples, in order to compare their meta-metabolome (xenometabolome + endometabolome) to the meta-metabolome of untreated control sediments (containing only endometabolome). As shown in Figure C.II 1, the approach deals with groups of samples that present heterogeneous complexities. The spiked samples are more complex than the untreated control samples due to the presence of BA's xenometabolome. This xenometabolome is a complex mixture of Bti metabolites, peptides, and formulation agents containing polymers. The present work explores the occurrence of the heterogeneous "group-to-group" ion suppression issue, and the associated risks and difficulties when determining environmental biomarkers. The objective of this investigation is to highlight an experimentally proven example of this challenging problem in untargeted metabolomics. Moreover, this study proposes a simple protocol dedicated to filtrate the relevant biomarker candidates, using a dilution-based approach [24–26].

Ultimately, the current contribution seeks to propose and discuss a simple workflow that is accessible for metabolomics developers and users, in order to help identify ion suppression and to avoid false positives and false negatives. This workflow allows for the filtration of reliable "biomarker" candidates before processing in subsequent complex steps such as metabolite annotation/characterization and absolute quantification by stable isotope-labeled reference standards.

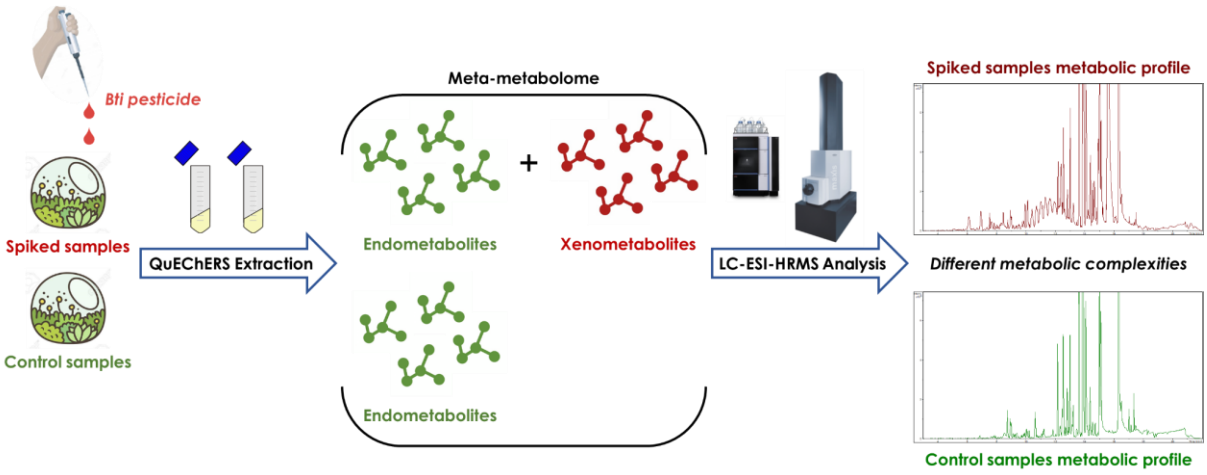


Figure C.II 1: Unlike for the untreated control sediments, the application of a complex formulated BA increases the complexity of spiked sediments' meta-metabolome.

2. Material and Methods

2.1. Chemicals, reagents and materials

The 1200 ITU mg^{-1} VectoBac® 12AS commercial formulated Bti solution containing *Bacillus thuringiensis*, subsp. *israelensis*, strain AM 65-52 (Valent BioSciences, Libertyville, IL, U.S.) was supplied by the "Entente Interdépartementale pour la Démoustication du littoral Méditerranéen" (EID Méditerranée, Montpellier, France). For sample preparation: Methanol (MeOH) HPLC grade, Acetonitrile (ACN) HPLC grade, and Sodium Chloride (NaCl) ≥ 99.0 % ACS reagent were purchased from Sigma-Aldrich (Saint-Quentin-Fallavier, France). Magnesium Sulfate ($MgSO_4$) anhydrous RE – Pure – was purchased from CARLO ERBA (Val de Reuil, France). 50 mL Falcon® tubes were purchased from Fisher Scientific (Illkirch, France). 0.22 µm Polytetrafluoroethylene (PTFE) filters and 2 mL vials were purchased *via* Analytic Lab (Castelnau-le-Lez, France). For LC-HRMS analyses: Water (H_2O) for HPLC LC-MS grade was purchased from VWR International (Fontenay-sous-Bois, France). Acetonitrile (ACN) for LC/MS and Formic Acid (FA) for LC-MS were purchased from CARLO ERBA (Val de Reuil, France).

2.2. Sediment and salty water collection

Sediment and salty water sampling took place in a lagoon located at the Scamandre regional natural reserve (Aigues-Mortes, France). Sample collection was the same as reported by Salvia *et al.* [29]. In brief, samples were collected at three different points separated by 10 to 15 meters. They were then homogenized and stocked at 4 °C in dark. The collected sediments contained approximately 35 % of water, they had never been exposed to Bti. The sampling was conducted in collaboration with the EID Méditerranée.

2.3. Microcosms set-up

Ten identical microcosms were prepared in the laboratory as described by Salvia *et al.* [29]. Each microcosm was filled with 15 g of wet sediments (with 35 % of water) and 4.75 mL of salty water, resulting a composition of Sediment/Water 50:50 (w/w). For the spiked samples, 5.5 µL of aqueous commercial solution of the Bti insecticide were applied on five microcosms (5 biological replicates), and 5.5 µL of sterilized water were applied to the other five microcosms constituting the untreated control samples. Microcosms were then incubated in the dark (to exclude photolysis factor from the study) at 22 ± 2 °C for 15 days until the extraction

was performed. The water content in the microcosm was maintained through the addition of sterile water until the extraction was performed.

2.4. Meta-metabolome extraction

A QuEChERS-based protocol [30–32] was applied for sediments meta-metabolome extraction. This protocol is the same previously developed and applied for the EMF approach [29]. In brief, the microcosm contents (*i.e.* 15 g of wet sediment + 4.75 mL of salty water) were transferred to 50 mL Falcon® tubes, then 15 mL of ACN were added and the tubes were vigorously shaken for 10 s on vortex mixer. The salts (1 g NaCl + 4 g $MgSO_4$) were then added and the tubes were immediately shaken for 15 s manually, and swirled for 30 s using the vortex device. After, tubes were centrifuged for 10 min at 20 °C with a rotational speed of 4500 RPM, using an Allegra X-30R Centrifuge (Beckman Coulter, Brea, CA, U.S.). 13.5 mL of the ACN layer were collected and transferred to new 50 mL Falcon® tubes. Extraction solvent was then evaporated under vacuum at 30 °C until dry, using an EZ-2^{plus} evaporator (Genevac, Ipswich, U.K.). The dry residue was then re-dissolved in 1.5 mL of MeOH, mixed for 10 s using the vortex device, and transferred to 2 mL vials after filtration through 0.22 µm PTFE filters.

2.5. LC-HRMS acquisitions

LC-HRMS analyses were performed using a Vanquish UHPLC$^+$ Focused LC system equipped with an online degasser, a binary pump system, a temperature-controlled autosampler, and a column compartment (Thermo Fisher Scientific, Waltham, MA, U.S.), coupled with a maXis Electrospray-Quadrupole/Time-of-Flight (ESI-Q/ToF) mass spectrometer, equipped with a Nitrogen-supplied Hexapole CID collision cell (Bruker Daltonics, Bremen, Germany; Billerica, MA, U.S.).

2.5.1. LC conditions

HPLC column Kinetex® Polar C18 (particle size: 2.6 µm, pore size: 100 Å, length: 100 mm, internal diameter: 2.1 mm) from Phenomenex (Torrance, CA, U.S.) was selected to achieve compounds separation. It was chosen in order to widen the band of the analyzed metabolites (in term of polarity), as the modified C18 columns has proven their performance in improving the retention efficiency for polar and semi-polar metabolites [33,34]. The column was equipped with a Phenomenex SecurityGuard™ ULTRA UHPLC Polar C18 2.1 mm column guard cartridge for protection during analyses. The mobile phase consisted of two phases of elution solvents: phase A is H_2O + 0.1 % FA (v/v), and phase B is ACN + 0.1 % FA (v/v).

Chromatographic separation was conducted by applying a classic single-ramp gradient dedicated for untargeted metabolic screening. The applied gradient was the following: 5 % (B) during 2 min, from 5 % to 100 % (B) in 15 min, 100 % (B) during 3 min, from 100 % to 5 % (B) in 2 min and 5 % (B) during 4 min. The flow rate was set to 0.5 mL min^{-1}, column oven temperature was maintained at 30 °C. The injection volume was 5 µL. Only the elution from 2.45 min to 22.00 min was introduced into the mass spectrometer. The rest was diverted to waste in order to reduce source and system fouling.

2.5.2. HRMS conditions

For the ESI-Q/ToF conditions, two main MS segments were established: a Full HRMS acquisition segment and a calibration segment. The acquisition segment was set between 2.45 min and 22.00 min of Retention Time (RT). Scan range was set between *m/z* 80 and *m/z* 1600 with 0.8 Hz of spectra rate (1.25 s for 1 scan), corresponding to 8786 spectra summation. No rolling average was applied. Spectra acquisition was in both Profile (Continuum) and Line (Centroid). Maximum Intensity was used for Line Spectra Calculation. "Focus" mode was active. For the ESI source, a positive ion polarity mode was applied with a capillary voltage of 3.5 kV and an end plate offset of -0.5 kV. Nitrogen nebulizer pressure was set to 3 Bar, the dry gas flow to 10 L min^{-1}, and the dry heater temperature to 200 °C. For Funnels, Quadrupole and Collision Cell, tune parameters were optimized in order to favor the transfer of ions with *m/z* between 80 and 1600 to the ToF analyzer. All these parameters are detailed in Section A.II-A.1.1. (Appendix II-A). For the calibration segment (2.00 min to 2.45 min of RT), a pre-run internal mass-scale calibration was performed in High Precision Calibration (HPC) mode using Sodium Formate (NaF) calibration solution (0.05 % FA + 0.50 mL NaOH 1.0 M in 50 mL of H_2O/iPA 50:50 (v/v)) that was automatically infused at the beginning of each injection. All parameters were the same as for the Full HRMS acquisition segment, except for the nebulizer pressure that was set to 0.7 Bar (the suitable pressure for the syringe infusion flow rate of 3 µL min^{-1}), and for the dry gas flow that was adjusted to 4 L min^{-1} in order to reduce the de-clustering. For the ToF analyzer, voltage settings are described in details in Section A.II-A.1.2. (Appendix II-A). The resolution was experimentally assessed at the Full Width at Half Maximum (FWHM) of *m/z* peaks along the scan range. All values are shown in Table A.II-A 1 (Appendix II-A).

2.6. Analytical sequence

The analytical sequence for the untargeted metabolic profiling consisted of one LC-Q/ToF batch. All samples were injected randomly to minimize the effect of instrumental drifts. A pool QC was injected between every 2 sample injections in order to control and correct the potential "intra-batch" drifts. Samples were maintained at 20 °C in the LC temperature-controlled autosampler along the analytical sequence. The pool QC was prepared by mixing similar volume aliquots from all the 10 sediment extracts (*i.e.* the 5 replicates of control and the 5 replicates of pesticide-spiked sediments).

The dilution was performed by a random selection of 3 replicates from each sample group (control and spiked samples). They were diluted in MeOH with different dilution ratios: 1/2, 1/4, 1/6 and 1/10. The diluted samples were then immediately injected in the same LC-HRMS conditions, from the lower to the higher concentration, respectively.

2.7. LC-HRMS/MS acquisitions

LC-HRMS/MS acquisitions were performed using two different mass spectrometers: the maXis ESI-Q/ToF and a Q Exactive™ Plus Heated-Electrospray-Quadrupole/Orbitrap™ (HESI-Q/Orbitrap) Fourier Transform (FT) mass spectrometer, equipped with a C-trap and a Nitrogen-supplied octapole HCD collision cell (Thermo Fisher Scientific, Bremen, Germany; Waltham, MA, U.S.). Both spectrometers are coupled to Vanquish UHPLC^{+} Focused LC systems. LC conditions applied for the MS/MS acquisitions are the same described in Section 2.5.1.

2.7.1. MS/MS methods for the LC-Q/ToF MS

For MS/MS acquisitions performed using the maXis Q/ToF MS: the two segments (acquisition and calibration – previously described in Section 2.5.2.) were applied with the same associated RT ranges and ESI source parameters (ESI+). The scan range was set between *m/z* 40 and *m/z* 650 with 2.0 Hz of spectra rate (0.50 s for 1 scan), corresponding to 5596 spectra summation. No rolling average was applied. Spectra acquisition was in both Profile (Continuum) and Line (Centroid). Maximum Intensity was used for Line Spectra Calculation. "Focus" mode was active. Tune parameters were optimized in order to favor the transfer of ions with *m/z* between 40 and 650 to the ToF. They were experimentally optimized by monitoring NaF calibration clusters profile. Funnel 1 RF and Multipole RF were set to 400 Vpp. The in-source CID (isCID) energy was equal to 0.0 eV. The applied Quadrupole ion energy was fixed to 4.0 eV. The Quadrupole Low Mass was equal to *m/z* 200. For the collision cell, the Collision RF was set to

300 Vpp and the collision energy was set to 8.0 eV if no precursor is selected for fragmentation. The transfer time was equal to 50.0 µs, and the pre-pulse storage was equal to 5.0 µs. For the ToF analyzer, voltage settings are the same described in Section A.II-A.1.2. (Appendix II-A). The resolution was also experimentally assessed at the Full Width at Half Maximum (FWHM) of *m/z* peaks along the scan range. All values are shown in Table A.II-A 2 (Appendix II-A). For the mass-scale calibration, the same method described in Section 2.5.2. was applied after it was adjusted for the new scan range (*m/z* 40-650). To perform the fragmentations, targeted "Auto MS/MS" acquisitions were achieved following the scheduled precursors list described in Table A.II-A 3 (Appendix II-A). Separated runs with 4 different collision energies were performed on samples belonging to the two studied environmental conditions (Control and Spiked). The applied CID energies are shown in Table A.II-A 3 (Appendix II-A). MS/MS experiments succeeded for only 3 of the selected features. For a last remaining feature, no fragments could be detected. Thus, new acquisitions using LC-Q/Orbitrap were performed.

2.7.2. MS/MS methods for the LC-Q/Orbitrap FT-MS

For MS/MS acquisitions performed using the Q Exactive™ Plus Q/Orbitrap FT-MS: a single acquisition segment was set between 2.45 min and 22.00 min of RT. First, a Full MS experiment in ESI+ mode was performed for a general screening. It aimed to acquire more precise exact *m/z* measures and isotope patterns (at a higher resolution), and to define the RT ranges of the targeted metabolites. After, an inclusion list (Table A.II-A 4 – Appendix II-A) was constructed for Parallel Reaction Monitoring (PRM) experiments (targeted MS/MS). Then, for additional investigations, a Full MS experiment in ESI– mode was performed to search for potential negatively-charged species pertaining to the targeted metabolites. After detecting negative ions for 3 out of 4 targeted metabolites, a new inclusion list (Table A.II-A 4 – Appendix II-A) was constructed for ESI– PRM acquisitions. Properties for both ESI+ and ESI– Full MS experiments were the same: the scan range was set between *m/z* 80 and *m/z* 1200. The resolution was equal to 140000 (at *m/z* 200), the Automatic Gain Control (AGC) target was set to 3e6 ions, and the Maximum Injection Time (IT) was set to 200 ms. For both ESI+ and ESI–, PRM properties were the following: the resolution was equal to 35000 (at *m/z* 200), the AGC target was set to 5e5 ions, the Maximum IT was set to 150 ms, the isolation window was equal to 0.50 *m/z*, and the fixed first mass was set to *m/z* 50. For HCD fragmentations, 3 separated runs with 3 different collision energies were performed on samples belonging to the two studied environmental conditions (Control and Spiked). The applied HCD energies are shown in Table A.II-A 4 (Appendix II-A). For all Full MS and PRM experiments in ESI+ mode, HESI source

tunes were as the following: Sheath gas flow rate was set to 35 a.u., Auxiliary gas flow rate was set to 25 a.u., Sweep gas flow rate was set to 2 a.u., capillary temperature was equal to 360 °C, the Aux gas heater temperature was equal to 200 °C, the spray voltage was equal to 3.2 |kV|, and the S-lens RF level was equal to 50.0. In ESI– mode, all parameters were the same as for ESI+, except for the Auxiliary gas flow rate that was set to 10 a.u., and the capillary temperature that was set to 320 °C.

2.8. Software and data processing

For the maXis Q/ToF MS: LC system, LC-MS hyphenation and analytical sequence piloting were performed using HyStar 3.2.49.4 (Bruker Daltonics). Q/ToF piloting and LC-HRMS/MS data acquisition were performed using otofControl 4.0.97.4560 (Bruker Daltonics). Raw LC-HRMS/MS data were acquired in ".d" folder format. A new post-acquisition mass-scale internal calibration (HPC mode) was performed using Compass DataAnalysis 4.3 (Bruker Daltonics), and then ".d" LC-HRMS files were converted to NetCDF using the same software in order to upload and process them using Galaxy Workflow4Metabolomics platform [35–37]. All the NetCDF files are published on the European Bioinformatics Institute (EMBL-EBI, Hinxton, U.K.) MetaboLights platform [38,39]. The preprocessing workflow and all its parameters are published on the Galaxy Workflow4Metabolomics platform [40]. The "XCMS" algorithm-based preprocessing [41] consisted of a "centWave" peak piking [42], "PeakDensity" peak grouping, loess/non-linear "PeakGroups" retention time adjustment (degree of smoothing: 0.8), peak filling and "CAMERA" peak annotation [43]. The considered signal value for ion features was the chromatographic peak area. After, an "intra-batch" signal correction was applied using the "Batch correction" function with a "loess" regression model [44] (0.8 of span), followed by a matrix cleanup according to feature's CV in pool QC injections (all features with area RSD upper than 30 % through pool QC injections were eliminated from the dataset) [3]. After generating the data matrix, statistical analyses were performed using the R-based MetaboAnalyst platform [45–47]. All applied command lines for data processing and statistical analyses are shown in Section A.II-A.4.1. (Appendix II-A). Compass DataAnalysis 4.3 (Bruker Daltonics) was used for manual raw LC-HRMS and MS/MS data processing. Welch Two Sample t-Test for independent means comparison was performed using the R Commander 2.4-2 "Rcmdr" package [48] of R 3.3.3 software. The command lines are shown in Section A.II-A.4.2. (Appendix II-A).

For the Q Exactive™ Plus Q/Orbitrap FT-MS: LC piloting, LC-MS hyphenation, analytical sequence piloting and LC-HRMS/MS data acquisitions were performed using Xcalibur 4.1.31.9 (Thermo Fisher Scientific). Data were acquired in RAW format. The Mass Spectrometer and the HESI source were configured using Q Exactive Plus – Orbitrap MS 2.9 build 2926 software (Thermo Fisher Scientific). RAW data were explored using Xcalibur 4.1.31.9 and FreeStyle 1.3 (Thermo Fisher Scientific).

For each validated biomarker candidate (Section 3.4.), available MS/MS spectra acquired from the two mass spectrometers at the different collision energies and ionization modes were manually transformed to ".ms" format for putative identifications using SIRIUS 4.4.29 software [49]. Data transformation is detailed in Section A.II-B.2., Figure A.II-B 1 and Figure A.II-B 2 (Appendix II-B). The identifications through molecular databases were based on two main criteria: i) the elemental composition determined using the exact *m/z*, the adduct type and the isotope patterns, and ii) the computational structural elucidation based on the MS/MS fragments detected at different collision energies. The tolerated *m/z* deviation for fragments was set to 10 ppm for the Q/ToF and 5 ppm for the Q/Orbitrap, as the resolution of the Orbitrap is relatively higher for small *m/z*. The database search and identifications were conducted using the SIRIUS 4.4.29 software and its CSI:FingerID feature [50]. The retained putative annotations were the propositions presenting the "first rank" score and pertaining to natural products.

3. Results and Discussions

Results will be presented and discussed following the chronological order of data analysis that led to doubtful observations during "biomarkers" mining. The proposition of biomarkers validation method will be then exhibited.

3.1. Multivariate statistical analysis

The generated data matrix was uploaded to MetaboAnalyst platform for statistical data analysis and visualization. It consists of 16 observations/injections (5 control samples, 5 spiked samples, and 6 QC) and 1091 variables/ion features. A scaling step was applied on dataset prior to multivariate analyses, using a Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable) [51].

The descriptive unsupervised Principal Component Analysis (PCA) [52] is first applied (Figure C.II 2A). It shows a significant difference between metabolic profiles of the studied conditions (Control (Ctr) *vs.* Spiked (Bti) sediments). This difference is represented by the complete separation of the two sample clusters according to the first principal component PC1 (explaining 83.2 % of variations). The discrimination is explained by the PCA loadings plot (Figure C.II 2B). A massive cloud of features (red ellipse) was concentrated on spiked samples' side (the left side of the PC1). Those features are thus more abundant in spiked samples, which proves the higher complexity of spiked sediments' metabolic profiles. After checking their boxplots and abundance histograms, features of the massive cloud were only present in spiked sediments and were not detected in untreated control samples. Thus, they were mainly considered as traces of xenometabolites, or potential extremely overexpressed endometabolites that are not detected in the control samples. Then, for in-depth mining of biomarkers, the explicative supervised Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) [53,54] was applied after excluding QC samples from the dataset. Results (Figure C.II 2C) show that the predictive component (p) of the OPLS-DA explained 83.2 % of variations (the same percentage observed for the PC1 of the PCA). This is mainly caused by the dominance of the massive number of features that are only present in spiked samples. In fact, these features were revealed by the S-Plot of the OPLS-DA (the red ellipse in Figure C.II 2D). The S-Plot application was also able to highlight other relevant discriminant features that were hard to detect using PCA loadings plot. Two zones of importance were outlined in the S-Plot: the dark green ellipse highlighting biomarker candidates with an overexpression in control samples, and the light green ellipse highlighting biomarker candidates with an overexpression

in spiked samples (Figure C.II 2D). S-Plot dimensions are explained in Figure A.II-A 2 (Appendix II-A).

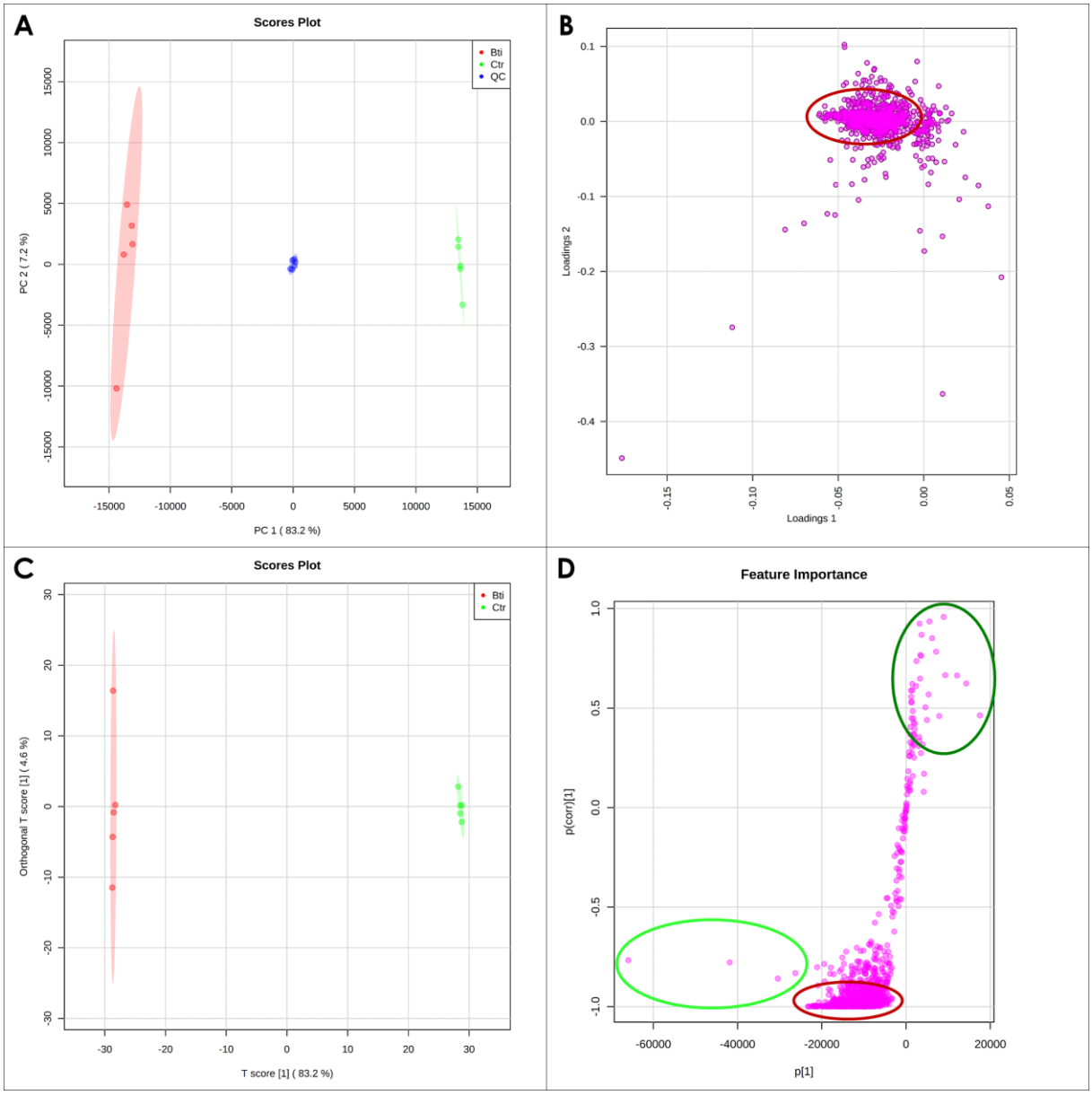


Figure C.II 2: The applied multivariate statistical analyses.
A: PCA, **B:** PCA loadings plot, **C:** OPLS-DA, **D:** OPLS-DA S-Plot.
PCA and OPLS-DA color codes: red for Spiked samples (Bti), green for Control samples (Ctr), blue for QC.
Red ellipses in B and D highlight the massive cloud of ion features potentially originating from Bti xenometabolome (or extremely overexpressed endometabolites).
The dark green ellipse in D highlights biomarker candidates overexpressed in Ctr.
The light green ellipse in D highlights biomarker candidates overexpressed in Bti.
Plots generated using MetaboAnalyst.

3.2. Biomarker candidates and raw data crosscheck

After examining the S-Plot, several features were considered as biomarker candidates. 14 features corresponding to 10 metabolites were finally selected for further investigations. They were selected according to the following criteria: i) they are the most discriminant features between the two compared groups (Spiked *vs.* Control) according to S-Plot, ii) they are endometabolites, as they are detected in both spiked and control samples. Features are summarized in Table C.II 1.

Table C.II 1: Biomarker candidates.
Features are sorted according to the descending order of their t-Test -Log_{10}[p-Val] applied on automatically processed data, from the top to the bottom. The t-Test with an unequal variance assumption was performed using MetaboAnalyst.
Compound formula prediction based on exact *m/z* and isotope pattern was performed using FreeStyle 1.3.
†: Feature code represents the ion nominal *m/z* preceded by "M", and ion's retention time (in seconds) preceded by "T".
‡: Adducts are annotated following the proposal of Damont *et al.* 2019 [19].
a, b, c: Features belonging to a same metabolite (based on "CAMERA" annotations).

Feature code†	**-Log10[p-Val] (t-Test)**	**RT (min)**	**Adduct type‡**	**Predicted Formula (M)**	**Theoretical exact *m/z***	**Experimental exact *m/z***	**Error (ppm)**	**Isotope Pattern Coverage (%)**
Features overexpressed in control sediments								
M461T812[a]	***4.14	13.54	$[M+H]^+$	$C_{30}H_{36}O_4$	461.26864	461.26784	-1.72	100.00
M314T729	**2.95	12.15	$[M+H]^+$	$C_{20}H_{43}NO$	314.34174	314.34104	-2.22	99.70
M483T812[a]	**2.74	13.54	$[M+Na]^+$	$C_{30}H_{36}O_4$	483.25058	483.24988	-1.46	100.00
M409T541[b]	**2.13	9.01	$[M+Na]^+$	$C_{22}H_{26}O_6$	409.16216	409.16180	-0.89	100.00
M410T541[b]	**2.01	9.01	$[M+Na]^+\_^{13}C$	$C_{22}H_{26}O_6$	410.16557	410.16505	-1.27	N/A
M304T652	*1.98	10.86	$[M+H]^+$	$C_{21}H_{37}N$	304.29988	304.29945	-1.41	99.57
M254T754	*1.45	12.57	$[M+H]^+$	$C_{16}H_{31}NO$	254.24784	254.24722	-2.43	99.69
M425T541[b]	*1.37	9.01	$[M+K]^+$	$C_{22}H_{26}O_6$	425.13610	425.13554	-1.31	99.54
M228T731	*1.27	12.19	$[M+H]^+$	$C_{14}H_{29}NO$	228.23219	228.23174	-1.97	99.53
M274T527	*1.06	8.78	$[M+H]^+$	$C_{16}H_{35}NO_2$	274.27406	274.27377	-1.06	98.50
Features overexpressed in spiked sediments								
M622T896	**2.84	14.93	$[M+H]^+$	$C_{37}H_{68}N_2O_5$	621.52010	621.51907	-1.65	99.90
M578T743	**2.56	12.38	$[M+H]^+$	$C_{35}H_{63}NO_5$	578.47790	578.47649	-2.43	99.13
M625T939[c]	**2.09	15.65	$[M+H]^+\_^{13}C$	$C_{37}H_{70}N_2O_5$	624.53918	624.53807	-1.78	N/A
M624T939[c]	**2.02	15.65	$[M+H]^+$	$C_{37}H_{70}N_2O_5$	623.53575	623.53474	-1.62	100.00

In order to crosscheck results, a manual exploring of the raw data is conducted. It aims to validate the results of the relative quantifications generated by the automated processing workflow. Indeed, information mining in the large and complex metabolomics data is extremely difficult without computational tools. Automated data processing and statistical analysis are essential for prioritizing the information. However, after filtering and prioritizing the information, raw data crosscheck must be considered to assess the quality of results. It allows detecting and/or avoiding potential errors and artefacts that can occur during the automated processing. These errors are not easily detectable; they can potentially lead to alter the subsequent experimental or processing steps, and preliminary or final conclusions as well.

Hence, raw data crosscheck is conducted by performing peak area integration of EICs of experimental exact *m/z* (± 0.0050) for each of the selected candidates, through all the analyzed samples. Results are shown in Figure C.II 3. Features' relative abundances are concordant with the results generated by the automated processing workflow (shown in Figure A.II-A 3 – Appendix II-A). A difference in degrees of significance is noticed, which is potentially related to the difference in peak integration algorithms (automated *vs.* manual). Only one feature (M304T652) was excluded at this stage, as its variation between groups was considered insignificant after raw data crosscheck.

Figure C.II 3: Boxplots showing the relative abundances of the selected biomarker candidates, generated by the manual processing of raw data.
Vertical axis represents the EIC peak area of experimental exact *m/z* of each feature. The green refers to abundances in control samples, red is for abundances in spiked samples.
Plots are sorted according to the descending order of features' scores of significance ($-\text{Log}_{10}$[p-Val] of Welch Two Sample t-Test applied on manually processed data), from the left to the right, and then from the top to the bottom. The first three rows represent plots of features overexpressed in control samples, the last row is for plots of features overexpressed in spiked samples.

3.3. Mass spectra: biomarkers or suppressed ions?

Mass Spectrometry detectors consist of complex systems combining several physical phenomena, chemical reactions and programed events, acting synergically. The large and complex datasets provided by MS are critical indicators for the quality of the acquired data. They can reveal several fundamental/instrumental-related artefacts and problems that influence results, as described in Section A.II-A.2. (Appendix II-A). Hence, the critical check of these MS data, with an awareness of fundamental aspects of the technique was considered for in-depth interpretation of results.

For each of the selected biomarker candidates, mass spectra were explored in the two sample conditions. Here, critical observations have been noticed for all features overexpressed in control sediments and for one feature (M578T743) overexpressed in spiked samples. As shown in Figure C.II 4, mass spectra in spiked samples reveal the co-elution of the selected metabolites with large multi-charged macromolecules, presenting massive isotope patterns of multi-charged ions (zoom in Figure C.II 5). These macromolecules are *a priori* polymers originating from the formulation agents of the pesticide (constituting 88.39 % of the VectoBac® 12AS commercial solution). According to the available literature, they are hypothetically federally-approved inert ingredients (*e.g.* emulsifiers, solvents, carriers) [55–57] based on biodegradable polymers (*e.g.* alginate, starch, cellulose, proteins) [58]. Their massive patterns explain the massive cloud of xenometabolites features observed previously in PCA loadings plot (Figure C.II 2B) and OPLS-DA S-Plot (Figure C.II 2D). These multi-charged ions and their isotope patterns generate very important redundancies in the data matrix, explaining the high number of ion features (1091 variables) in the dataset. This fact also explains the significant clustering of the two sample groups in the PCA, and the tendency of the OPLS-DA model to the overfitting (Figure A.II-A 4 – Appendix II-A).

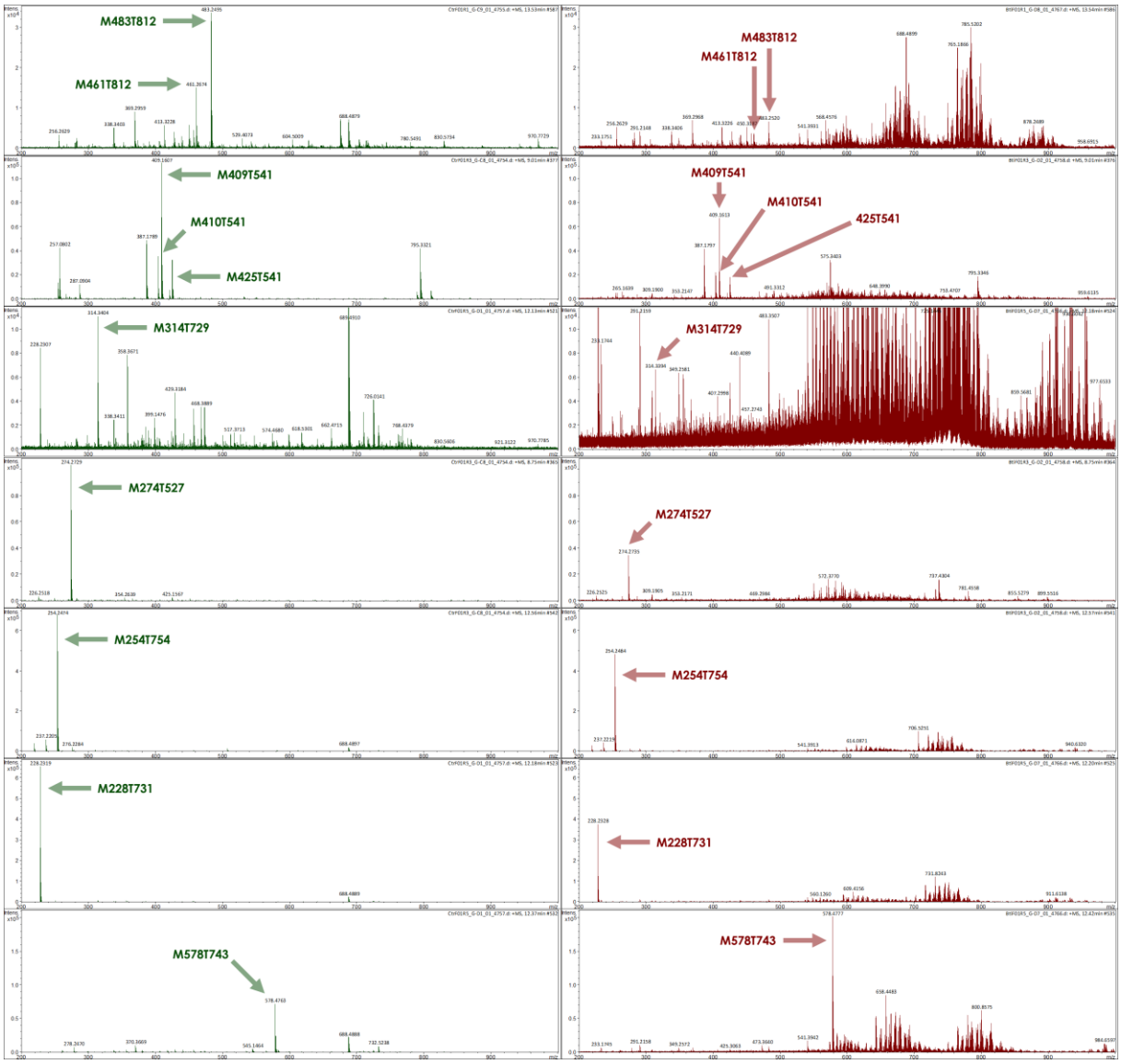


Figure C.II 4: Spectra in spiked samples (red) show the co-elution of the metabolites with multi-charged macromolecules, represented by the massive isotope patterns.
For spectra of each metabolite, intensity scale is fixed to the same value in both conditions (control in green and spiked in red). Spectra range zoom: *m/z* 200 to *m/z* 1000.
Spectra generated using Compass DataAnalysis 4.3 software.

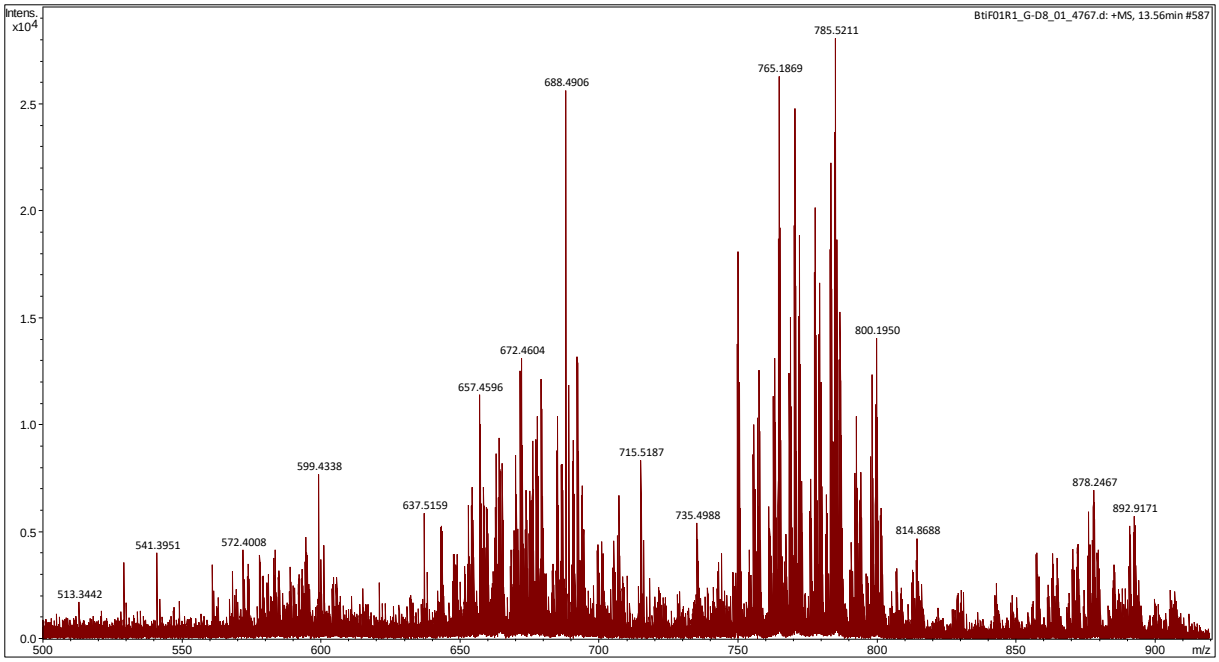


Figure C.II 5: Zoom on the massive isotope patterns of the co-eluting multi-charged macromolecules (spectrum range: *m/z* 500 to *m/z* 950).
High number of charges can be estimated since the high resolving power of the ToF MS (Table A.II-A 1 – Appendix II-A) was not sufficient to separate the recovering multi-charged isotope peaks. In conclusion, a large number of Carbons and Heteroatoms (*e.g.* Oxygen, Nitrogen) can be estimated, despite the complexity of the patterns.
Spectrum generated using Compass DataAnalysis 4.3 software.

On the other hand, concerning the main question regarding biomarkers determination; co-elutions with such macromolecules clearly put in doubt the reliability of the selected metabolites as "biomarkers". This is due to the potential ion suppression phenomenon occurring in the spiked samples. Such occurrence causes a decrease in metabolites signals and provokes a false variation that is not related to the studied biological factor (the impact of the pesticide on sediments microbiome). It is however related to a MS detection fundamental-instrumental factor. Indeed, as these relatively-small metabolites are co-eluting with multi-charged macromolecules, the hypothesis of ion suppression occurrence is highly supported by the mechanisms explaining this phenomenon [24,26] (previously summarized in Section 1.). Mainly, it can be related to the precipitation of small molecules with nonvolatile macromolecules, and the competition to access the available charge. Literature provides further evidences to support this hypothesis. In fact, it is frequently reported that molecules with higher mass usually suppress the signal of smaller molecules [25,26,59,60]. In addition, according to Furey *et al.* [26], polymers and carbohydrates can act as ion suppressors. Several previous works also reported the ion suppression phenomenon issued from the co-elution of analytes of interest with polymers-based drug formulation agents (excipients). Those studies were in the

framework of targeted LC-MS/MS quantitative analysis for drugs development [61–63]. False positives and false negatives were reported.

It is worth to mention that a contaminant compound (detected in both samples and blank injections) was revealed co-eluting with multi-charged macromolecules (Figure A.II-A 5 – Appendix II-A). It shows significantly higher abundances in control samples (Figure A.II-A 6 – Appendix II-A). Therefore, this compound named "M457T675" will be considered for the examination in the subsequent processing steps, as it may present an additional indicator for ion suppression occurrence.

3.4. Dilution-based approach for "biomarkers" validation

To filtrate reliable biomarker candidates and eliminate irrelevant suppressed ions, ion suppression influence on discrimination between groups should be reduced. Several strategies to overcome or reduce this issue were widely reported and described in the literature [24–26]. Their applicability for the current study was assessed. For instance, sample cleanup is not a suitable approach for untargeted metabolomics. It can engender a loss of potentially-relevant metabolic information. Solutions based on modifying LC and/or MS parameters can be constraint to robustness issues and the requirements of high throughput analysis (*e.g.* simple and reproducible protocols, fast analyses, "wide-band" information collection). Also, untargeted approaches require compromising these parameters, as the "optimal" conditions are complex and difficult to be defined in such contexts (conversely to the targeted approaches, where the quantification of targeted analytes is the base to define "optimum"). On the other hand, strategies based on reference standards or stable isotope-labeled standards [64] are non-practical for the current case. This is due to the lack of information allowing the simple identification of sediments' endogenous metabolites, as metabolome databases for environmental matrices (*e.g.* soil, sediment, sludge) are restrained. The application of "generic" internal standards (IS), *i.e.* compounds considered analogues to known biomolecules or families of biomolecules, can help to assess matrix effect in samples. However, this action still present several drawbacks. In fact, metabolites response in ESI-MS can differ between different molecules or families of molecules, and between the different LC elution ranges. For instance, if the IS is suppressed conversely to other metabolites, false positives can occur if IS-based post-acquisition normalization of concentrations is performed [16].

Finally, the adopted method was the dilution approach. Despite its drawbacks for traces analysis, this strategy was widely reported and recommended as "straightforward" method

[24–26]. Indeed, the dilution of sample extracts reduces the matrix effect, as it decreases the concentration of all matrix components. This enhances the efficiency of LC separation, and limits the occurrence of noise, interferences, and ion suppression in MS. Therefore, the approach was considered applicable and optimal for the current study. As described in Section 2.6., three replicates were randomly selected for dilution from each group of samples. They were diluted in MeOH at 1/2, 1/4, 1/6 and 1/10 of dilution ratios, and then analyzed with the same LC-HRMS method (described in Section 2.5.). Diluted samples were immediately injected in order to avoid any potential sample stability concerns at different dilution levels over the time. Original samples were also re-analyzed in the same analytical sequence to assure higher reproducibility and reliable comparisons.

After LC-HRMS data acquisition, EICs of experimental *m/z* of the selected features (including the contaminant M457T675) were integrated in all original and diluted samples. EIC area integrations were performed within an *m/z* window of ± 0.0050. Means and standard deviations (SD) of peak areas were calculated. Then, in order to enhance data visualization, the following equations (1) and (2) were applied to means and SDs, respectively:

$$(1): \overline{X_a} = \overline{X_r} \times f$$

$$(2): S_a = S_r \times f$$

Where "f" is the factor of dilution (*i.e.* 1/dilution ratio), "$\overline{X_r}$" and "S_r" are respectively the real means and SDs of EIC peak areas, and "$\overline{X_a}$" and "S_a" are respectively the "apparent" means and SDs calculated to visualize dilution curves.

Potential aberrations that can affect EIC area integrations, mainly related to shifts in *m/z* measuring through dilutions were assessed before highlighting dilution profiles. This by examining the effect of charge quantity-in-space and the saturation of ToF's Secondary Electron Multiplier detector. As a result, these aberrations were not significantly observed and did not affect the EIC area integrations through dilutions. Assessments and results are detailed in Section A.II-A.3. of the Appendix II-A.

For features overexpressed in control samples: M274T527 is the only feature showing a significant difference between groups after ten-time dilution (Figure C.II 6). Curves of its peak areas' "apparent means" ($\overline{X_a}$) do not show any significant increase following the increase of dilution factor in both control and spiked samples (slopes ≈ 0). Thus, this metabolite does not undergo a significant ion suppression, as its signal is not affected by the dilution. It is thus

validated as a reliable biomarker candidate. However, for the 5 remaining metabolites and their occurrences, the signal difference between groups is no more significant after ten-time dilution (2 of the metabolites and their occurrences are shown in Figure C.II 6). Curves show an enhancement of signal following the increase of dilution factor in spiked samples (slopes > 0). This means that these metabolites were undergoing an ion suppression in spiked samples. The decrease of their abundances in the spiked group is thus related to the ion suppression phenomenon and not to a biological factor. In addition, results reveal that the influence of matrix effect on ion signals was more important in spiked samples, when compared to control samples. This demonstrates the "group-to-group" heterogeneous matrix effect, due to the higher complexity of spiked samples' meta-metabolome, containing the complex xenometabolites of the Bti pesticide. This observation is also noticed for features overexpressed in the spiked group as shown in Figure C.II 7. The signal of these metabolites was enhanced in spiked samples and the difference between groups was increased at the 1/2 and 1/4 dilutions. Thus, the important matrix effect in spiked samples is not only provoking false positives, but could also engender potential false negatives by an exclusive signal extinction and loss in significance of potentially-reliable biomarker candidates that are overexpressed in spiked samples.

It is worth mentioning that ion suppression occurrence is also proved by the dilution profile of the M457T675 contaminant (Figure A.II-A 8 – Appendix II-A). The applied dilution does not affect the concentration of this compound originating from the organic elution solvent (ACN). However, its signal was significantly enhanced in spiked samples by the dilution, which led to lose the variation significance between the two groups. This presents an additional evidence on the occurrence of a heterogeneous "group-to-group" ion suppression and the potential provocation of false positives.

For additional validations regarding the 4 features proven as valid biomarker candidates, putative identifications were performed. The MS/MS acquisitions were performed using ESI+ Q/ToF, ESI+ Q/Orbitrap, and ESI– Q/Orbitrap Full HRMS and MS/MS experiments. All the related workflows, methods and processing are discussed in details in Sections A.II-B.1. and A.II-B.2. (Appendix II-B). The features were putatively identified at the level "2a" of identification confidence (according to the scale defined by Schymanski *et al.* [65]). Results are shown in Table C.II 2. The putative identifications correspond to compounds that can be originating from nature. For the 3 features overexpressed in the spiked samples (*i.e.* M578T743, M622T896 and M624T939), the three different types of acquisitions were able to provide MS/MS spectral data. All these separated data led to the same putative annotations, *i.e.*, for

each feature, all the 3 different acquisition types proposed the same candidate. For the fourth feature overexpressed in control samples (*i.e.* M274T527), only ESI+ Q/Orbitrap experiments succeeded to provide MS/MS spectra. A putative annotation was thus made using these available MS/MS data.

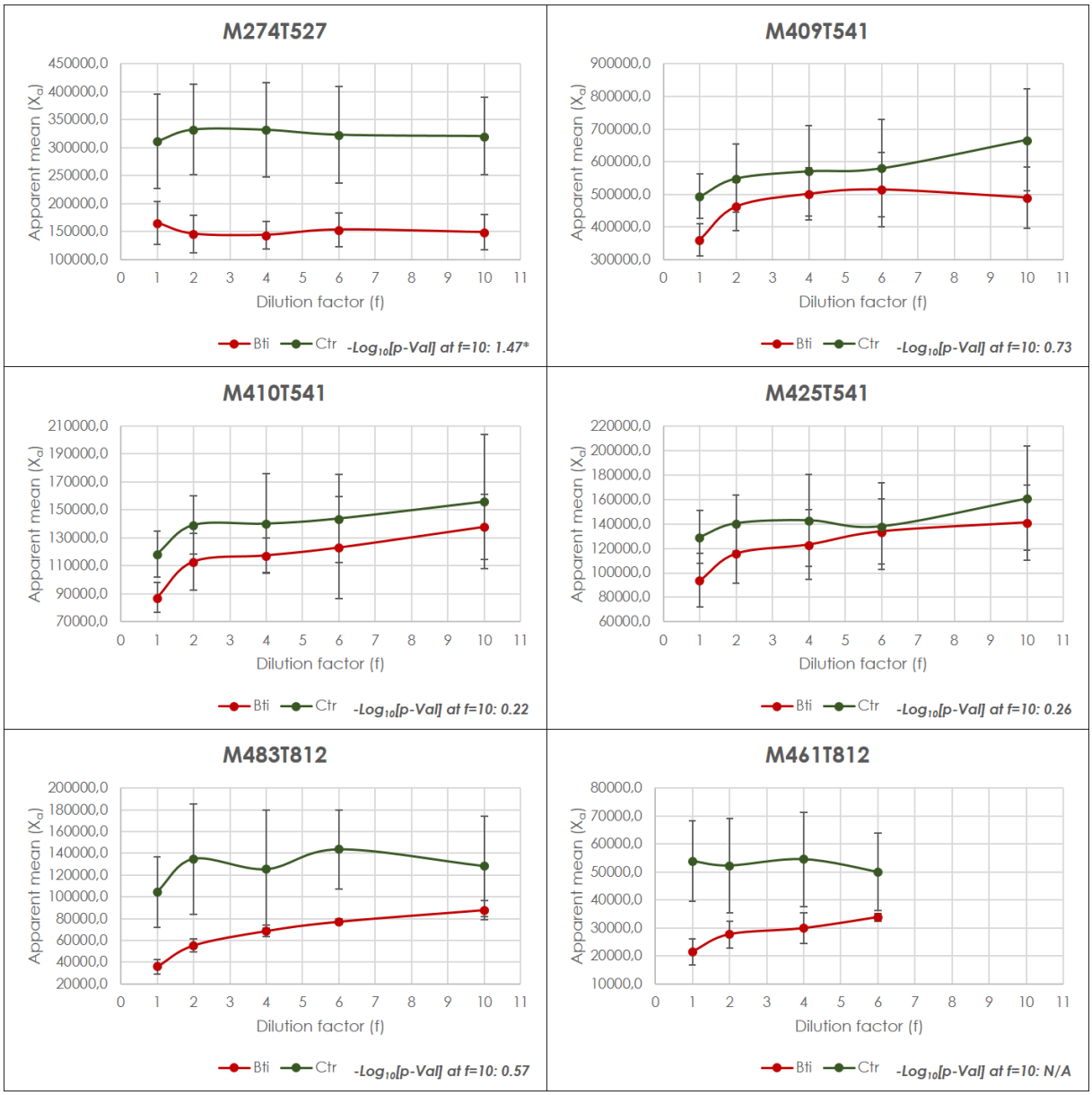


Figure C.II 6: "Apparent means" evolution versus dilution factor ($\overline{X_a} = \overline{X_r} \times f$) for features overexpressed in control samples.
Green curves represent the evolution of means in control samples (Ctr). Red curves are for spiked samples (Bti). M461T812 feature was below the detection limit at f = 10 (Means were not calculated).

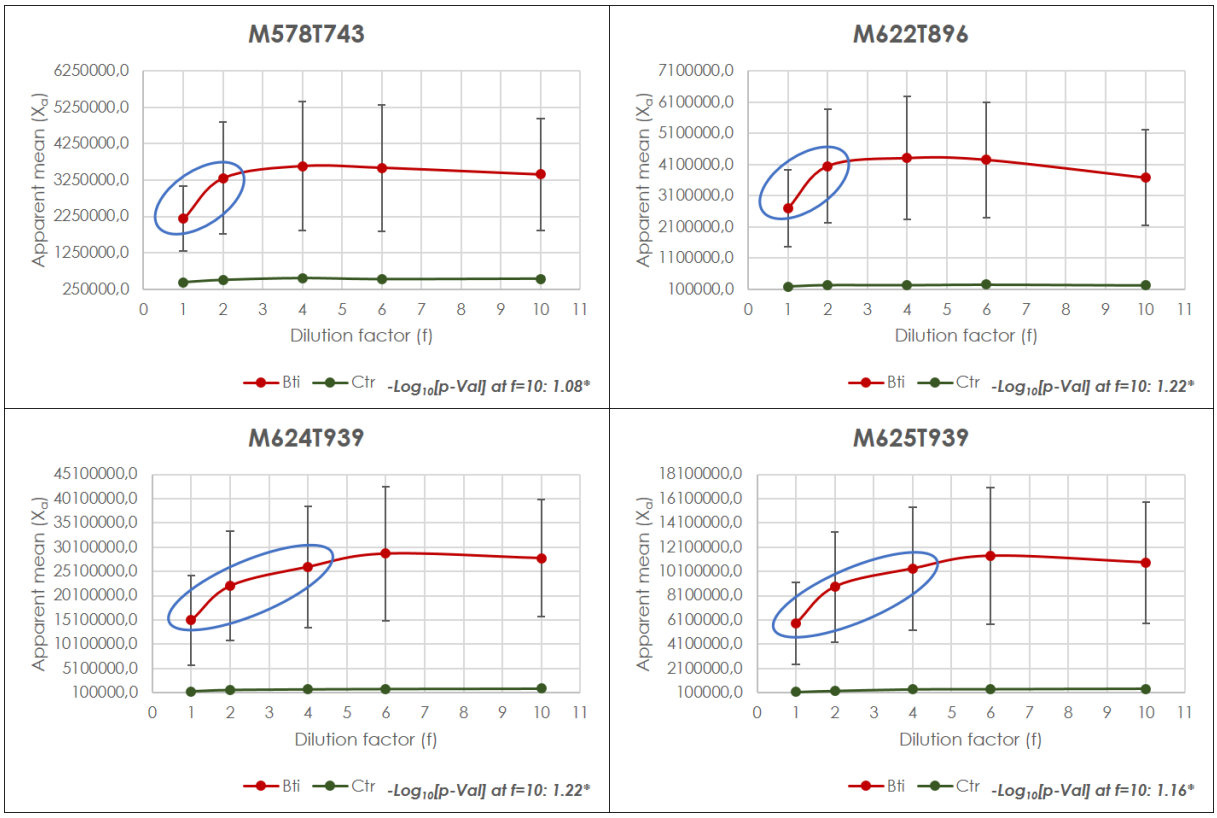


Figure C.II 7: "Apparent means" evolution versus dilution factor ($\overline{X_a} = \overline{X_r} \times f$) for features overexpressed in spiked samples.
Blue ellipses highlight the signal enhancement following the dilution in the spiked samples.
Green curves represent the evolution of means in control samples (Ctr). Red curves are for spiked samples (Bti).

Table C.II 2: Putative identifications of biomarker candidates.
N/A: Not Available. *: The "first rank" annotation that was proposed by the 3 different/separated spectral data was represented by a structural form only (presented in Figure A.II-A 9 – Appendix II-A). The database search that was performed on SIRIUS 4.4.29 qualified it as a "natural product". A manual database search was thus performed using Reaxys® [66]. It was performed in order to find names or information that could be related to the proposed structure (or the SMILES or the InChI™ Key). However, no name or information could be found, which demonstrates the difficulty in identifying metabolites in such environmental chemistry contexts.

Feature code	Name	InChI™ Key	Acquisition	CSI:FingerID Score	SIRIUS Similarity Score (%)
Biomarker candidates overexpressed in control sediments					
M274T527	Hexadecasphinganine	ZKLREJQHRKUJHD	ESI+ Q/Orbitrap	-101.249	50.673
Biomarker candidates overexpressed in spiked sediments					
M578T743	N/A*	VRQPOJLXNNVWEV	ESI+ Q/ToF	-360.882	33.187
			ESI+ Q/Orbitrap	-295.738	36.851
			ESI– Q/Orbitrap	-390.739	32.901
M622T896	[1-Oxo-1-[(2-oxocyclopropyl)amino]octadecan-3-yl] 6-(decanoylamino)hexanoate	NEDUMUADZBTZIB	ESI+ Q/ToF	-270.040	45.594
			ESI+ Q/Orbitrap	-271.118	46.541
			ESI– Q/Orbitrap	-324.782	43.281
M624T939	3-Hydroxy-2-[2-[[(E)-octadec-9-enoyl]amino]hexadecanoylamino]propanoic acid	NUFXBUXRBVOVLJ	ESI+ Q/ToF	-206.410	55.300
			ESI+ Q/Orbitrap	-177.645	63.157
			ESI– Q/Orbitrap	-274.166	46.334

3.5. Summary: a pragmatic approach to avoid ion suppression risks

The risk caused by the heterogeneous "group-to-group" ion suppression is experimentally proved. Therefore, a simple approach can be suggested in order to encounter such risks in (meta-)metabolomics. Particularly when studying xenobiotics and dealing with samples exhibiting heterogeneous complexities. The approach seeks for two ultimate objectives. The first is to help detecting such ion suppressions and thus avoiding biased decisions and conclusions caused by potential false positives/negatives. The second aim is to filtrate relevant biomarker candidates before processing in complex subsequent experiments, and particularly before processing to the mandatory step: the absolute quantification using stable isotope-labeled standards [8,9,64,67]. Such information prioritizing can be helpful to save time and resources. In fact, stable isotope-labeled standards are expensive, hardly accessible and mostly lacking when working in fields like environmental chemistry, marine biology, or plant secondary metabolome. This is due to the wide biochemical diversity and the restrained metabolome databases or MS spectral libraries for such fields. Metabolites annotation is relatively complicated in such studies. Compounds characterization is thus needed to identify novel relevant biomarkers, which is also complex and needs information to be prioritized.

Thus, the suggested approach is designed to be a pragmatic solution and easily accessible for metabolomics developers and users. It is summarized in Figure C.II 8.

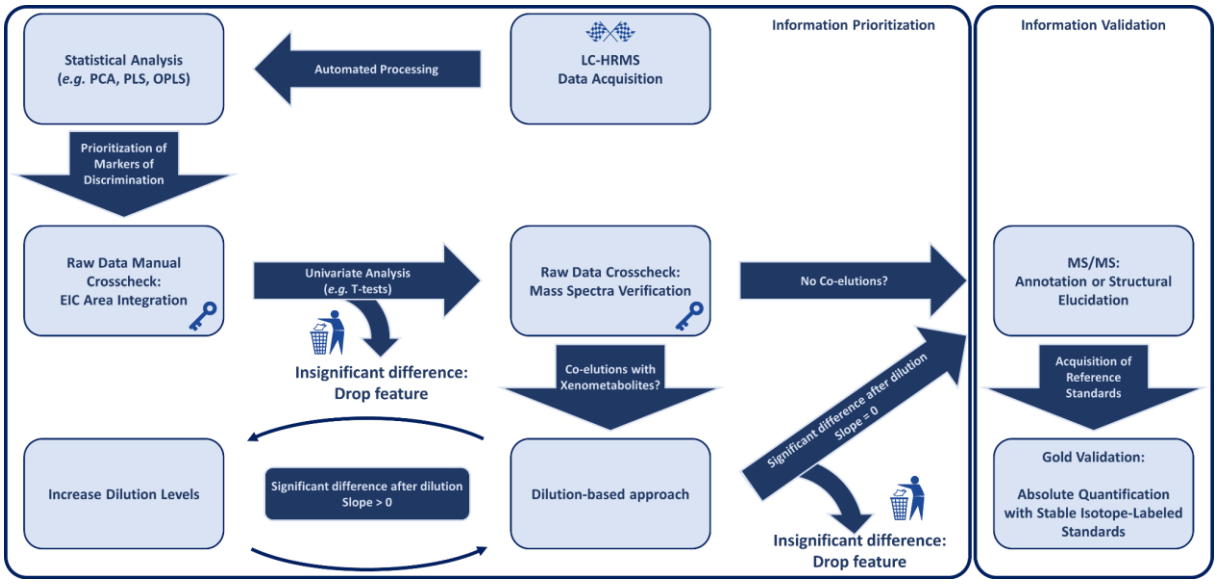


Figure C.II 8: The suggested approach; a pragmatic and easily accessible solution.
Key-steps of the approach are raw data manual crosscheck and mass spectra verifications. Raw data crosscheck allows to eliminate insignificant features after applying appropriate statistical tests (test type depends on sample size and the studied factor). Mass spectra verification allows to reveal potential MS-related artefacts (*e.g.* ion suppression).

For the dilution-based method, at least 5 levels of concentrations are recommended to construct reliable dilution curves. However, if the feature is still significant after dilutions but the slope of its curve is > 0, this means that the metabolite is undergoing an ion suppression and its significance should be examined after performing further dilutions until flattening the curve (slope = 0). Here, an advantage can be highlighted for the designed visualization model and its equations (Equations (1) and (2)). Indeed, it can anticipate ion suppression using slopes as indicators. This is due to the non-linear response of compounds undergoing ion suppression. For this same reason, Broadhurst *et al.* [17] did not recommended the technique using dilutions for pool QC, aiming to eliminate artefacts from metabolomics datasets.

It is worth mentioning that in some particular cases, the examination of a selected biomarker candidate can be considered critical by the investigators. Additional crosscheck steps can be hence performed to investigate ion suppression effects. For instance, punctual tests based on changing LC gradient or ESI polarity can be done [24–26]. For the current work, negative ESI mode was tested using the Q/ToF MS but neither the prioritized features, nor the multi-charged ions of macromolecules were detected.

4. Concluding Remarks

In the current work, the heterogeneous "group-to-group" ion suppression phenomenon was studied and experimentally proved. It should be taken into consideration as a serious potential problem for biomarkers determination, particularly, when dealing with samples exhibiting heterogeneous complexities. A simple workflow is thus suggested to encounter ion suppression risks on biomarkers determination in such cases. It is based on manual raw data crosscheck and mass spectra verification after prioritizing the discriminant features by the automated data processing. Then, a dilution-based approach and a dedicated data visualization model can be applied to reveal the suspected suppressed ions and to filtrate the significant metabolites before the complex subsequent experiments. The approach is easily accessible for metabolomics developers and users. It allows saving time and resources by prioritizing relevant information. It is a crucial step that must be performed before drawing conclusions and further investigations such as metabolites annotation/characterization and absolute quantification.

For the EMF approach, the *a priori* knowledge about the heterogeneous complexity between the different groups of samples (polluted *vs.* unpolluted) leads to highly expect such selective ion suppression occurrences (mainly in the polluted samples). Therefore, it is highly recommended that the environmental biomarkers validation procedure passes through the suggested verification approach in order to avoid the potential false positives. Once the biomarker candidates are selected after the multivariate and the univariate analyses are done, the chromatograms and the mass spectra should be verified. Then, in case of co-elutions between these biomarker candidates and xenometabolites, the dilution-based approach should be applied in order to determine if the intensity difference between the two groups is the result of a potential biological effect or due to the ion suppression.

To be noted that after the dilution is performed and the diluted samples are analyzed, it can be interesting to process their data in order to investigate if some new and different biomarker candidates appear. These new investigations applied on the diluted samples can potentially lead to the revelation of potential biomarkers that are overexpressed in the polluted group but were suppressed in the concentrated samples due to their co-elution with xenometabolites (*i.e.*, they were false negatives).

Acknowledgments

Authors would like to acknowledge Jeanine Almany (École Pratique des Hautes Études) for providing English language editing, and Dr. Jean-Philippe Antignac (Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement) for his constructive critics and comments regarding the manuscript. Authors also acknowledge two anonymous reviewers for their constructive critics that helped improving the current work.

Acknowledgments to collaborators from the EID Méditerranée, Christophe Lagneau and Benoit Frances, for supplying the commercial formulated Bti pesticide, and for their valuable help with water and sediment samplings. Acknowledgments to Alexandre Verdu, Engr. (Bruker Daltonics) for providing technical information and explanations regarding the maXis Q/ToF. Acknowledgments to Jennifer Sola, M.Sc. (Université de Perpignan Via Domitia) for her precious assistance during the experimental work.

This work was supported by the Marie Skłodowska-Curie Actions EnvFate project (Grant agreement ID: 746656, funded under H2020-EU.1.3.2.). Its post-doctoral fellowship grant was awarded to CP. The funding institution had no role in the experimental design, the data processing, or in writing and reviewing the manuscript.

Ph.D. fellowship grant was awarded to HG by the French Ministry of Higher Education, Research and Innovation (MESRI), *via* the Doctoral School ED305 "Energie et Environnement" (Université de Perpignan Via Domitia).

The LC-Q/ToF and LC-Q/Orbitrap methods developments and analyses had been performed using the Biodiversité et Biotechnologies Marines (Bio2Mar) facilities – Métabolites Secondaires Xénobiotiques Métabolomique Environnementale (MSXM) platform at the Université de Perpignan Via Domitia (http://bio2mar.obs-banyuls.fr/).

List of Abbreviations

a.u.: Arbitrary Unit

ACN: Acetonitrile

AGC: Automatic Gain Control

APCI: Atmospheric Pressure Chemical Ionization

BA: Biocontrol Agent

Bti: *Bacillus thuringiensis israelensis*

CID: Collision-Induced Dissociation

CV: Coefficient of Variation

EIC: Extracted Ion Chromatogram

EMF: Environmental Metabolic Footprinting

ESI: Electrospray Ionization

FA: Formic Acid

FT: Fourier Transform

FWHM: Full Width at Half Maximum

HCD: Higher-energy C-trap Dissociation

HESI: Heated-Electrospray Ionization

HPC: High Precision Calibration

HPLC: High Performance Liquid Chromatography

HRMS: High Resolution Mass Spectrometry

InChI: The IUPAC International Chemical Identifier

iPA: Isopropyl-Alcohol

IS: Internal Standard

isCID: in-source CID

IT: Injection Time

ITU: International Toxic Units

LC: Liquid Chromatography

MeOH: Methanol

MS/MS: Tandem Mass Spectrometry

MS: Mass Spectrometry

NaF: Sodium Formate

NetCDF: Network Common Data Form

OPLS-DA: Orthogonal Projections to Latent Structures Discriminant Analysis

PCA: Principal Component Analysis

ppm: Parts-per-Million

PRM: Parallel Reaction Monitoring

PTFE: Polytetrafluoroethylene

p-Val: p-Value

Q/ToF: Quadrupole/Time-of-Flight

QC: Quality Control

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

RF: Radio Frequency

RPM: Revolutions per Minute

RSD: Relative Standard Deviation

RT: Retention Time

S/N: Signal-to-Noise ratio

SD: Standard Deviation

SMILES: Simplified Molecular-Input Line-Entry System

UHPLC: Ultra High Performance Liquid Chromatography

Vpp: Peak-to-Peak Voltage

References

[1] M. Bedair, L.W. Sumner, Current and emerging mass-spectrometry technologies for metabolomics, TrAC Trends in Analytical Chemistry. 27 (2008) 238–250. https://doi.org/10.1016/j.trac.2008.01.006.

[2] Z. Lei, D.V. Huhman, L.W. Sumner, Mass Spectrometry Strategies in Metabolomics, J. Biol. Chem. 286 (2011) 25435–25442. https://doi.org/10.1074/jbc.R111.238691.

[3] A. Roux, D. Lison, C. Junot, J.-F. Heilier, Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review, Clinical Biochemistry. 44 (2011) 119–135. https://doi.org/10.1016/j.clinbiochem.2010.08.016.

[4] E.G. Armitage, C. Barbas, Metabolomics in cancer biomarker discovery: Current trends and future perspectives, Journal of Pharmaceutical and Biomedical Analysis. 87 (2014) 1–11. https://doi.org/10.1016/j.jpba.2013.08.041.

[5] L. Brennan, H. Gibbons, A. O'Gorman, An Overview of the Role of Metabolomics in the Identification of Dietary Biomarkers, Curr Nutr Rep. 4 (2015) 304–312. https://doi.org/10.1007/s13668-015-0139-1.

[6] J.-L. Wolfender, G. Glauser, J. Boccard, S. Rudaz, MS-based Plant Metabolomic Approaches for Biomarker Discovery, Natural Product Communications. 4 (2009) 1417–1430. https://doi.org/10.1177/1934578X0900401019.

[7] J.G. Bundy, M.P. Davey, M.R. Viant, Environmental metabolomics: a critical review and future perspectives, Metabolomics. 5 (2009) 3–21. https://doi.org/10.1007/s11306-008-0152-0.

[8] J. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Translational biomarker discovery in clinical metabolomics: an introductory tutorial, Metabolomics. 9 (2013) 280–299. https://doi.org/10.1007/s11306-012-0482-9.

[9] S.A. Goldansaz, A.C. Guo, T. Sajed, M.A. Steele, G.S. Plastow, D.S. Wishart, Livestock metabolomics and the livestock metabolome: A systematic review, PLoS ONE. 12 (2017) e0177675. https://doi.org/10.1371/journal.pone.0177675.

[10] J.-L. Ren, A.-H. Zhang, L. Kong, X.-J. Wang, Advances in mass spectrometry-based metabolomics for investigation of metabolites, RSC Adv. 8 (2018) 22335–22350. https://doi.org/10.1039/C8RA01574K.

[11] X. Zhang, Q. Li, Z. Xu, J. Dou, Mass spectrometry-based metabolomics in health and medical science: a systematic review, RSC Adv. 10 (2020) 3092–3104. https://doi.org/10.1039/C9RA08985C.

[12] P. Filzmoser, B. Walczak, What can go wrong at the data normalization step for identification of biomarkers?, Journal of Chromatography A. 1362 (2014) 194–205. https://doi.org/10.1016/j.chroma.2014.08.050.

[13] J. Boccard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Analytica Chimica Acta. 769 (2013) 30–39. https://doi.org/10.1016/j.aca.2013.01.022.

[14] J. Boccard, J.-L. Veuthey, S. Rudaz, Knowledge discovery in metabolomics: An overview of MS data handling, J. Sep. Sci. 33 (2010) 290–304. https://doi.org/10.1002/jssc.200900609.

[15] E.C. Considine, G. Thomas, A.L. Boulesteix, A.S. Khashan, L.C. Kenny, Critical review of reporting of the data analysis step in metabolomics, Metabolomics. 14 (2018) 7. https://doi.org/10.1007/s11306-017-1299-3.

[16] Y. Wu, L. Li, Sample normalization methods in quantitative metabolomics, Journal of Chromatography A. 1430 (2016) 80–95. https://doi.org/10.1016/j.chroma.2015.12.007.

[17] D. Broadhurst, R. Goodacre, S.N. Reinke, J. Kuligowski, I.D. Wilson, M.R. Lewis, W.B. Dunn, Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies, Metabolomics. 14 (2018) 72. https://doi.org/10.1007/s11306-018-1367-3.

[18] M.D. Luque de Castro, F. Priego-Capote, The analytical process to search for metabolomics biomarkers, Journal of Pharmaceutical and Biomedical Analysis. 147 (2018) 341–349. https://doi.org/10.1016/j.jpba.2017.06.073.

[19] A. Damont, M.-F. Olivier, A. Warnet, B. Lyan, E. Pujos-Guillot, E.L. Jamin, L. Debrauwer, S. Bernillon, C. Junot, J.-C. Tabet, F. Fenaille, Proposal for a chemically consistent way to annotate ions arising from the analysis of reference compounds under ESI conditions: A prerequisite to proper mass spectral database constitution in metabolomics, J Mass Spectrom. 54 (2019) 567–582. https://doi.org/10.1002/jms.4372.

[20] М. Александров, Л. Галль, Н. Краснов, В. Николаев, В. Павленко, В. Шкуров, Экстракция ионов из растворов при атмосферном давлении-новый метод масс-спектрометрического анализа биоорганических веществ, Дан СССР. 277 (1984) 379–383. http://mass-spektrometria.ru/download/2005(2)/T2N1.pdf (accessed October 18, 2020).

[21] M.L. Alexandrov, L.N. Gall, N.V. Krasnov, V.I. Nikolaev, V.A. Pavlenko, V.A. Shkurov, Extraction of ions from solutions under atmospheric pressure as a method for mass spectrometric analysis of bioorganic compounds, Rapid Communications in Mass Spectrometry. 22 (2008) 267–270. https://doi.org/10.1002/rcm.3113.

[22] V.G. Zaikin, A.A. Sysoev, Mass Spectrometry in Russia, Eur J Mass Spectrom (Chichester). 19 (2013) 399–452. https://doi.org/10.1255/ejms.1248.

[23] M. Yamashita, J.B. Fenn, Electrospray ion source. Another variation on the free-jet theme, J. Phys. Chem. 88 (1984) 4451–4459. https://doi.org/10.1021/j150664a002.

[24] J.-P. Antignac, K. de Wasch, F. Monteau, H. De Brabander, F. Andre, B. Le Bizec, The ion suppression phenomenon in liquid chromatography–mass spectrometry and its consequences in the field of residue analysis, Analytica Chimica Acta. 529 (2005) 129–136. https://doi.org/10.1016/j.aca.2004.08.055.

[25] F. Gosetti, E. Mazzucco, D. Zampieri, M.C. Gennaro, Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry, Journal of Chromatography A. 1217 (2010) 3929–3937. https://doi.org/10.1016/j.chroma.2009.11.060.

[26] A. Furey, M. Moriarty, V. Bane, B. Kinsella, M. Lehane, Ion suppression; A critical review on causes, evaluation, prevention and applications, Talanta. 115 (2013) 104–122. https://doi.org/10.1016/j.talanta.2013.03.048.

[27] R. King, R. Bonfiglio, C. Fernandez-Metzler, C. Miller-Stein, T. Olah, Mechanistic investigation of ionization suppression in electrospray ionization, J Am Soc Mass Spectrom. 11 (2000) 942–950. https://doi.org/10.1016/S1044-0305(00)00163-X.

[28] C. Patil, C. Calvayrac, Y. Zhou, S. Romdhane, M.-V. Salvia, J.-F. Cooper, F.E. Dayan, C. Bertrand, Environmental Metabolic Footprinting: A novel application to study the impact of a natural and a synthetic β-triketone herbicide in soil, Science of The Total Environment. 566–567 (2016) 552–558. https://doi.org/10.1016/j.scitotenv.2016.05.071.

[29] M.-V. Salvia, A. Ben Jrad, D. Raviglione, Y. Zhou, C. Bertrand, Environmental Metabolic Footprinting (EMF) vs. half-life: a new and integrative proxy for the discrimination between control and pesticides exposed sediments in order to further characterise pesticides' environmental impact, Environ Sci Pollut Res. 25 (2018) 29841–29847. https://doi.org/10.1007/s11356-017-9600-6.

[30] M. Anastassiades, S.J. Lehotay, D. Štajnbaher, F.J. Schenck, Fast and easy multiresidue method employing acetonitrile extraction/partitioning and "dispersive solid-phase extraction" for the determination of pesticide residues in produce, Journal of AOAC International. 86 (2003) 412–431. PMID: 12723926.

[31] S.J. Lehotay, Quick, Easy, Cheap, Effective, Rugged, and Safe Approach for Determining Pesticide Residues, in: J.L. Martínez Vidal, A.G. Frenich (Eds.), Pesticide Protocols, 1st ed., Humana Press, Totowa, NJ, 2006: pp. 239–261. https://doi.org/10.1385/1-59259-929-X:239.

[32] J.L. Martínez Vidal, A. Garrido Frenich, eds., Pesticide Protocols, 1st ed., Humana Press, Totowa, NJ, 2006. https://doi.org/10.1385/159259929X (accessed March 23, 2020).

[33] Z. Liu, S. Rochfort, A fast liquid chromatography–mass spectrometry (LC–MS) method for quantification of major polar metabolites in plants, Journal of Chromatography B. 912 (2013) 8–15. https://doi.org/10.1016/j.jchromb.2012.10.040.

[34] H. Ghosson, A. Schwarzenberg, F. Jamois, J.-C. Yvin, Simultaneous untargeted and targeted metabolomics profiling of underivatized primary metabolites in sulfur-deficient barley by ultra-high performance liquid chromatography-quadrupole/time-of-flight mass spectrometry, Plant Methods. 14 (2018) 62. https://doi.org/10.1186/s13007-018-0329-0.

[35] Galaxy Workflow4Metabolomics, Galaxy Workflow4Metabolomics. (n.d.). https://galaxy.workflow4metabolomics.org/ (accessed February 28, 2020).

[36] F. Giacomoni, G. Le Corguille, M. Monsoor, M. Landi, P. Pericard, M. Petera, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E.A. Thevenot, C. Caron, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics. 31 (2015) 1493–1495. https://doi.org/10.1093/bioinformatics/btu813.

[37] Y. Guitton, M. Tremblay-Franco, G. Le Corguillé, J.-F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E.A. Thévenot, Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics, The International Journal of Biochemistry & Cell Biology. 93 (2017) 89–101. https://doi.org/10.1016/j.biocel.2017.07.002.

[38] H. Ghosson, Y. Guitton, A. Ben Jrad, C. Patil, D. Raviglione, M.-V. Salvia, C. Bertrand, MTBLS1784: Electrospray Ionization and heterogeneous matrix effects in Liquid Chromatography-Mass Spectrometry-based meta-metabolomics: a biomarker or a suppressed ion?, MetaboLights. (2020). https://www.ebi.ac.uk/metabolights/MTBLS1784 (accessed June 10, 2020 – Status: In Curation).

[39] K. Haug, K. Cochrane, V.C. Nainala, M. Williams, J. Chang, K.V. Jayaseelan, C. O'Donovan, MetaboLights: a resource evolving in response to the needs of its scientific community, Nucleic Acids Research. 48 (2019) D440–D444. https://doi.org/10.1093/nar/gkz1019.

[40] H. Ghosson, Y. Guitton, A. Ben Jrad, C. Patil, D. Raviglione, M.-V. Salvia, C. Bertrand, Electrospray Ionization and heterogeneous matrix effects in Liquid Chromatography-Mass Spectrometry-based meta-metabolomics: a biomarker or a suppressed ion? - W4M Workflow and Parameters, Galaxy Workflow4Metabolomics. (2020). https://workflow4metabolomics.usegalaxy.fr/u/hikmatghosson/w/w4m1784 (accessed September 16, 2020).

[41] 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.

[42] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics. 9 (2008) 504. https://doi.org/10.1186/1471-2105-9-504.

[43] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets, Anal. Chem. 84 (2012) 283–289. https://doi.org/10.1021/ac202450g.

[44] F.M. van der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema, Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping, J. Proteome Res. 8 (2009) 5132–5141. https://doi.org/10.1021/pr900499r.

[45] MetaboAnalyst, MetaboAnalyst. (n.d.). https://www.metaboanalyst.ca/ (accessed February 28, 2020).

[46] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Research. 46 (2018) W486–W494. https://doi.org/10.1093/nar/gky310.

[47] J. Chong, D.S. Wishart, J. Xia, Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis, Current Protocols in Bioinformatics. 68 (2019) e86. https://doi.org/10.1002/cpbi.86.

[48] J. Fox, The R Commander: A Basic-Statistics Graphical User Interface to R, Journal of Statistical Software. 14 (2005) 1–42. https://doi.org/10.18637/jss.v014.i09.

[49] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A.V. Melnik, M. Meusel, P.C. Dorrestein, J. Rousu, S. Böcker, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information, Nat Methods. 16 (2019) 299–302. https://doi.org/10.1038/s41592-019-0344-8.

[50] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, Searching molecular structure databases with tandem mass spectra using CSI:FingerID, Proc Natl Acad Sci USA. 112 (2015) 12580–12585. https://doi.org/10.1073/pnas.1509788112.

[51] R.A. van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics. 7 (2006) 142. https://doi.org/10.1186/1471-2164-7-142.

[52] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2 (1901) 559–572. https://doi.org/10.1080/14786440109462720.

[53] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics. 16 (2002) 119–128. https://doi.org/10.1002/cem.695.

[54] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, NY, 1966: pp. 391–420. https://ci.nii.ac.jp/naid/20001378860/en/.

[55] United States Department of Agriculture, Control/Eradication Agents for the Gypsy Moth - Human Health and Ecological Risk Assessment for Bacillus thuringiensis var. kurstaki (B.t.k.). FINAL REPORT, United States Department of Agriculture, Newtown Square, PA, 2004. https://www.fs.usda.gov/naspf/sites/default/files/volume-lll-chapters-1-8-appendixes-f-l-risk-assessments.pdf (accessed September 22, 2020).

[56] D.M. Weeks, M.J. Parris, A *Bacillus thuringiensis kurstaki* Biopesticide Does Not Reduce Hatching Success or Tadpole Survival at Environmentally Relevant Concentrations in Southern Leopard Frogs (*Lithobates sphenocephalus*), Environ Toxicol Chem. 39 (2020) 155–161. https://doi.org/10.1002/etc.4588.

[57] United States Environmental Protection Agency, Inert Ingredients Overview and Guidance, United States Environmental Protection Agency. (n.d.). https://www.epa.gov/pesticide-registration/inert-ingredients-overview-and-guidance (accessed September 22, 2020).

[58] E. Zomer, Floating sustained release pesticide granules, US 2004/0185079 A1, 2004. https://patents.google.com/patent/US20040185079A1/en (accessed September 22, 2020).

[59] T.M. Annesley, Ion Suppression in Mass Spectrometry, Clinical Chemistry. 49 (2003) 1041–1044. https://doi.org/10.1373/49.7.1041.

[60] J.L. Sterner, M.V. Johnston, G.R. Nicol, D.P. Ridge, Signal suppression in electrospray ionization Fourier transform mass spectrometry of multi-component samples, Journal of Mass Spectrometry. 35 (2000) 385–391. https://doi.org/10.1002/(SICI)1096-9888(200003)35:3<385::AID-JMS947>3.0.CO;2-O.

[61] X.S. Tong, J. Wang, S. Zheng, J.V. Pivnichny, P.R. Griffin, X. Shen, M. Donnelly, K. Vakerich, C. Nunes, J. Fenyk-Melody, Effect of Signal Interference from Dosing Excipients on Pharmacokinetic Screening of Drug Candidates by Liquid Chromatography/Mass Spectrometry, Anal. Chem. 74 (2002) 6305–6313. https://doi.org/10.1021/ac025988p.

[62] W.Z. Shou, W. Naidong, Post-column infusion study of the ‘dosing vehicle effect’ in the liquid chromatography/tandem mass spectrometric analysis of discovery pharmacokinetic samples, Rapid Communications in Mass Spectrometry. 17 (2003) 589–597. https://doi.org/10.1002/rcm.951.

[63] P.J. Larger, M. Breda, D. Fraier, H. Hughes, C.A. James, Ion-suppression effects in liquid chromatography–tandem mass spectrometry due to a formulation agent, a case study in drug discovery bioanalysis, Journal of Pharmaceutical and Biomedical Analysis. 39 (2005) 206–216. https://doi.org/10.1016/j.jpba.2005.03.009.

[64] E. Ciccimaro, I.A. Blair, Stable-isotope dilution LC–MS for quantitative biomarker analysis, Bioanalysis. 2 (2010) 311–341. https://doi.org/10.4155/bio.09.185.

[65] 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.

[66] Reaxys®, Reaxys. (n.d.). https://www.reaxys.com/ (accessed September 30, 2020).

[67] C.H. Johnson, J. Ivanisevic, G. Siuzdak, Metabolomics: beyond biomarkers and towards mechanisms, Nat Rev Mol Cell Biol. 17 (2016) 451–459. https://doi.org/10.1038/nrm.2016.25.

Chapter III

"Can we footprint the Volatilome?"

Preamble

The present chapter addresses **the analysis of the volatile part of pesticides residues** after their application on an environmental matrix. In fact, studying those volatile xenometabolites is required and can afford a number of advantages in order to understand the environmental fate.

First, the study of the volatile part of chemical substances is clearly **recommended by the OECD guidelines**. It is essential to assess the transformation of the tested substance and it should be analyzed by appropriate methods. Moreover, detecting and characterizing the volatile substances and the volatile transformation by-products of a pesticide are important in order to **estimate the exposure of farmers, workers, insects and plants to those residues that might represent certain toxicity**.

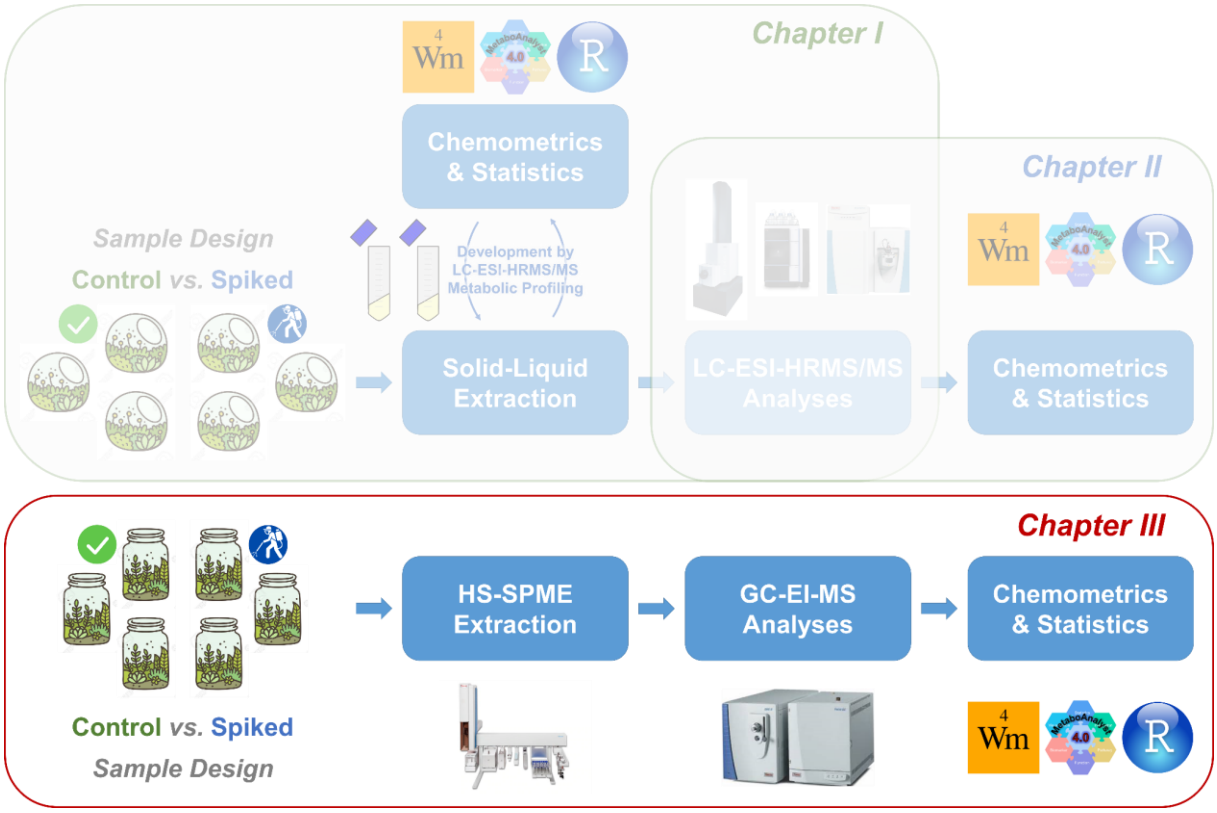
On the other hand, **detecting and characterizing volatile by-products of a pesticide help determining some of its degradation pathways**. This determination constitutes a complementary information to the analysis of the non-volatile part, by correlating the different characterized volatile and non-volatile by-products, which helps improving the understanding of the environmental fate.

Moreover, the **determination of the "dissipation time" of pesticide's volatile part can be a complementary indicator for the "resilience time"**.

To address this objective, the main challenge is to find an **appropriate approach capable to deal with the complex biopesticides**. These products consist of a wide variety of molecules including **several unknown compounds**. The proposed solution is a novel **untargeted metabolomics approach** that combines the **Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry** (HS-SPME-GC-MS) to the **computational data preprocessing** and the advanced **chemometric and statistical tools**.

Thus, **an automated and non-destructive online HS-SPME-GC-MS method** will be set-up for analyzing a **typical complex bioherbicide**; the "*Myrica gale* extract", applied on soil. The method will succeed to overcome the need of **a large number of samples**, by engineering a simple sampling design that allows analyzing one soil microcosm for several time points **without causing its destruction**. Then, the method will be applied for **38-days kinetic study**. The appropriate chemometric and statistical workflow will be then developed for the

investigation of the **volatile xenometabolome**, the **dissipation time** and **method sensitivity**. This application will serve to prove the novel concept of "Volatile-EMF".



Online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted volatile metabolomics for studying emerging complex biopesticides: A proof of concept

Hikmat Ghosson[1,2,*], Delphine Raviglione[1,2], Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,3,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: *Equal Contribution (Last Co-authors)*

Publication

Ghosson, H. *et al. Anal. Chim. Acta* **(2020)**, 1134:58–74. doi:10.1016/j.aca.2020.08.016

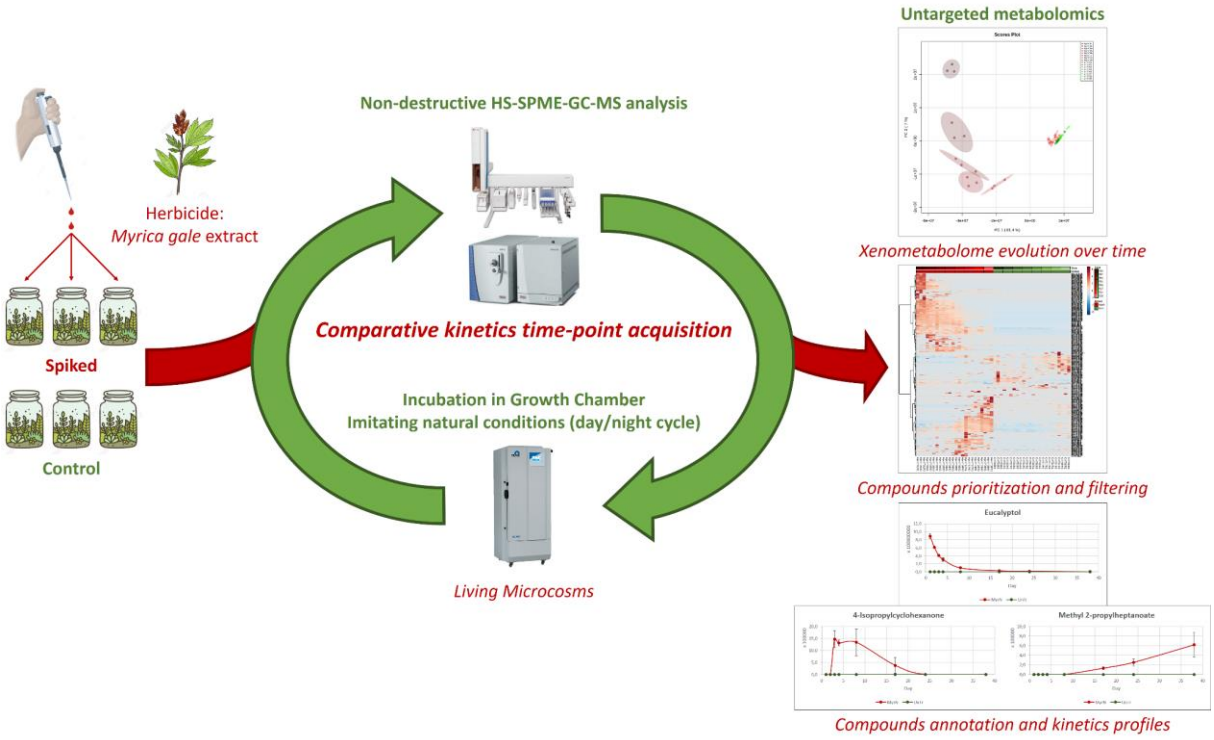
Keywords

Untargeted Metabolomics; Complex Biopesticides; Solid Phase Microextraction; Headspace; Gas Chromatography-Mass Spectrometry

Highlights

- Novel HS-SPME-GC-MS-based metabolomics approach is introduced for studying environmental fate of complex biopesticides.
- The *Myrica gale* methanolic extract was selected as a typical bioherbicide applied on soil for the proof of the concept.
- A green, non-destructive automated method was developed and applied for a comparative 38-day kinetics experiment.
- Untargeted analyses allowed explaining xenometabolome evolution through the time and prioritizing 101 xenometabolites.
- 96 xenometabolites were putatively identified, including 63 compounds reported for the first time in the studied herbicide.

Graphical Abstract



Abstract

This work introduces a novel online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted metabolomics approach, suggested as an alternative tool to study the environmental fate of volatile xenometabolites in emerging complex biopesticides, *e.g.* the *Myrica gale* methanolic extract herbicide containing several unknown metabolites. A "living" microcosm sample was designed for non-destructive analysis by a 35-minute HS-SPME automated extraction and a 36-minute GC-MS run. A 38-day kinetics study was then applied on two groups of soil samples: control and spiked. Statistical tools were used for the comparative kinetics study. The Principal Component Analysis revealed and explained the evolution and the dissipation of the herbicide volatile xenometabolome over time. The time-series Heatmap and Multivariate Empirical Bayes Analysis of Variance allowed the prioritization of 101 relevant compounds including 22 degradation by-products. Out of them, 96 xenometabolites were putatively identified. They included 63 compounds that are identified as herbicide components for the first time. The Orthogonal Projections to Latent Structures Discriminant Analysis and its Cross-Validation test were used to assess the total dissipation of the herbicide volatile residues and method sensitivity. The reproducibility of the method was also assessed. The highest inter-samples *(n = 3)* Peak Area RSD was 7.75 %. The highest inter-samples *(n = 3)* and inter-days *(n = 8)* Retention Time SD were 0.43 s and 3.44 s, respectively. The work presents a green, non-laborious and high-throughput approach. It required a small number of environmental samples (6 microcosms) that were analyzed 8 times and were not destroyed during the study.

1. Introduction

Pesticides research, development and production are constantly expanding since these chemicals and agents are essential for several anthropogenic and economic activities (*e.g.* agriculture, food production and protection, disease vectors control). Their development however, faces numerous problems due to their potential impact on human health [1–3] and ecology [4,5]. These issues reinforce the requirement and the importance of prior in-depth studies of their fate, impacts and risks on health and environment. Also, the development of new pesticides of natural origins, known as "biopesticides" or "biocontrol agents" (BA), is one of the suggested alternatives to chemical/synthetic pesticides, as they are presumed to be less harmful for human health and environment. Moreover, their dissipation is likely to be relatively fast [6].

Extracted from plants or different types of microorganisms, these emerging natural products are mostly based on one or several bioactive compounds which usually act in a synergic and/or pleiotropic mode of action. Their complex (bio)chemical nature containing several different and unknown molecules and/or macromolecules is requiring new conceptual and analytical challenges for the assessment of their transformation and dissipation. The classic concepts of fate assessment, such as the DT50 approach [7], are non-applicable for such types of complex pesticides. These classic targeted approaches are limited to known compounds and molecules. In addition, the DT50 approach does not consider the transformation products (TPs) of the pesticide, in particular the unknown TPs. Additional protocols and approaches are therefore needed in order to assess the pesticide transformation in the environment, and to study the impact of its application on environmental biodiversity.

New analytical proxies were thus suggested as alternative approaches for the emerging complex biopesticides, mainly based on untargeted metabolomics strategies [8,9]. A new approach called Environmental Metabolic Footprinting (EMF) was recently introduced by Patil *et al.* [10] and Salvia *et al.* [11]. This new approach presents the application of the untargeted metabolomics as a universal tool for kinetics studies in order to assess both the fate and impact of different types of complex pesticides. This aims to introduce an integrative concept called "resilience time".

In the two previous mentioned works [10,11], kinetics studies were performed on an important number of samples by applying destructive Solid-Liquid Extractions (SLE) followed by Liquid Chromatography-Mass Spectrometry (LC-MS) analysis. They were restricted to the solid phase

of the environmental matrices (soil and sediments). However, studying the volatile part of the xenometabolome, *i.e.* pesticide compounds and their TPs, is essential for the risk assessment of these emerging biopesticides, and in particular for products based on plant essential oils, which contain an important amount of volatile and semi-volatile organic compounds. The OECD guidelines for the testing of chemicals and their transformation recommend the consideration of the volatile part [12,13]. In fact, studying pesticide volatile residues can provide complementary information to better understand its environmental fate. In addition, pesticide volatile residues screening allows to assess the exposure risk to pesticide compounds for farmers/workers, insects and plants, as well as the exposure to their TPs that might be more toxic. Therefore, the aim of the present work is to introduce the concept of a new untargeted metabolomics-based approach, dedicated to analyze and study the volatile residues of emerging complex biopesticides applied on environmental matrices, by using online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry (HS-SPME-GC-MS).

HS-SPME is an appropriate technique for volatile organic compounds analysis. It is based on extracting and isolating these analytes from the sample by adsorbing and concentrating them on the layer of a coated fiber. Thus, they can be eventually desorbed and introduced in the analytical instrument with or without the need of extraction solvents [14,15]. Since its introduction in 1989 by Belardi & Pawliszyn [16], the SPME is still being widely developed and extensively used for different types of targeted and untargeted analytical approaches as broadly described by Reyes-Garcés *et al.* [17]. For pesticides research, several works have been reported and were mainly focused on targeted screening and quantification of pesticide residues in different environmental and food matrices [18–20]. Untargeted screening to study the transformation of pesticides and to identify their by-products was also reported, but in a much fewer number of publications [18,21,22].

SPME presents several advantages as a green, non-destructive and cost-effective technique. Its automation provides additional advantages, particularly for metabolomics approaches, mainly by enhancing the robustness and the reproducibility of the applied extraction method. Moreover, reducing the laborious time-consuming manual work and sample preparation steps is essential for high throughput analyses and to minimize errors related to sample handling. Otherwise, as a green non-destructive method, the application of the HS-SPME reduces the number of environmental samples needed, by making it possible to analyze the same sample for several time points, particularly in case of kinetics tracking study. This can also enhance the

performance of the approach by reducing sample preparation and random biological variations-related biases.

On the other hand, GC-MS analytical technique provides several advantages concerning untargeted metabolomics. The GC is a suitable separation technique for volatile and semi-volatile organic compounds. It is well known for its significant analytical robustness, affording high chromatographic resolution and precise retention time repeatability [23]. GC also provides a tool for compounds' identification by allowing the calculation of Kováts Retention Index (RI) [24], which is an advantage for the identification of unknown xenometabolites. Mass spectrometers are highly sensitive detectors capable of characterizing and quantifying compounds. In this work, the chosen detector is a Single Quadrupole MS, equipped with an Electron Impact (EI) ionization system. The main advantages of this spectrometer are the large dynamic range of the Quadrupole mass analyzer, its high scan frequency, and the ability of the EI to provide reproducible fragmentations for the analyzed compounds [23]. This presents an essential tool for characterizing unknown compounds by fast spectral library search and/or by structural elucidation.

All of these advantages were considered for the development of an online HS-SPME-GC-MS method, which was dedicated for studying the environmental fate of an emerging bioherbicide applied on soil: the *Myrica gale* methanolic extract.

Introduced by Popovici *et al.* [25,26], the herbicide composition was partially identified by several studies [25–30]. Its bioactive compound is Myrigalone A, an allelochemical, mixed with several other compounds: mainly triketones and terpenes. The herbicide mode of action was described by Oracz *et al.* [31]. This research work revealed a potential synergic activity between Myrigalone A and terpenes. This activity was recently confirmed and explained by Khaled *et al.* [32]. Therefore, an optimal herbicide activity requires the application of the total complex mixture of the plant extract. However, several components in this complex mixture are still unknown, and their transformation in nature is not deeply understood. Thus, the untargeted metabolomics approach is a potential solution for studying the environmental fate of this bioherbicide. Therefore, in order to prove the concept of the suggested HS-SPME-GC-MS-based untargeted metabolomics approach, the *Myrica gale* methanolic extract was selected as a typical complex bioherbicide in order to study the dissipation of its volatile residues after its application on soil, through a 38-days kinetics study. The study targets exclusively volatile

residues that are spontaneously released to the gas phase above soil (the headspace) during imitated environmental conditions applied to microcosm samples.

2. Material and Methods

2.1. Chemicals

Methanol HPLC grade was purchased from VWR International (Fontenay-sous-Bois, France). The dry methanolic extract of *Myrica gale* was prepared as described Popovici *et al.* [25]. The spiking herbicide solution for application on soil samples was prepared at a concentration of 72 mg mL^{-1} of dry extract dissolved in Methanol (containing 18 mg mL^{-1} of the bioactive compound Myrigalone A). C7-C30 Saturated Alkanes mix (1000 µg mL^{-1} of each component in Hexane) was purchased from Sigma-Aldrich (Saint-Quentin-Fallavier, France).

2.2. Soil material

Soil sample was collected from an arable field at the agricultural domain of the "Institut Universitaire de Technologie" (IUT) of Perpignan, France (42°40'55.1"N 2°53'51.2"E). The surface layer (15 cm) of soil was collected on 3 different points separated by 1.5 meter. After collection, the soil was homogenized and passed through a 2 mm sieve. Then, it was stocked in the dark at 4 °C until the experiment. The soil composition analysis and characterization were performed by Arterris Laboratory (Toulouges, France) accredited by the French Accreditation Committee (Cofrac). Results were the following: 13.9 % of clay, 60.5 % of silt, 25.6 % of sand, 20 % of soil humidity, 1.7 % of organic matter, 0.98 % of organic Carbon, 15.5 meq 100 g^{-1} cation exchange capacity (CEC), 214 % Ca^{2+}/CEC and pH of 8.1 in water. According to the Soil Textural Triangle of the United States Department of Agriculture [33], this soil is classified as a silt loam soil. It was never been contaminated or exposed to herbicides.

2.3. Soil samples set-up

Samples consisted of 6 g of soil weighted in 20 mL HS-SPME vials (Thermo Fisher Scientific, Courtabœuf, France). This weight was optimized in order to keep 2/3 of the vial volume as headspace. After, vials were hermetically closed by a crimped septum, and two 18G×1 ½" (1.2 × 38 mm) Agani™ needles (Terumo®, Leuven, Belgium) were implanted on the extremity sides of the septum (Figure A.III-A 1 – Appendix III-A). This is to assure aerobic conditions by allowing air exchange between the internal headspace and the outside. The prepared soil vials were incubated in a GC 401 growth chamber (Nüve, Saracalar, Turkey) for 24 hours before the spiking in order to reestablish the biological and microbial activity. Incubation conditions were 24 hours day/night cycle with alternation of light/dark, 28 °C/18 °C of temperature, and 40 % RH/65 % RH of humidity (Figure A.III-A 2 – Appendix III-A). The soil moisture was

maintained at 20 % during the incubation and throughout the experiment, following a standardized environmental protocol implemented and published in previous works [10,34], aiming to assure conditions that are comparable to real environmental cases. The aim of implementing this sample design was to assure a "living system". As mentioned previously, samples will be used for several kinetic time points, so measures were taken to ensure that they will not be destroyed during the study.

The preparation of herbicide-spiked soil samples was performed by applying the *Myrica gale* methanolic extract with a dose equivalent to 300 µg of the active compound (Myrigalone A) per gram of soil (1.2 mg of dry *Myrica gale* methanolic extract per gram of soil). This corresponds to ten-times the agronomical field dose, following testing guidelines recommendations [12,13] in order to assess their transformation and risks on health and environment in an extreme pollution scenario [34].

2.4. Headspace-Solid Phase Microextraction development

Automated Headspace-Solid Phase Microextraction (HS-SPME) was performed using a TriPlus™ RSH™ autosampler (Thermo Fisher Scientific, Waltham, U.S.). The extraction method was developed by optimizing the following conditions and parameters: the SPME fiber coating, the incubation time, the extraction time, and the extraction temperature. Tests were performed by analyzing herbicide-spiked soil samples (prepared following the protocol described in Section 2.3.).

SPME fiber coating tests were performed by comparing 6 different types of coatings: 100 µm Polydimethylsiloxane (100 µm PDMS, Fused Silica, 23 Ga, Autosampler), 7 µm Polydimethylsiloxane (7 µm PDMS, Fused Silica, 24 Ga, Autosampler), 85 µm Polyacrylate (85 µm PA, Fused Silica, 23 Ga, Autosampler), 65 µm Polydimethylsiloxane/Divinylbenzene (65 µm PDMS/DVB, Stableflex, 23 Ga, Autosampler), 85 µm Carboxen/Polydimethylsiloxane (85 µm CAR/PDMS, Stableflex, 23 Ga, Autosampler), and 50/30 µm Divinylbenzene/Carboxen/Polydimethylsiloxane (50/30 µm DVB/CAR/PDMS, Stableflex, 23 Ga, Autosampler), all purchased from Supelco (Bellefonte, U.S.). Tests were performed by applying the following HS-SPME conditions: 5 min of incubation time, 30 min of extraction time, and 40 °C of extraction temperature.

Next, the duration of sample incubation before the SPME extraction (incubation time) was assessed in order to choose the optimal condition. 3 different incubation times were tested using

the selected 50/30 µm DVB/CAR/PDMS fiber: 5 min, 15 min, and 30 min (extraction time: 30 min, extraction temperature: 40 °C).

After, the exposure duration of the SPME fiber to the Headspace (extraction time or adsorption time) was assessed. 7 different values were tested: 5 min, 10 min, 20 min, 30 min, 40 min, 50 min and 60 min (incubation time: 5 min, extraction temperature: 40 °C, fiber coating: 50/30 µm DVB/CAR/PDMS).

Regarding extraction temperature, 3 values were tested in order to assess the impact of increasing temperature on volatile metabolic profiles. Tested temperatures are the following: 40 °C, 60 °C, and 80 °C (incubation time: 5 min, extraction time: 30 min, fiber coating: 50/30 µm DVB/CAR/PDMS).

Finally, a dose response curve was applied after adapting optimal conditions. This in order to examine fiber's over-saturation. 6 different herbicide doses were applied on 6 different batches of soil samples (with 3 biological replicates for each dose batch), and then analyzed and compared to control untreated soil samples (3 biological replicates) in order to assess method's sensitivity. The 6 applied doses corresponded to: 10^{-3}-time, 10^{-2}-time, 10^{-1}-time, 1-time, 10-times, and 20-times the agronomic field dose of the herbicide.

For all optimization tests and method's application, the incubated sample vial was shaken vigorously throughout the incubation and the extraction procedures in order to enhance the homogenization of sample temperature.

2.5. Gas Chromatography-Mass Spectrometry

Gas Chromatography-Mass Spectrometry analyses were performed on a Focus GC system coupled to an Electron Impact-Single Quadrupole DSQ II Mass Spectrometer (Thermo Fisher Scientific, Waltham, U.S.; Bremen, Germany). An Agilent J&W DB-5MS GC column was used for separation (length: 30 m, inner diameter: 0.25 mm, film thickness: 0.25 µm, Agilent Technologies, Santa Clara, U.S.). Desorption was performed in Splitless mode for a duration of 1 min at an inlet temperature of 230 °C, followed by a 5 min post-injection fiber conditioning at 260 °C in order to prevent fiber carryovers. The 36-min GC run was developed for an optimal compounds separation. It consisted of a 1 mL min^{-1} constant flow method with Helium as carrier gas. The oven temperature was programed as the following: an initial temperature of 60 °C was held for 1 min, and was then followed by a first ramp of 10 °C min^{-1} in order to reach 100 °C. After, a second ramp of 3 °C min^{-1} was applied and held until a temperature of 182 °C was

reached. Finally, the last ramp of 25 °C min^{-1} was applied until a temperature of 230 °C was reached. This end temperature was held for 2 min in order to prevent any potential column carryover. GC-MS transfer line temperature was maintained on 240 °C throughout the run.

The MS acquisition method was a Full MS scan for positive ions with an *m/z* range of 40-400. The scan rate was 5 scans s^{-1} (2027.11 amu s^{-1}). The source temperature was set to 250 °C, the applied electron energy was -70 eV, and the detector gain was equal to 30000 (1362 V).

2.6. Software and data processing

GC-MS piloting and data acquisition were performed using Xcalibur 3.0.63 (Thermo Fisher Scientific, Waltham, U.S.). Data were acquired in RAW format and then converted to ANDI format (NetCDF) in order to upload and process them using Galaxy Workflow4Metabolomics platform [35–37]. Data are published on the MetaboLights platform (EMBL-EBI, European Bioinformatics Institute, Hinxton, U.K.) [38,39]. The automated processing workflow used the metaMS package (Galaxy Version 2.1.1) [40] dedicated for GC-MS data. All of its conditions and parameters were published on the platform [41]. In brief, a "matchedFilter" algorithm was used for peak piking, with a Full Width at Half Maximum (FWHM) of 5 (Gaussian model peak) [42]. In addition, GC-MS peaks were considered for peak piking only if: i) their pseudo-spectra contained a minimum of 5 *m/z* features, ii) if these peaks were present in at least 70 % of samples belonging to a defined condition. Between the different injections/runs, the similarity threshold between peaks pseudo-spectra was set to 0.7, and maximum peak Retention Time (RT) variation was set to 15 s in order to prevent any potential splitting of a metabolite feature into two different features. After generating the data matrix, statistical analyses were performed using the R-based MetaboAnalyst platform [43–45]. Xcalibur 4.1.31.9 (Thermo Fisher Scientific, Waltham, U.S.) and AMDIS 2.72 (National Institute of Standards and Technology, Gaithersburg, U.S.) were used for the deconvolution of MS spectra and the manual data processing to cross-check the results obtained by the automated processing. Compass DataAnalysis 4.3 (Bruker Daltonik GmbH, Bremen, Germany) was used for EIC peak area integration and for counting molecular features' number. NIST 14 library search for putative identification of compounds was performed using NIST MS Search 2.2 (National Institute of Standards and Technology, Gaithersburg, U.S.). Welch Two Sample T-test for independent means comparison was performed using the R Commander 2.4-2 "Rcmdr" package [46] of R 3.3.3 software.

2.7. Application for a kinetics study

After all analytical conditions were optimized and set-up, a 38-day kinetics tracking study was conducted to prove the concept of the suggested approach. The studied environmental samples consisted of two different groups of soil vials/microcosms (described in the Section 2.3.) with 3 replicates of each: an untreated control soil (UnTr), and an herbicide-spiked soil (MyrN). After spiking, samples were incubated in the growth chamber with the day/night cycle conditions mentioned in Section 2.3., in order to imitate natural conditions for herbicide transformation in soil.

Next, 8 different kinetic time points were analyzed: day 1, day 2, day 3, day 4, day 8, day 17, day 24, and day 38 after spiking. The same soil samples were analyzed by the HS-SPME-GC-MS developed method for all the 8 time points. The order of injections of the different samples was randomized in order to reduce the impact of potential analytical drifts. Blank injections were performed during each time point analysis, by extracting and analyzing the headspace of an empty 20 mL vial using the same HS-SPME-GC-MS method. For Kováts RI calculation, 20 µL of the C7-C30 Alkanes mix solution were introduced to a 20 mL vial, then it was analyzed by applying the same HS-SPME-GC-MS method.

After each analysis, soil microcosms were re-incubated in the growth chamber until the next kinetics time point.

3. Results and Discussions

3.1. Headspace-Solid Phase Microextraction optimization

The HS-SPME method was optimized in order to establish a compromise between three major criteria: i) assuring an optimal sensitivity for a wide-range detection of different types of volatile compounds, ii) applying non-destructive conditions to soil samples, iii) preventing an induced volatilization of compounds that are relatively less volatile in the imitated environmental conditions, as the approach targets exclusively volatile residues that are spontaneously released to the gas phase above soil.

For the selection of the SPME fiber coating, Results of tests are shown in Figure A.III-A 3 and Table A.III-A 1 (Appendix III-A). PDMS/DVB, DVB/CAR/PDMS and CAR/PDMS showed better results in term of total TIC area and number of molecular features when compared to the 2 PDMS and the PA coatings. In addition, CAR/PDMS fiber coating showed the highest total TIC area and the highest number of molecular features, followed by the DVB/CAR/PDMS, and then the PDMS/DVB.

Nonetheless, performances of PDMS/DVB, DVB/CAR/PDMS and CAR/PDMS coatings were in-depth examined. A data matrix was generated by processing GC-MS raw data of fiber tests (using the same processing method described in Section 2.6.), and then a Heatmap analysis was applied on the dataset. Heatmap (Figure A.III-A 4 – Appendix III-A) shows that PDMS/DVB and CAR/PDMS coatings differ by their specificity for different types of herbicide compounds (as highlighted with yellow boxes in the Figure A.III-A 4). However, DVB/CAR/PDMS coating is able to extract simultaneously a part of compounds that are extracted with the PDMS/DVB exclusively, and another part of compounds that are extracted with the CAR/PDMS exclusively (as outlined by the green boxes in the Figure A.III-A 4). Therefore, for the current work, the use of the DVB/CAR/PDMS coating is considered as the best compromise between the highest sensitivity and the widest molecular diversity.

Regarding the duration of sample incubation before the SPME extraction (incubation time), Figure A.III-A 5 and Table A.III-A 2 (Appendix III-A) show that the increase of incubation time decreases the sensitivity of the method (in term of total TIC area and number of molecular features). This decrease of sensitivity can be hypothetically explained by the accumulation of a higher ratio of water vapor in the headspace. This may prevent the optimal adsorption of some compounds to the SPME fiber, such as L-α-bornyl acetate containing an Ester function, and epi-γ-Eudesmol and α-Terpineol both containing a Hydroxyl function (Table A.III-A 3 –

Appendix III-A). Therefore, an incubation time of 5 min was chosen as an optimum for sensitivity.

Concerning the exposure duration of the SPME fiber to the Headspace (extraction time or adsorption time), results in Figure A.III-A 6 and Table A.III-A 4 (Appendix III-A) show that a significant difference (in term of total TIC area and number of molecular features) is observed when comparing 5 min, 10 min and 20 min, *vs.* 30 min, 40 min, 50 min and 60 min. For those last 4 values of extraction time, total TIC areas and numbers of molecular features seem to be no more evolving. Therefore, a 30 min extraction time was chosen as a compromise between sensitivity and short-time analysis.

Regarding extraction temperature, this parameter is constrained by two problematics: i) the application of relatively high temperatures risks to deteriorate the environmental samples. These risks should be avoided as the current study aims to implement a non-destructive method. ii) As mentioned previously, the scope of the approach is to target exclusively volatile residues that are spontaneously released to the headspace during the imitated environmental conditions. Applying relatively high temperature can provoke an induced volatilization of compounds that are relatively less volatile in those conditions, which should be avoided in order to prevent a conceptual bias. The provocation of this induced volatilization was proved by testing 3 extraction temperatures: 40 °C, 60 °C, and 80 °C. According to results in Figure A.III-A 7 and Table A.III-A 5 (Appendix III-A), the increase of extraction temperature led to a decrease in signal for compounds eluted between 40 °C and 130 °C (0 min to 11 min of RT), meanwhile an increase in signal for compounds eluted between 130 °C and 230 °C (11 min to 21 min) was observed. Therefore, beside its destructive aspect, increasing extraction temperature seems to decrease method's sensitivity for the relatively volatile compounds, meanwhile it increases the signal of compounds that are relatively less volatile in environmental conditions.

On the other hand, temperatures below 40 °C were non-applicable in the current work due to problems in stabilizing incubator temperature. This problem risks deteriorating the reproducibility of the extraction. Thus, 40 °C is considered as the optimal compromise for extraction temperature.

To sum up, the optimal HS-SPME conditions applied for the study are the following: 50/30 µm DVB/CAR/PDMS as fiber coating, 5 min of incubation time, 30 min of extraction time, and 40 °C of extraction temperature. To assess the over-saturation of the fiber with these conditions, a dose response curve was examined. Results in Figure A.III-A 8 and Figure A.III-A 9

(Appendix III-A) show that at 20-times the field dose, the Total TIC area and the number of detected molecular features are still increasing. This means that at the optimized HS-SPME conditions, the fiber is not yet over-saturated when analyzing 10-times the field dose (*i.e.* the dose applied for the kinetics study), as the fiber is still able to adsorb higher number and quantity of compounds.

It is worth mentioning that despite the important influence of moisture ratio on the detection of several volatile metabolites, the variation of this parameter is constrained by the complexity of the environmental context. In fact, the moisture ratio fixed at 20 % throughout the current study aims to assure conditions that are comparable to real environmental cases (following previously published protocols [10,34]). Setting a moisture ratio that does not represent the standardized environmental/biochemical conditions question of the study risks to change the abiotic and biotic transformation pathways of xenometabolites during the kinetics study. In addition, the variation of moisture ratio can *de facto* provoke the volatilization of metabolites that are relatively less volatile when the standardized environmental conditions are in-place.

3.2. Herbicide residues detection and low matrix background

After 1 day of spiking, a rich profile of extracted herbicide volatile residues was detected by HS-SPME-GC-MS analysis, as shown in Figure C.III 1. The detected analytes were eluted between 60 °C and 175 °C (1 min to 30 min of RT), presenting a complex volatile fingerprint with several major and minor compounds.

In contrast to the spiked soil, the HS-SPME extract of the untreated control soil samples did not contain an important number of detected compounds (Figure C.III 1). Compared to the blank GC-MS profiles, there was no significant difference. In both groups, all detected peaks mainly consisted of silicon-derivate compounds. These compounds are probably issued from the bleeding of septum, fiber and/or GC column.

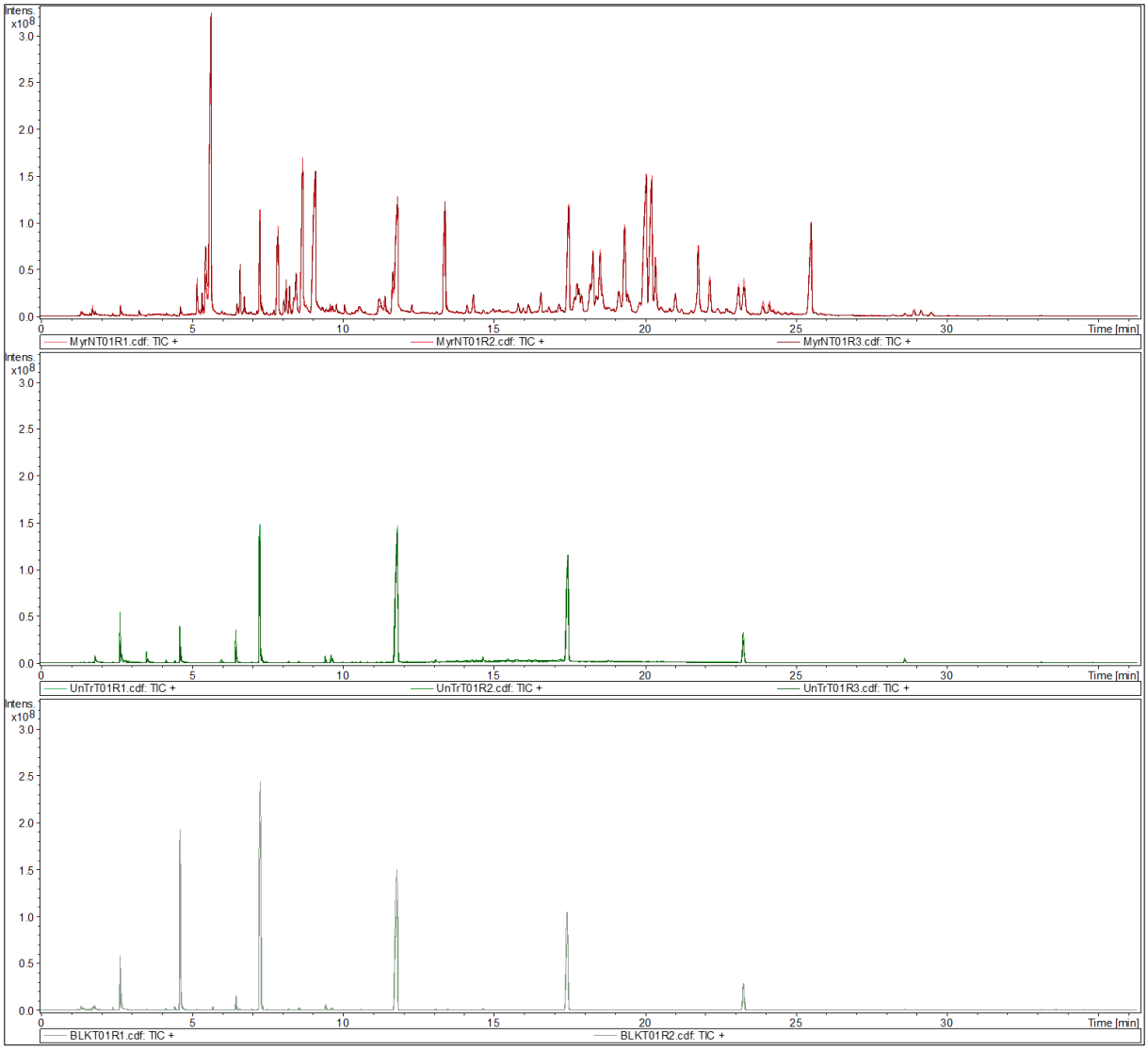


Figure C.III 1: GC-MS chromatograms of HS-SPME extracts after one day of spiking for spiked samples "MyrN" (red), control untreated "UnTr" samples (green) and blanks (grey).
For the two sample groups, figures consist of three overlaid chromatograms (TIC) of the three biological replicates. For blanks, two chromatograms of the two analytical replicates are overlaid.
Chromatograms were performed with Compass DataAnalysis 4.3 software. The intensity scale is fixed to 3.50E8.

The poor GC-MS profile of the untreated control soil HS-SPME extracts reveals the difficulty in extracting and/or detecting endogenous metabolites originating from soil. Thus, this method is not suitable for studying the impact of the applied herbicide on the soil biodiversity. The advantage, however, is the selectivity of the HS-SPME-GC-MS method to the residues of *Myrica gale* extract in the current study, leading to a low matrix background. This can improve the study of the environmental fate of the herbicide, by enhancing the detection, the quantification and the identification of volatile compounds issued from its xenometabolome, and preventing matrix effects and interferences originating from the matrix.

3.3. Untargeted metabolomics analyses

To prove the concept of the suggested untargeted approach, the 38-day kinetics tracking was performed by applying the HS-SPME-GC-MS analysis on the two groups of samples; the control untreated soil and the soil spiked with the *Myrica gale* extract herbicide (as described in Section 2.7.). After the end of the kinetics tracking and the acquisition of all data, RAW files were converted to ANDI format (NetCDF) and then uploaded on the Galaxy Workflow4Metabolomics platform for data preprocessing (Section 2.6.). The generated data matrix consisted of 64 analyzed samples (24 untreated control samples, 24 spiked samples, and 16 blank injections), and 376 variables. Each of these variables represents a "picked" pseudo-spectrum after it was defined by retention time-based clustering of its *m/z* fragment ion signals using CAMERA package [40,47]. *In fine*, depending on the applied parameters of the preprocessing [41], each variable should represent a relevant detected compound (without neglecting the high possibility of considering noise and artefacts). This acquired data matrix was used for the statistical analyses.

3.3.1. Principal Component Analysis

First, the Principal Component Analysis (PCA) was applied. All kinetics time points of both untreated control (UnTr) and spiked (MyrN) samples were integrated. The PCA played an important role for understanding the results that were acquired with this approach. It shows that over time, the volatile metabolic profiles of the spiked samples tend to converge with those of the untreated control samples (Figure C.III 2). According to the first principal component axis (PC1), the later kinetics time points, *i.e.* days 17, 24 and 38 after spiking, were more similar to the control profiles in comparison with the earlier kinetics time points. This means that after 17 days of herbicide application, an important dissipation of its xenometabolome had occurred. In fact, the PC1 that explains 81.4 % of variations, consists of the regression of the main features issued from the xenometabolome. This was confirmed by exploring the loadings of the PC1, revealed by the loading plot of the PCA and the Biplot (Figure C.III 3). The 6 most significant features of the PC1 were only present in the extracts of spiked soils as shown in Figure C.III 4. They were more abundant particularly in the earlier kinetics time points, *i.e.* days 1, 2, 3, 4 and 8 after spiking.

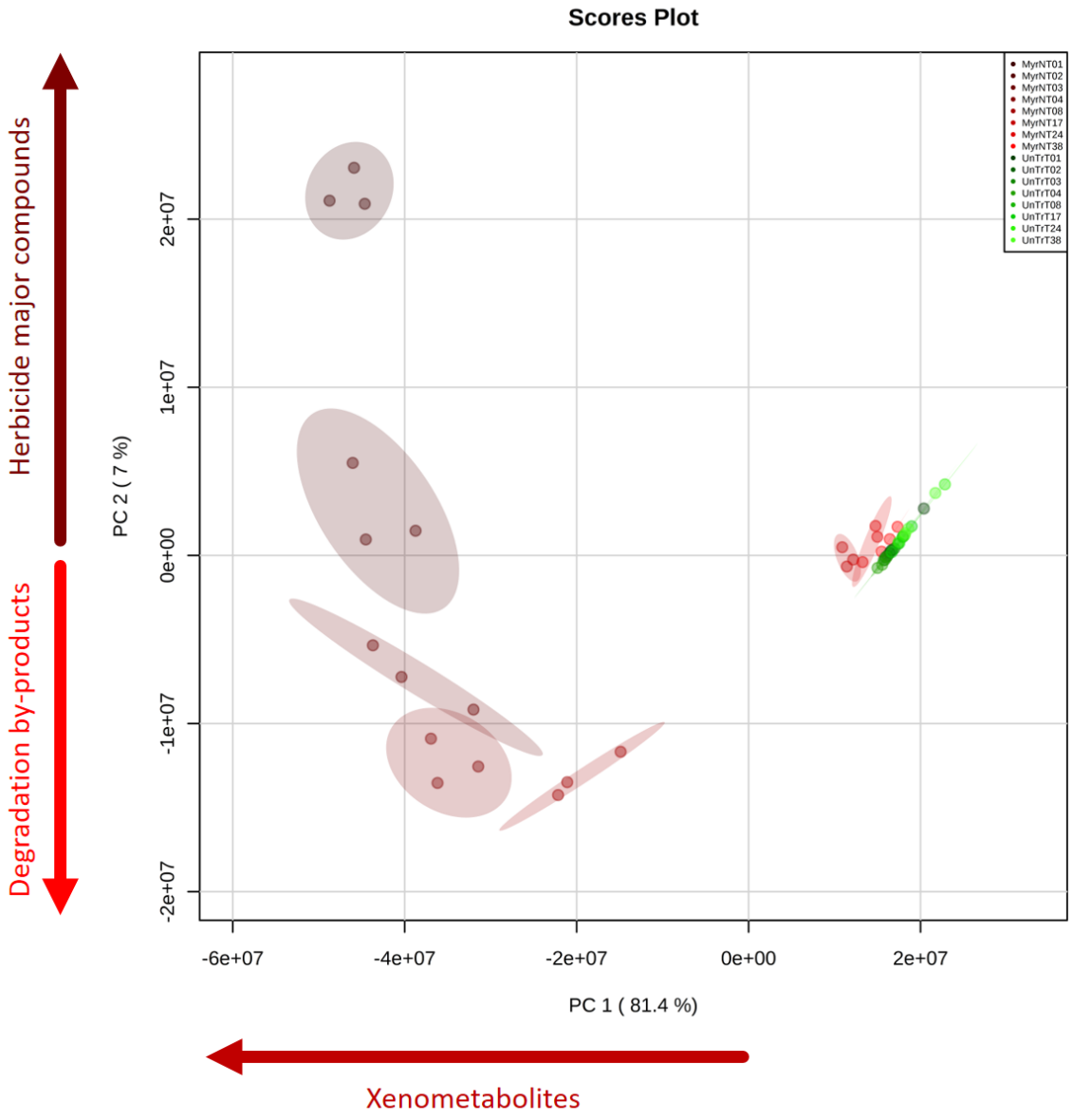


Figure C.III 2: Principal Component Analysis (PCA).
Plot generated using MetaboAnalyst.

Another important result regarding the degradation of the herbicide was revealed by the PCA. In fact, a progressive evolution of the volatile profiles of the earlier kinetics time points, *i.e.* from day 1 (T01) to day 8 (T08), was significantly observed on the second principal component axis (PC2). The explanation of this result is that the PC2, which accounts for 7 % of variations, consists of two main types of volatile xenometabolites: the major herbicide volatile compounds contained in the *Myrica gale* extract, and the volatile degradation by-products issued from herbicide compounds. These two "families" of xenometabolites constituted the two opposed sides of the PC2 as shown in Figure C.III 2. This explanation was confirmed by the loadings of the PC2 (Figure C.III 3). The 6 most significant features of the upper part of the PC2 axis were the compounds of the herbicide. Their highest abundance was at day 1 (T01), and then started

to decay over time (Figure C.III 5A). For the 6 features with the highest contributions in variation on the lower part of the PC2 axis, their abundance increased over time, before starting to decay in the later kinetics time points (Figure C.III 5B). Thus, these features represent the by-products issued from the degradation of the herbicide mixture.

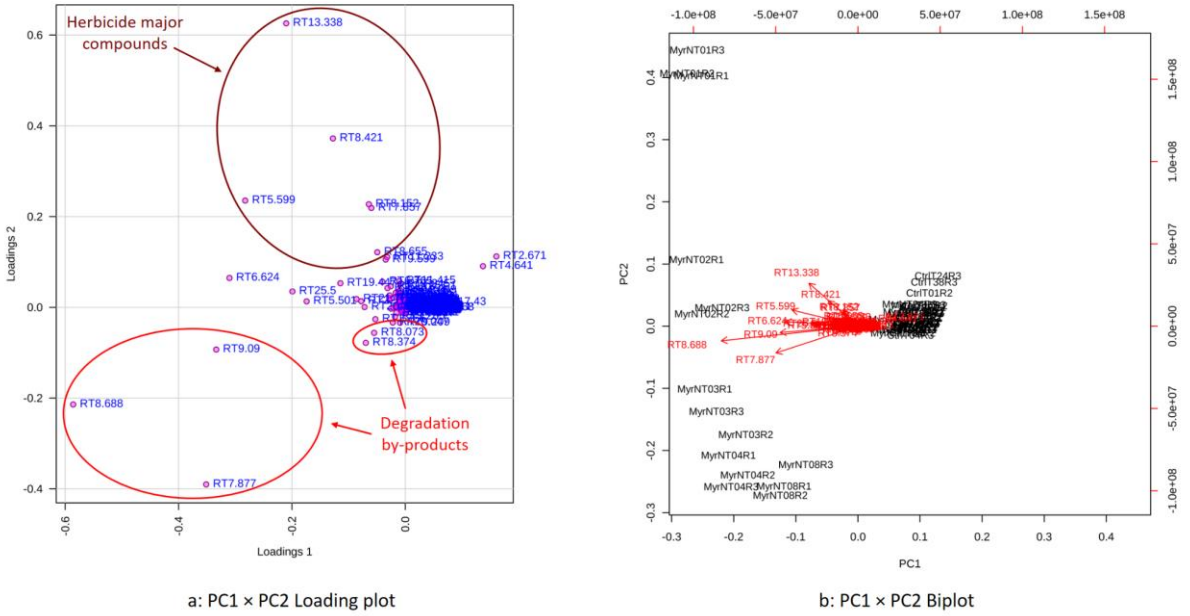


Figure C.III 3: Loading plot and Bioplot showing correlations between samples and features of the PC1 and the PC2.
Plots generated using MetaboAnalyst.

It is worth mentioning that according to PCA, there was no significant difference between volatile profiles of the untreated control samples over time. This proves another advantage of reducing the matrix background, by eliminating soil biochemical evolution factor from analyses. Therefore, tracking and understanding herbicide's environmental fate are enhanced from a chemical-analytical point of view.

Ultimately, PCA provided a general understanding of the evolution of xenometabolome through the time. In-depth analyses were then conducted to explain this evolution by filtering and tracking xenometabolites over time, in order to identify their nature and to annotate them.

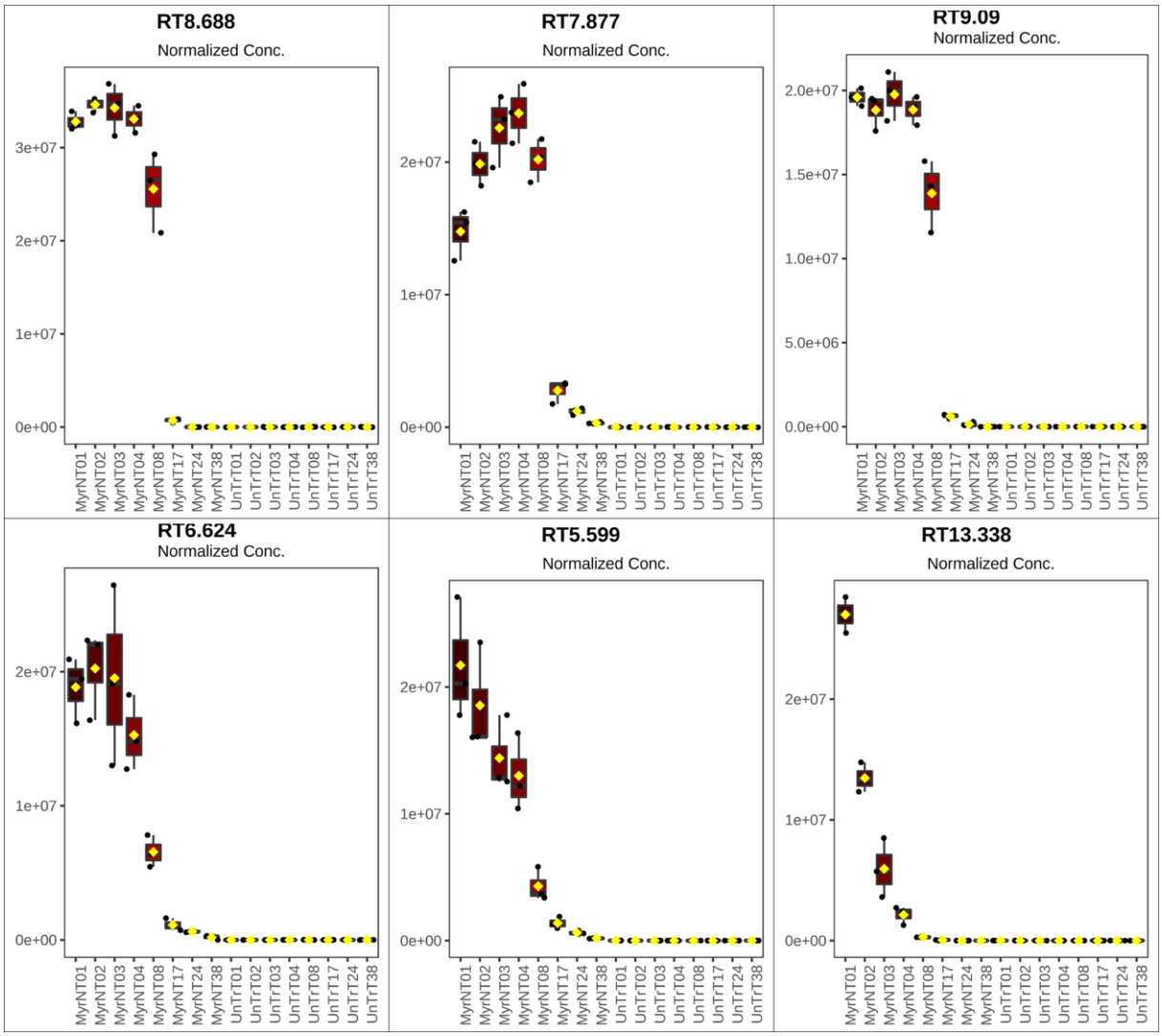


Figure C.III 4: Boxplots of features with the highest contributions in variation of PC1. The abundance in the two groups of soil samples and their evolution over time are represented.
Boxplots show the null abundance of these features in all control samples (UnTr). The abundance decay over time in spiked samples is also shown (MyrNT01 to MyrNT38, respectively).
Features plots are sorted according to the descending order of their PC1 scores (in absolute value), from the left to the right, and then from the top to the bottom.
Plots generated using MetaboAnalyst.

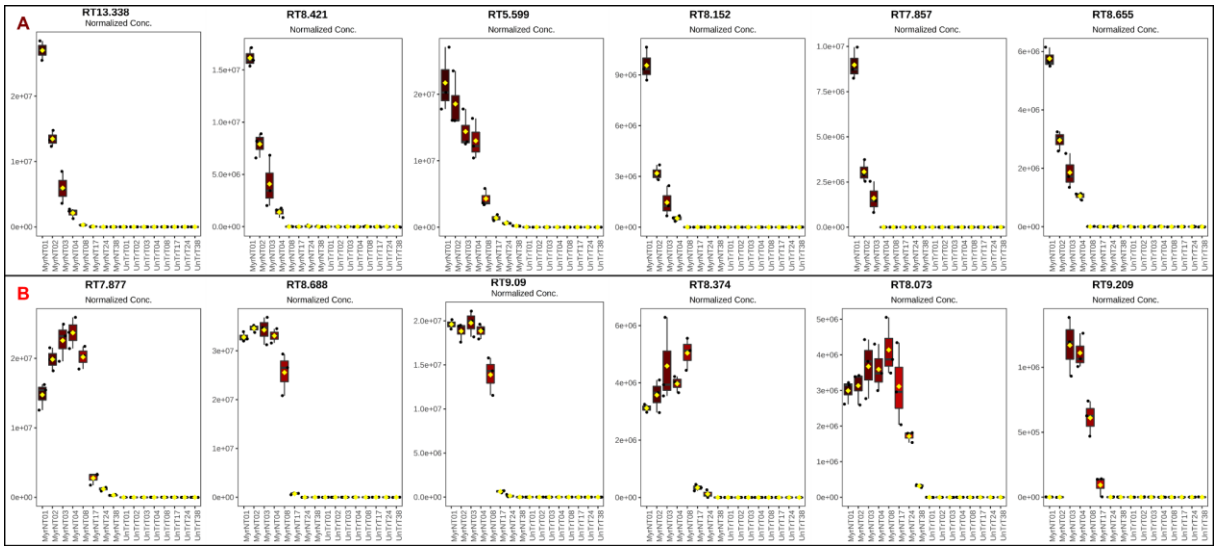


Figure C.III 5: Boxplots of features with the highest contributions in variation on the two opposite sides of the PC2.
A: Most significant features on the upper part of PC2 axis. Their highest abundance is at day 1 (MyrNT01), and then it starts to decay through time.
B: Most significant features on the lower part of PC2 axis. Their abundance increases through time until starting to decay by days 8 and/or 17 (MyrNT08 and/or MyrNT17).
All these features show a null intensity in the control samples (UnTr).
Features plots are sorted according to the descending order of their PC2 scores (in absolute value), from the left to the right.
Plots generated using MetaboAnalyst.

3.3.2. Xenometabolome features prioritization by time-series Heatmap

As mentioned previously, PCA represented a good tool for an overview understanding of xenometabolome evolution through the time. However, only major molecular traces were revealed by this model. In-depth xenometabolome discovery required different statistical tools dedicated to prioritize and filtrate molecular features of the detected volatile xenometabolites. Therefore, a second different statistical analysis was exploited in this work: the time series-based Heatmap.

The time series-based Heatmap considers all features present in the data matrix, in order to visualize the evolution of their abundances over time. Moreover, it performs a Dendrogram-based clustering in order to group the correlated variables (molecular features), basing on their evolution profiles over the time, and their abundances in the different sample conditions. Thus, it facilitates the “fishing” of features according to their chemical/biochemical nature.

Results of the applied time series-based Heatmap are shown in Figure C.III 6. The Heatmap was applied on all of the kinetics time points (day 1 to day 38) for both untreated control soil and herbicide-spiked soil groups. In this Heatmap, samples were not clustered but sorted according to treatment condition as the first factor, and then according to time evolution as the

second. Features however, were clustered without *a priori*, according to the correlation of their abundances between the different samples.

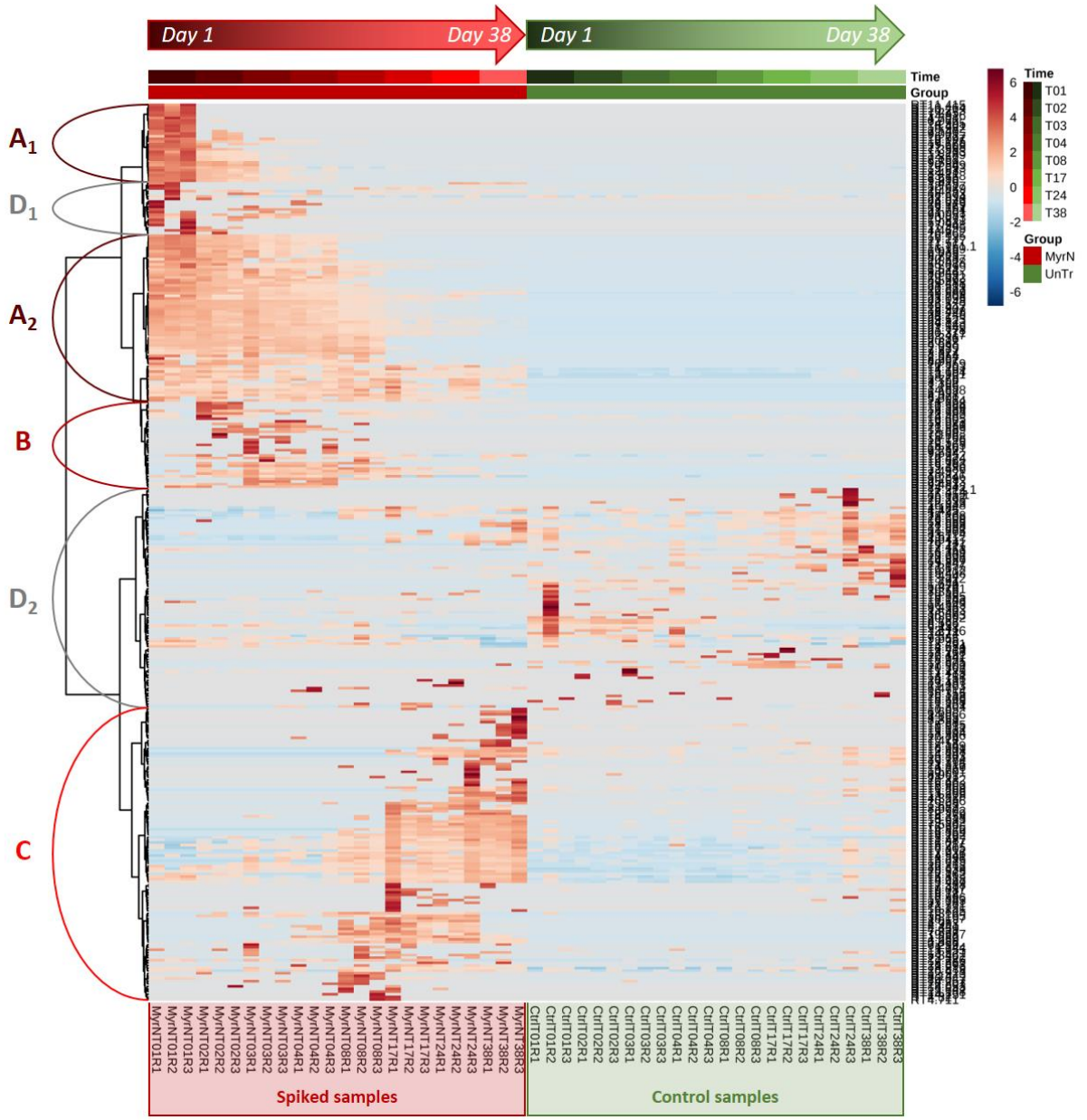


Figure C.III 6: Time series-based Heatmap. Clustering algorithm: Ward, distance measure: Euclidean. Plot generated using MetaboAnalyst.

The clustering of features led to the identification of 4 main zones of interest. Zone A, divided into two sub-zones, A_1 and A_2, consisted of compounds that were only present in the spiked samples. The majority of these features were at their highest level of abundance at day 1 and then started to decay over time. Thus, they are considered as components of the *Myrica gale* extract herbicide. The difference between the two A sub-zones was that in comparison to the A_2 sub-zone features, A_1 sub-zone features presented a higher intensity at the beginning of the kinetics tracking, and their decay over time was faster. A_2 sub-zone features however, had lower

intensities at the beginning of the tracking compared to A_1 sub-zone features. They decayed more slowly and some of these features were still present on day 38.

Zone B consisted of features that appeared at the middle of the kinetics tracking. Therefore, those features were considered as degradation by-products. Zone C features were also considered as degradation by-products. They appeared at the end of the kinetics tracking, however.

Zone D, also divided into two sub-zones, D_1 and D_2, represented features that were abundant in both control and spiked samples. Most of these features were considered as noise and artefacts as they showed a random dispersion of abundances between replicates. For sub-zone D_1, features were identified as random artefacts and noise issued from the complex xenometabolome profile. As this complexity is relatively higher in spiked samples at day 1 and day 2, this can explain the reason why this noise is higher in those samples, and less intense in the control samples. For sub-zone D_2 the most relevant of these features were examined by a fast putative annotation using NIST 14 library. All of these features were silicon-derivate compounds. Thus, they were considered as products of septum, fiber and column bleeding. These features were also found in blank injections, which confirmed this hypothesis. It is worth mentioning that the significant features of the D_2 sub-zone were more abundant in the untreated control soil samples. This can be explained by the possible fact that in the spiked samples, the adsorption sites of the SPME fiber were less “available” due to the presence of a rich volatile xenometabolome, bleeding compounds originating from the vial septum were thus less able to fixate on the fiber.

Afterwards, as important numbers of features were prioritized by the Heatmap, a verification procedure was performed in order to filter and remove the eventual hidden artefacts. This procedure was achived using the Multivariate Empirical Bayes Analysis of Variance (MEBA) approach for time series, based on the timecourse method [48], and designed for the comparison of temporal profiles across different conditions or groups of treatments. The results are shown in the Appendix III-B.

3.3.3. Xenometabolome kinetics tracking and putative compounds identification

All of the prioritized significant features, revealed by the Heatmap and verified by the MEBA, were manually tracked over time by integrating their GC-MS pseudo-spectra peak areas over all the RAW files. This was done in order to draw their time evolution curves according to the 38-day kinetics tracking. In addition, this manual tracking is recommended in order to

crosscheck the automatically generated results and to avoid any false positives that may occur due to the potential errors of the automated data preprocessing.

The manual tracking finally led to consider 101 features as relevant, including 22 features that were considered as degradation by-products according to the kinetics profiles/curves evolution over time. All of these kinetics profiles are shown in Appendix III-B. Two orthogonal tools were used for putative identification of compounds: the EI-MS fragmentation patterns search on NIST 14 spectral library, and the calculation of Kováts RI that were compared to RI values reported in the NIST library. Kováts RI calculation was performed following the method of Lucero *et al.* [49]. Out of the 101 relevant features, 96 compounds including 20 degradation by-products, representing 99.83 % of the total TIC area after blank subtraction were putatively identified on the levels "2" and "3" of identification confidence according to Sumner *et al.* [50]. The most abundant compounds and all identified degradation by-products are shown in Table C.III 1. Detailed annotations of all the 101 prioritized features are summarized in Table A.III-A 6 (Appendix III-A). Furthermore, out of the 96 annotated compounds, 33 were reported in the literature as *Myrica gale* essential oil components [28–30]. All of these 33 compounds found in the literature were abundant at day 1 after spiking, representing 67.82 % of the total TIC area after blank subtraction. Meanwhile 63 compounds (47 herbicide components and 16 degradation by-products) are identified for the first time as compounds originating from the *Myrica gale* extract. Figure C.III 7 shows kinetic profiles of the 6 major compounds identified: Eucalyptol, L-terpinen-4-ol, α-Terpineol, α-Terpineol acetate, 3,7(11)-Selinadiene and Germacrone. The rest of the xenometabolites were predominantly terpenes, aromatic and aliphatic esters, alcohol and ketones (Table A.III-A 6 – Appendix III-A).

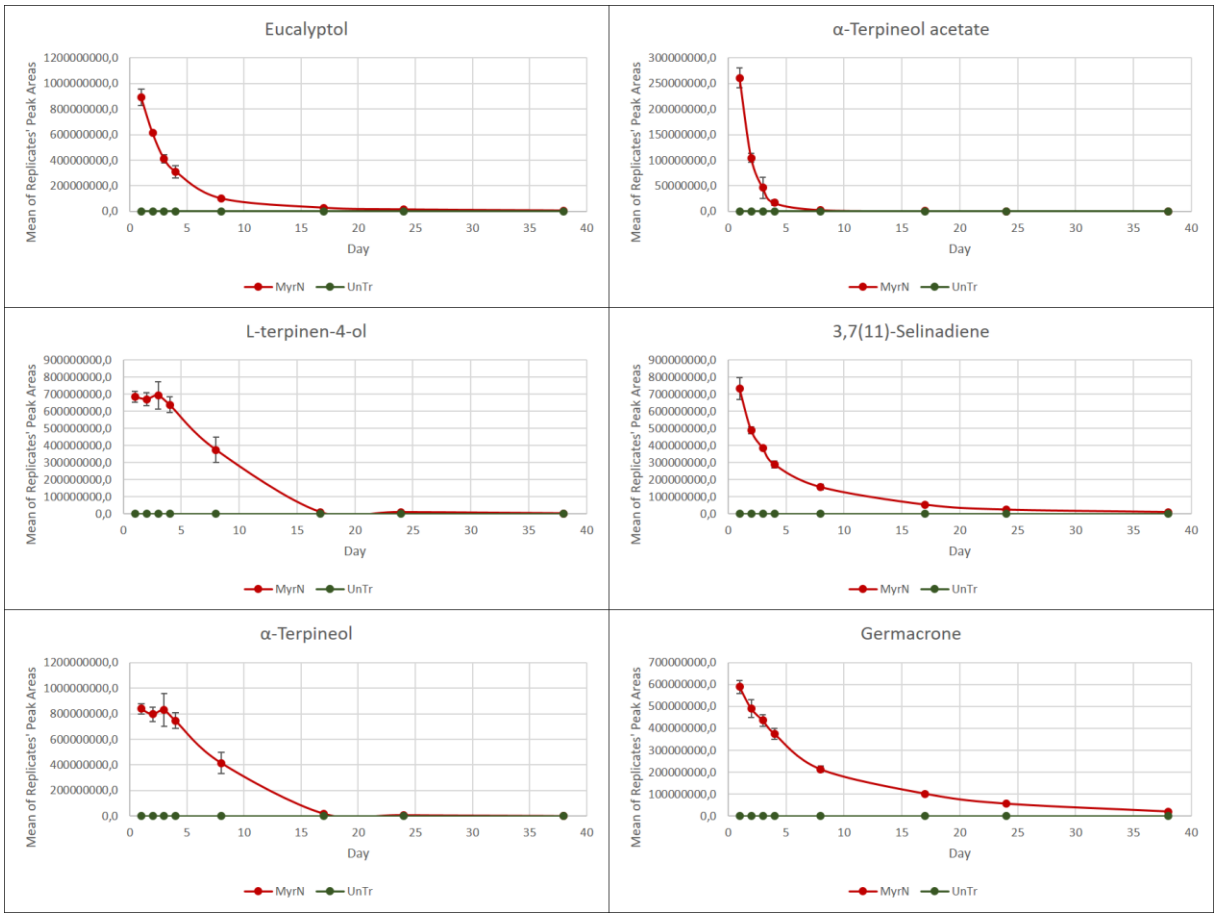


Figure C.III 7: Kinetic profiles of the 38-day degradation tracking of the 6 major compounds detected. The Peak Areas represent the sum of EICs of the major EI-fragments/ions.

Otherwise, several of the identified degradation by-products could be hypothetically related to the detected *Myrica gale* extract compounds. For instance, Figure C.III 8 shows the kinetics profiles of 2,3-Dehydro-1,8-cineole, Camphor, and Camphene hydrate, that are hypothetically the by-products of Eucalyptol and Borneol after oxidation, and Camphene after hydration, respectively.

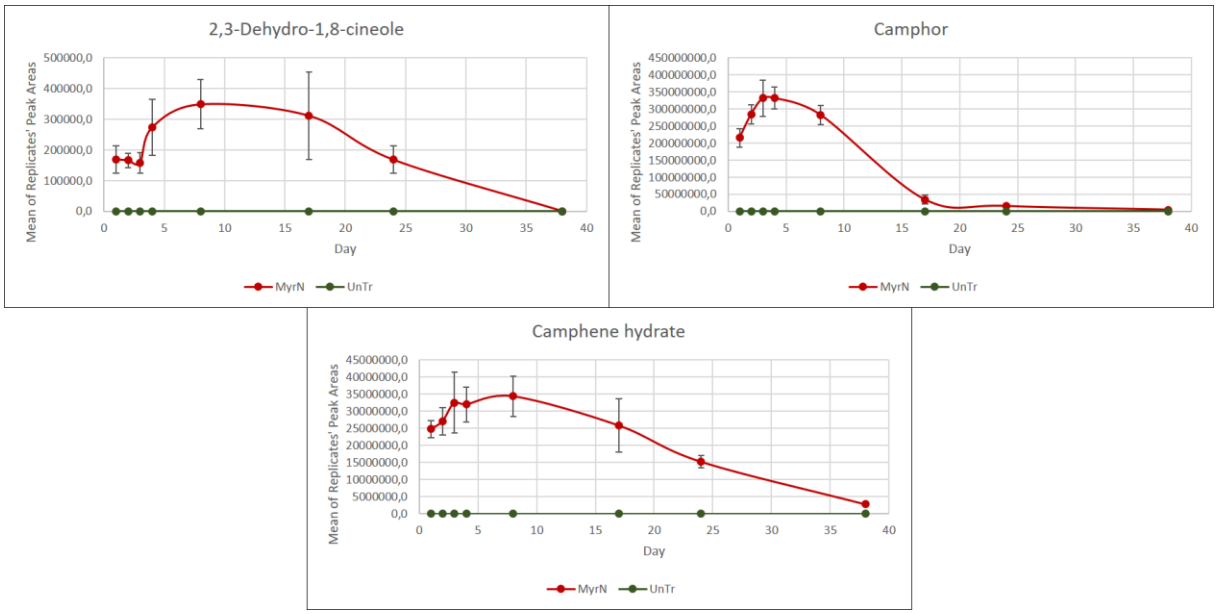


Figure C.III 8: Kinetics profiles of the 38-day degradation tracking of the 3 volatile degradation by-products: 2,3-Dehydro-1,8-cineole, Camphor, and Camphene hydrate.
The Peak Areas represent the sum of EICs of the major EI-fragments/ions.

It is worth to mention that several degradation by-products (14 out of 22, representing 5.49 % of the total TIC area after blank subtraction), were detected at the day 1 after spiking. This can be explained by three different hypotheses: i) the degradation of their parents was very fast so they started to appear after 1 day of the application of the herbicide, ii) they were already present in the applied herbicide mixture due to a slight degradation of their parents during the extraction and/or the stock of the *Myrica gale* extract, iii) these compounds are not only degradation by-products but also essential components of the *Myrica gale* extract. This last case can be considered for the Camphor and the Camphene hydrate that were reported in the literature as components of the *Myrica gale* essential oil, as well as their hypothetic parents, *i.e.* Borneol and Camphene, respectively.

Thereby, this fast putative identification, based on fast library search for EI-MS fragmentation spectra and the Kováts RI calculations, presents one of the main advantages of this approach. Indeed, it allowed studying the complex mixture of the emerging natural herbicide, where several of its unknown components and TPs could be putatively identified. The EI fragmentation patterns and their reproducibility allowed the fast annotation of several of these unknown xenometabolites by a simple spectral library search, despite the low resolution of Quadrupole mass analyzer in measuring ions' *m/z* values. Kováts RI calculations assured higher identification confidence by providing an additional and orthogonal tool for metabolites characterization.

Table C.III 1: Summary of putative identifications of the most relevant features (herbicide xenometabolites with EICs area/total TIC area ≥ 1 %, and identified degradation by-products).
†: The given-code represents the retention time of the compound (in minutes) preceded by the Retention Time "RT" abbreviation.
‡: If MF ≥ 700, and Δ between experimental and NIST RI ≤ 10, the considered level of identification confidence is "2". If MF < 700, or Δ between experimental and NIST RI > 10, the considered level of identification confidence is "3" (levels defined by Sumner *et al.* 2007 [50]).
⏚: The percentage of the "sum of major fragments EICs area/total TIC area" ratio, calculated at day 1 after spiking.
N/A: Not Available. N/C: Not Calculated. The relative intensity was not calculated for degradation by-products that were not detected at day 1.

Compound given-code†	**Putative identity‡ (level 2 or 3 of identification confidence)**	**MF**	**RI (Experimental)**	**RI (NIST)**	**Relative intensity (%)⏚**	**Reference**
Myrica gale methanolic extract components						
RT5.492	p-Cymene	881	1026	1025 ± 2	1.56	[28–30]
RT5.501	o-Cymene	902	1029	1022 ± 2	1.46	N/A
RT5.599	Eucalyptol	910	1036	1032 ± 2	11.09	[28–30]
RT8.421	Borneol	903	1172	1166 ± 7	1.25	[30]
RT8.688	L-terpinen-4-ol	934	1179	1182 ± 0	7.09	[29,30]
RT9.090	α-Terpineol	913	1193	1189 ± 2	7.94	[29,30]
RT11.233	Methyl hydrocinnamate	928	1267	1279 ± 2	1.28	N/A
RT11.751	2-Undecanone	922	1285	1294 ± 2	1.60	[28]
RT13.338	α-Terpineol acetate	934	1342	1350 ± 3	4.75	[30]
RT17.570	Aromadendrene, dehydro-	784	1466	1464 ± 1	5.74	N/A
RT18.448	(+)-β-Selinene	853	1488	1486 ± 3	2.49	[30]
RT18.573	α-Selinene	904	1492	1494 ± 3	1.70	[30]
RT19.299	β-Cadinene	850	1516	1518 ± 10	3.90	[29]
RT20.047	γ-Selinene	907	1541	1544 ± N/A	8.30	N/A
RT20.170	3,7(11)-Selinadiene	913	1547	1542 ± 3	5.96	[28]
RT21.778	Aristolene epoxide	817	1585	N/A	2.84	N/A
RT22.148	cis-β-Elemenone	905	1595	1593 ± 3	1.42	[29,30]
RT23.095	1,4-Benzenedipropanol, α,α',γ,γ,γ',γ'-hexamethyl-	710	1622	N/A	1.81	N/A
RT25.500	Germacrone	930	1691	1693 ± 3	5.81	[28–30]
Degradation by-products						
RT2.425	Methyl isovalerate	831	777	773 ± 5	0.04	N/A
RT3.050	Tyranton	834	843	838 ± 8	<0.01	N/A
RT4.259	Butanoic acid, 2,2-dimethyl-3-oxo-, methyl ester	800	946	936	N/C	N/A
RT4.333	Methyl 2-methylhexanoate	770	952	953 ± 2	N/C	N/A
RT4.441	Camphene	932	957	952 ± 2	0.02	[29,30]
RT4.459	β-Pinene	678	960	979 ± 2	<0.01	[29,30]
RT4.953	2,3-Dehydro-1,8-cineole	819	991	991 ± 2	<0.01	N/A
RT6.690	Methyl 2-propylheptanoate	720	1096	1155 ± N/A	N/C	N/A
RT6.885	3-Acetyl-2,5-dimethylfuran	590	1101	1099 ± 4	0.01	N/A
RT7.536	Methyl octanoate	584	1133	1126 ± 2	N/C	N/A
RT7.877	(+)-Camphor	932	1148	1143 ± 9	2.80	[28]
RT8.073	Camphene hydrate	870	1155	1148 ± 2	0.31	[28]
RT8.140	3-Isopropyl-2-methylcyclopentanone	715	1157	1174 ± N/A	N/C	N/A
RT8.267_2	cis-p-Menthan-3-one	809	1164	1164 ± 6	0.35	N/A
RT8.677	2(3H)-Benzofuranone, hexahydro-3a,7a-dimethyl-, cis-	729	1182	N/A	N/C	N/A
RT9.209	Tetrahydrocarvone	856	1200	1208 ± N/A	0.01	N/A
RT11.495	8,9-Dehydrothymol methyl ether	733	1281	1247 ± N/A	0.02	N/A
RT12.486	5-Methoxy-4,4,6-trimethyl-7-oxabicyclo[4.1.0]heptan-2-one	645	1314	N/A	N/C	N/A
RT16.983	Selinan	621	1450	1476 ± 12	0.01	N/A
RT20.371	3,5,11-Eudesmatriene	859	1547	1495 ± N/A	1.92	N/A

3.3.4. Dissipation assessment by Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA)

Regarding limitations of classic concepts for environmental fate assessment of complex biopesticides, the targeted tracking is not applicable for the present study as described previously. Thus, the comparison of volatile metabolic profiles of both spiked samples and untreated control samples can be considered as an alternative concept to determine the dissipation of volatile compounds of the studied bioherbicide. The total dissipation is considered when the difference between the volatile metabolic profiles of the compared groups is no more significant. Therefore, the choice of the comparative statistical approach should lay on its ability to reveal the minor differences between the compared profiles. In addition, the significance of those minor differences should be also assessed to avoid the misleading conclusions or the loss of information.

The PCA model is a suitable tool for a holistic overview of the acquired data, as for revealing the major differences. However, minor differences will be hidden and it is difficult to determine them with this descriptive multivariate analysis. Thus, a discriminant analysis is needed for this aim. In this work, Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) [51,52], and its Cross-Validation (CV) test, were considered to quarry and validate the significance of minor differences that are still present after 38 days of kinetics tracking between spiked samples and control samples.

PCA, OPLS-DA and the CV test of OPLS-DA were applied to compare the volatile metabolic profiles of both spiked soil and untreated control soil samples at day 38 after spiking. This in order to check if the total dissipation of the herbicide has occurred. First, the PCA showed a discrimination between the two conditions according to both PC1 and PC2, explaining 84.9 % and 13.2 % of variations, respectively (Figure A.III-A 10 – Appendix III-A). However, PCA loadings showed that the significance of the two major discriminant features of the PC1 was unreliable, as an important intra-group variation had been noticed (Figure A.III-A 11 – Appendix III-A). Contrariwise, the three major features of the PC2 showed a significant difference between groups (Figure A.III-A 11 – Appendix III-A). Two of those features were considered as persistent xenometabolites as they were not detected in the untreated control samples (RT5.501: o-Cymene and RT1.766: Methyl benzyl sulfoxide). The third feature showing a higher abundance in the untreated control samples was identified as a silicon derivate compound issued from bleeding. It was also detected in the blank injections.

Results of PCA led to conclude that in this descriptive unsupervised multivariate analysis, minor significant discriminations between groups risk to be hidden by the random contaminations and artefacts. Therefore, the OPLS-DA and its CV test were applied as explicative supervised multivariate analyses, in order to reveal significant differences related to the two pre-defined groups (untreated control samples *vs.* spiked samples). These differences in variables (molecular features) will be revealed by the predictive component (p) of the OPLS-DA. Moreover, the significance of these features will be assessed by introducing the confidence dimension represented by the orthogonal (o) component of the OPLS-DA.

As described in Figure A.III-A 12 (Appendix III-A), the T score shows that the predictive (p) component explains 55.3 % of variations between the volatile profiles of spiked soil and the control soil. The orthogonal component, that explains systematic "groups-independent" variations, represents 16.3 % of variations (Orthogonal T score). Thus, the CV test was performed to assess the significance of "between-groups" and "groups-independent" systematic discriminations. The "between-groups" discrimination is assessed by calculating the correlation "R2Y" of the two groups of samples to the variation explained by the p component, and by the prediction/significance estimated by the Q2 value. The "groups-independent" systematic variations are also assessed by calculating the R2Y and Q2 applied to the orthogonal component.

The CV test results shown in Appendix III-A (Figure A.III-A 13) were the following: for the p component, R2Y and Q2 were 98.7 % and 92.3 %, respectively. For the o component, R2Y and Q2 were 1.25 % and 1.94 %, respectively. These results show both R2Y and Q2 above 90 % (for p), with R2Y higher than R2X and Q2, meaning that the OPLS-DA model is valid. Thus, the discrimination between the two defined groups of samples is significant. In addition, there is a high confidence in the significance of discrimination as the "groups-independent" systematic variations were not significant (R2Y and Q2 below 50 % for o, with R2Y lower than R2X and Q2).

Hence, this result means that at day 38 after spiking, the total dissipation of the volatile xenometabolome was not reached. Therefore, to reveal the persistent xenometabolites, the OPLS-DA S-Plot can be used as shown in Figure C.III 9. The S-Plot showed several persistent herbicide compounds that were still significantly abundant in the volatile profiles of spiked soil at day 38, *e.g.* RT1.766: Methyl benzyl sulfoxide, RT5.501: o-Cymene, RT25.500: Germacrone, RT20.371: 3,5,11-Eudesmatriene, RT4.333: Methyl 2-methylhexanoate,

RT19.447: Calamenene, RT8.073: Camphene hydrate, RT22.148: cis-β-Elemenone. The last 6 mentioned features were difficult to reveal using the PCA loading plot. In addition, the kinetics curves proved coherent results with the S-Plot by showing the persistence of these features after 38 days of herbicide application (Appendix III-B).

Another advantage of the OPLS-DA S-Plot was its ability to explain the high risk/insignificance of artefacts and contamination features (RT2.671 and RT4.641) previously revealed by the PCA loadings, despite their high contribution in discrimination between groups. This is thanks to the confidence/reliability dimension represented by the p(corr)[1] axis, as explained in Figure C.III 9.

It is worth mentioning that the determination of the total dissipation time of the *Myrica gale* extract herbicide necessitated a longer kinetics study. This however was not in the scope of the present work.

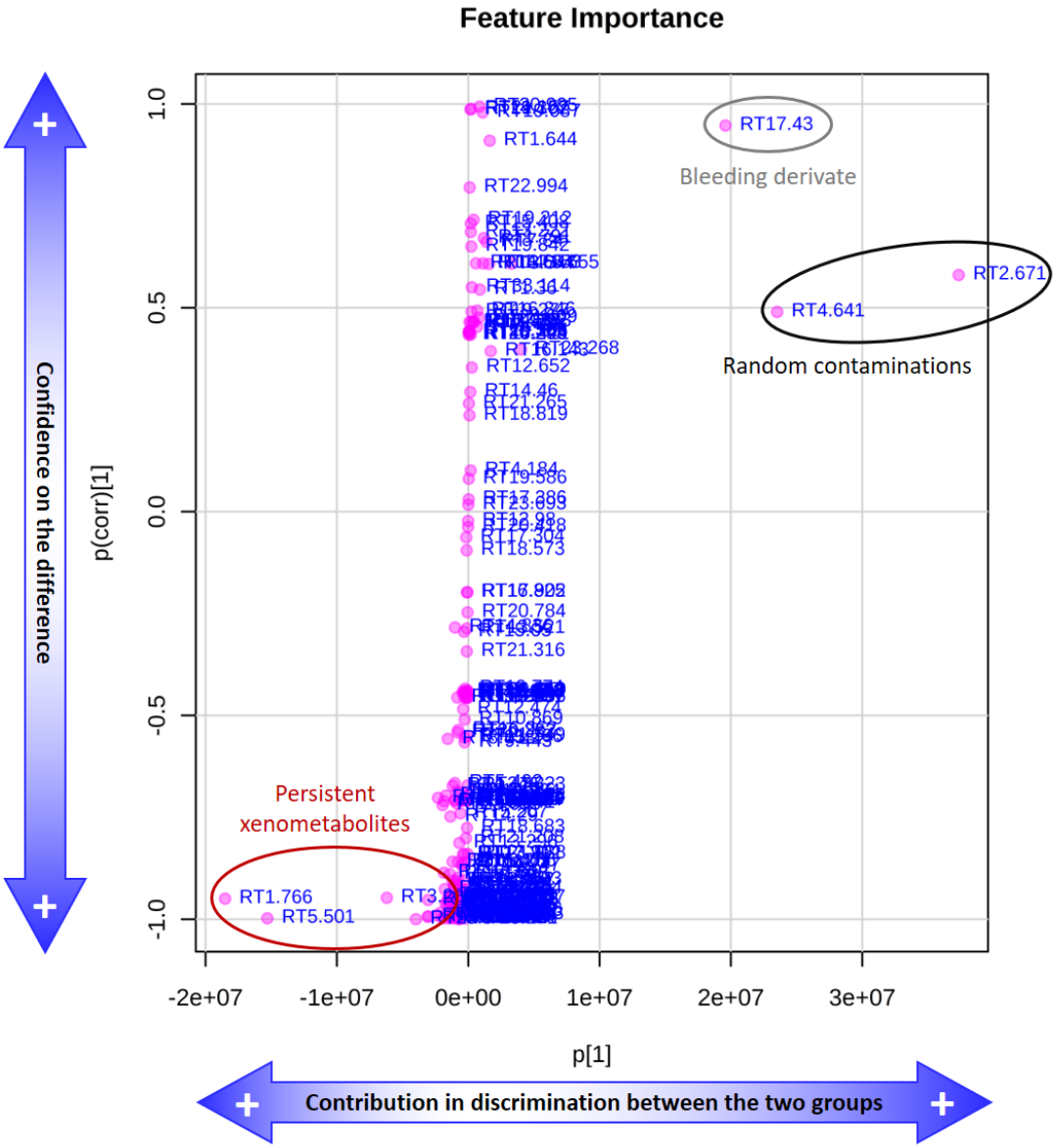


Figure C.III 9: OPLS-DA S-Plot showing the markers of discrimination between the two defined groups of soil samples. The further the feature from the 0 of the p[1] axis, the higher the magnitude of its variation between the two groups. The further the feature from the 0 of the p(corr)[1] axis, the lower its intra-group variation, thus the higher the confidence of its variation significance [53–55].
Plot generated using MetaboAnalyst.

3.4. Reproducibility and sensitivity assessment

3.4.1. Reproducibility

The analytical repeatability and reproducibility of the HS-SPME-GC-MS method were assessed by selecting several important compounds to calculate their Retention Time (RT) and Peak Area (PA) deviations. The 5 chosen compounds were distributed on the chromatogram RT range (Table C.III 2). "Inter-samples" Peak Area Relative Standard Deviation (RSD) and Retention Time Standard Deviation (SD) were calculated using the 3 biological replicates at the same day (day 1). The highest PA RSD was 7.75 % for the 3,7(11)-Selinadiene, and the highest RT SD was 0.43 s for Germacrone (Table C.III 2). This proved that the method was highly reproducible. "Inter-days" Retention Time Standard Deviation (SD) was also calculated for the selected compounds using the same sample that was injected 8 times with the following time gaps: 1, 2, 3, 4, 8, 17, 24, 38 days. The highest SD was for the Germacrone with 3.44 s of deviation (Table C.III 2). "Inter-days" PA variation, however, was not assessed due to the difficulty of the application of Internal Standards (IS) with this type of approach. In fact, as the sample is a living system analyzed for several time points for a period of 38 days, a degradation of IS may occur during the experiment.

Table C.III 2: "Inter-samples" and "inter-days" variations of Peak Area (PA) and Retention Time (RT) over the experiment. "Inter-samples" PA RSD and "inter-samples" RT SD were calculated using the 3 biological replicates. Inters-days" RT SD was calculated after the same sample was injected 8 times with the following time gaps: 1, 2, 3, 4, 8, 17, 24, 38 days.

Compound	**Retention Time (min)**	**PA RSD (%) "Inter-samples" *(n = 3)***	**RT SD (sec) "Inter-samples" *(n = 3)***	**RT SD (sec) "Inter-days" *(n = 8)***
Eucalyptol	5.60	6.79	0.25	1.61
L-terpinen-4-ol	8.69	4.20	0.18	1.00
α-Gurjunene	17.17	5.34	0.36	0.92
3,7(11)-Selinadiene	20.17	7.75	0.18	2.53
Germacrone	25.50	2.86	0.43	3.44

3.4.2. Sensitivity

As the current study is suggesting an untargeted metabolomics-based approach, classic protocols for targeted method validation are not reasonable (*e.g.* absolute quantification of targeted compounds using reference standards and calibration curves). Thus, a different concept was applied to assess the sensitivity at day 1 of the kinetics study, based on a comparative approximation related to herbicide field dose. Untreated control samples were compared to each dose level of the spiked soil samples (described in Section 2.4.). Comparisons were performed

using 3 indicators: the Total TIC area integration, the number of the detected molecular features, and by applying OPLS-DA Cross-Validation tests using a data matrix generated after raw data of dose response curve were processed (using the same processing method described in Section 2.6.). Results in Table C.III 3 show that the method is able to discriminate between the 2 conditions (spiked *vs.* untreated) from 20-times and until 10^{-1}-time the field dose, as significant differences between the compared conditions are observed for Total TIC areas and for numbers of the detected molecular features (Welch Two Sample T-test p-Values < 5 %), and as a reliable predictivity of the OPLS-DA model is observed for these dose levels (Q2 of the p1 > 50 %). Therefore, a minimum of sensitivity relative to herbicide dose is estimated between 10^{-1}-time and 10^{-2}-time the field dose at day 1 of the kinetics study.

Table C.III 3: Summary of OPLS-DA CV test and Welch Two Sample T-test results.

Applied dose (the field dose)	**p1**			**o1**			**T-test (-Log$_{10}$[p-Value])**	
	R2X	***R2Y***	***Q2***	***R2X***	***R2Y***	***Q2***	***Total TIC area***	***Number of Molecular Features***
20-times	70.30	99.90	98.30	07.68	00.09	00.28	*** 3.67	*** 5.48
10-times	69.40	100.00	98.30	07.64	00.02	00.45	*** 3.93	*** 6.54
1-time	56.60	99.80	94.80	12.70	00.21	00.97	** 3.15	*** 3.49
10^{-1}-time	44.60	99.40	88.50	16.30	00.55	03.76	*** 3.51	** 2.57
10^{-2}-time	26.20	84.80	48.30	24.30	14.00	11.10	0.70	1.19
10^{-3}-time	26.40	75.00	43.60	24.50	24.30	00.43	0.60	0.42

3.5. Sample design: a living system after 8 extraction operations

As previously mentioned, the sample design described in Section 2.3. was optimized in order to create a “living system”, such that the same prepared samples (soil vials) could then be tracked by several kinetics time point analyses, as the HS-SPME extraction is a non-destructive method.

After the application of 8 extractions on each vial/sample during the 38-day kinetics study, green plants were observed on top of the soil layer of untreated control samples. The development of this plant layer was progressive during the kinetics study and was even observed 44 days after the end of the kinetics tracking as shown in Figure C.III 10. This indicates that the implemented sample design and the optimized HS-SPME extraction was successful in providing appropriate conditions to sustain a living micro-ecosystem.

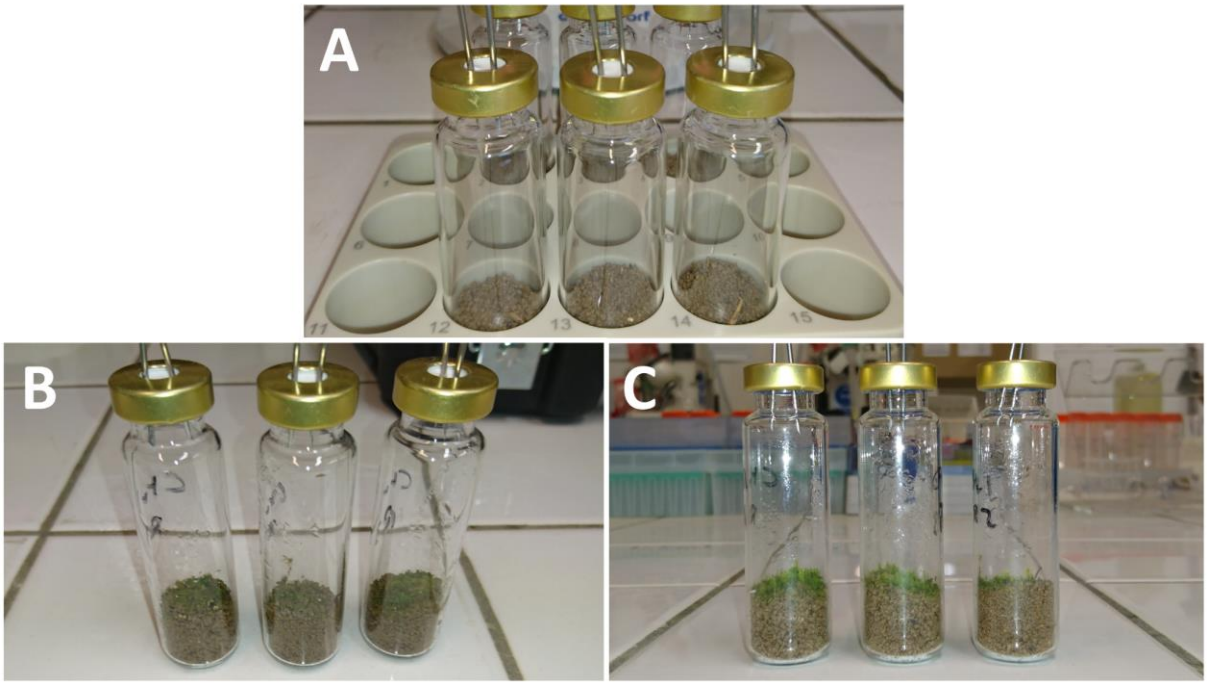


Figure C.III 10: The evolution of the untreated control soil vials/microcosms during the experiment.
Photo A was taken before the kinetics tracking began. Photo B was taken at the end of the kinetics tracking (at day 38, after analyses). Photos C was taken 44 days after the last time point was analyzed (*i.e.* after 82 days of the beginning of the kinetics tracking).

4. Conclusions

The present work aimed to introduce a novel HS-SPME-GC-MS-based untargeted metabolomics approach dedicated to study the environmental fate of complex biopesticides. The approach was developed and applied to study the volatile residues of the *Myrica gale* methanolic extract; an emerging natural herbicide applied to soil, consisting of a complex mixture of identified and non-identified compounds. The developed analytical method was proved reliable in the detection of a rich volatile profile originating from the herbicide, with a low matrix background and a significant robustness. This allowed the fast putative identification of 96 xenometabolites including 33 compounds reported in the literature, 47 compounds identified for the first time as *Myrica gale* extract components, and 16 new degradation by-products, by a 38-day kinetics tracking experiment. A comparison of herbicide-spiked and untreated control soil samples over time demonstrated the advantages of applying the untargeted metabolomics and its statistical tools as an alternative concept for complex pesticides study. The evolution of the herbicide volatile xenometabolome over time can be visualized and explained by using the PCA. The time-series Heatmap method is a suitable tool for prioritizing the relevant xenometabolites and sorting them according to their temporal evolution in the different groups of samples. This is done in order to characterize and identify new xenometabolites and TPs, which can then help to better understand the environmental fate of the herbicide, as well for assessing its potential risk and toxicity on the health and the environment. The OPLS-DA and its CV test provided a sensitive and confident determination of minor discriminations between the different groups of samples in order to assess the total dissipation of the herbicide volatile xenometabolome and method sensitivity.

The developed approach successfully revealed all of the significant results and conclusions through an analysis of only 6 environmental samples that were not destroyed throughout the course of the study. Thus, this non-destructive green automated method has now been shown to be capable of cost-effective high throughput analyses. Further analytical and technical developments can be performed in order to expand its potential application in environmental fate studies and emerging pesticides research.

Acknowledgments

Authors would like to acknowledge Jeanine Almany (École Pratique des Hautes Études) for providing English language editing (as well as constructing comments) which improved the manuscript. Acknowledgments for Dr. Nicolas Le Yondre (Université de Rennes 1), Dr. Chandrashekhar Patil (Université de Perpignan Via Domitia), Anaïs Amiot, M.Sc. (S.A.S. AkiNaO), and Mathieu Lazarus, M.Sc. (Université de Perpignan Via Domitia) for their scientific and technical advices. Authors also acknowledge two anonymous reviewers for their constructive reviews and critical comments that helped improving the current work.

This work was supported by the French National Research Agency (ANR) under TRICETOX project (ANR-13-CESA-0002), and the European Regional Development Fund (ERDF) under the Interreg POCTEFA PALVIP project (POCTEFA 2014-2020). The funding institutions had no role in the experimental design, the data processing, or in writing and reviewing the manuscript.

Ph.D. fellowship grant was awarded to HG by the French Ministry of Higher Education, Research and Innovation (MESRI), *via* the Doctoral School ED 305 "Energie et Environnement" (Université de Perpignan Via Domitia).

The HS-SPME-GC-MS method developments and analyses had been performed using the Biodiversité et Biotechnologies Marines (Bio2Mar) facilities – Métabolites Secondaires Xénobiotiques Métabolomique Environnementale (MSXM) platform at the Université de Perpignan Via Domitia (http://bio2mar.obs-banyuls.fr/).

List of Abbreviations

ANDI: Analytical Data Interchange

BA: Biocontrol Agent

CAR: Carboxen

CAS: Chemical Abstracts Service

CEC: Cation Exchange Capacity

CV: Cross-Validation

DVB: Divinylbenzene

EI: Electron Impact

EIC: Extracted Ion Chromatogram

EMF: Environmental Metabolic Footprinting

FWHM: Full Width at Half Maximum

GC: Gas Chromatography

HPLC: High Performance Liquid Chromatography

HS: Headspace

IS: Internal Standard

LC: Liquid Chromatography

MEBA: Multivariate Empirical Bayes Analysis of Variance

MF: Matching Factor

MS: Mass Spectrometry

NetCDF: Network Common Data Form

OECD: Organization for Economic Co-operation and Development

OPLS-DA: Orthogonal Projections to Latent Structures Discriminant Analysis

PA: Peak Area

PA: Polyacrylate

PC1: First Principal Component

PC2: Second Principal Component

PCA: Principal Component Analysis

PDMS: Polydimethylsiloxane

RH: Relative Humidity

RI: Kováts Retention Index

RSD: Relative Standard Deviation

RT: Retention Time

SD: Standard Deviation

SLE: Solid-Liquid Extraction

SPME: Solid Phase Microextraction

TIC: Total Ion Chromatogram

TP: Transformation Product

References

[1] R. McKinlay, J.A. Plant, J.N.B. Bell, N. Voulvoulis, Endocrine disrupting pesticides: Implications for risk assessment, Environment International. 34 (2008) 168–183. https://doi.org/10.1016/j.envint.2007.07.013.

[2] J. Dich, S.H. Zahm, A. Hanberg, H.-O. Adami, Pesticides and cancer, Cancer Causes & Control. 8 (1997) 420–443. https://doi.org/10.1023/A:1018413522959.

[3] B.N. Ames, L.S. Gold, Pesticides, Risk, and Applesauce, Science. 244 (1989) 755–757. https://www.jstor.org/stable/1703501 (accessed October 18, 2020).

[4] M. Gavrilescu, Fate of Pesticides in the Environment and its Bioremediation, Eng. Life Sci. 5 (2005) 497–526. https://doi.org/10.1002/elsc.200520098.

[5] H.M.G. van der Werf, Assessing the impact of pesticides on the environment, Agriculture, Ecosystems & Environment. 60 (1996) 81–96. https://doi.org/10.1016/S0167-8809(96)01096-1.

[6] C. Bertrand, C. Prigent-Combaret, A. Gonzales-Coloma, Chemistry, activity, and impact of plant biocontrol products, Environ Sci Pollut Res. 25 (2018) 29773–29774. https://doi.org/10.1007/s11356-018-3209-2.

[7] European Food Safety Authority, EFSA Guidance Document for evaluating laboratory and field dissipation studies to obtain DegT50 values of active substances of plant protection products and transformation products of these active substances in soil, EFSA Journal. 12 (2014) 3662. https://doi.org/10.2903/j.efsa.2014.3662.

[8] K.A. Aliferis, M. Chrysayi-Tokousbalides, Metabolomics in pesticide research and development: review and future perspectives, Metabolomics. 7 (2011) 35–53. https://doi.org/10.1007/s11306-010-0231-x.

[9] M.R. Viant, U. Sommer, Mass spectrometry based environmental metabolomics: a primer and review, Metabolomics. 9 (2013) 144–158. https://doi.org/10.1007/s11306-012-0412-x.

[10] C. Patil, C. Calvayrac, Y. Zhou, S. Romdhane, M.-V. Salvia, J.-F. Cooper, F.E. Dayan, C. Bertrand, Environmental Metabolic Footprinting: A novel application to study the impact of a natural and a synthetic β-triketone herbicide in soil, Science of The Total Environment. 566–567 (2016) 552–558. https://doi.org/10.1016/j.scitotenv.2016.05.071.

[11] M.-V. Salvia, A. Ben Jrad, D. Raviglione, Y. Zhou, C. Bertrand, Environmental Metabolic Footprinting (EMF) vs. half-life: a new and integrative proxy for the discrimination between control and pesticides exposed sediments in order to further characterise pesticides' environmental impact, Environ Sci Pollut Res. 25 (2018) 29841–29847. https://doi.org/10.1007/s11356-017-9600-6.

[12] OECD, Test No. 307: Aerobic and Anaerobic Transformation in Soil, 2002. https://doi.org/10.1787/9789264070509-en.

[13] OECD, Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems, 2002. https://doi.org/10.1787/9789264070523-en.

[14] J. Pawliszyn, Handbook of Solid Phase Microextraction, First Edition, Elsevier, 2012. https://doi.org/10.1016/C2011-0-04297-7.

[15] J. Pawliszyn, Theory of Solid-Phase Microextraction, Journal of Chromatographic Science. 38 (2000) 270–278. https://doi.org/10.1093/chromsci/38.7.270.

[16] R.P. Belardi, J.B. Pawliszyn, The Application of Chemically Modified Fused Silica Fibers in the Extraction of Organics from Water Matrix Samples and their Rapid Transfer to Capillary Columns, Water Quality Research Journal. 24 (1989) 179–191. https://doi.org/10.2166/wqrj.1989.010.

[17] N. Reyes-Garcés, E. Gionfriddo, G.A. Gómez-Ríos, Md.N. Alam, E. Boyacı, B. Bojko, V. Singh, J. Grandy, J. Pawliszyn, Advances in Solid Phase Microextraction and Perspective on Future Directions, Analytical Chemistry. 90 (2018) 302–360. https://doi.org/10.1021/acs.analchem.7b04502.

[18] M. Llompart, M. Celeiro, C. García-Jares, T. Dagnac, Environmental applications of solid-phase microextraction, TrAC Trends in Analytical Chemistry. 112 (2019) 1–12. https://doi.org/10.1016/j.trac.2018.12.020.

[19] D. Liang, W. Liu, R. Raza, Y. Bai, H. Liu, Applications of solid-phase micro-extraction with mass spectrometry in pesticide analysis, Journal of Separation Science. 42 (2019) 330–341. https://doi.org/10.1002/jssc.201800804.

[20] J. Beltran, F.J. López, F. Hernández, Solid-phase microextraction in pesticide residue analysis, Journal of Chromatography A. 885 (2000) 389–404. https://doi.org/10.1016/S0021-9673(00)00142-4.

[21] J.L. Martínez Vidal, P. Plaza-Bolaños, R. Romero-González, A. Garrido Frenich, Determination of pesticide transformation products: A review of extraction and detection methods, Journal of Chromatography A. 1216 (2009) 6767–6788. https://doi.org/10.1016/j.chroma.2009.08.013.

[22] V. Andreu, Y. Picó, Determination of pesticides and their degradation products in soil: critical review and comparison of methods, TrAC Trends in Analytical Chemistry. 23 (2004) 772–789. https://doi.org/10.1016/j.trac.2004.07.008.

[23] M. Bedair, L.W. Sumner, Current and emerging mass-spectrometry technologies for metabolomics, TrAC Trends in Analytical Chemistry. 27 (2008) 238–250. https://doi.org/10.1016/j.trac.2008.01.006.

[24] L.S. Ettre, The Kováts Retention Index System, Anal. Chem. 36 (1964) 31A-41A. https://doi.org/10.1021/ac60214a727.

[25] J. Popovici, C. Bertrand, D. Jacquemoud, F. Bellvert, M.P. Fernandez, G. Comte, F. Piola, An Allelochemical from Myrica gale with Strong Phytotoxic Activity against Highly Invasive Fallopia x bohemica Taxa, Molecules. 16 (2011) 2323–2333. https://doi.org/10.3390/molecules16032323.

[26] J. Popovici, C. Bertrand, G. Comte, Use of a Myrica gale plant for producing a herbicide agent, US008734858B2, 2014. https://patents.google.com/patent/US8734858B2/en (accessed October 18, 2020).

[27] J. Popovici, G. Comte, E. Bagnarol, N. Alloisio, P. Fournier, F. Bellvert, C. Bertrand, M.P. Fernandez, Differential Effects of Rare Specific Flavonoids on Compatible and Incompatible Strains in the Myrica gale-Frankia Actinorhizal Symbiosis, Applied and Environmental Microbiology. 76 (2010) 2451–2460. https://doi.org/10.1128/AEM.02667-09.

[28] J. Popovici, C. Bertrand, E. Bagnarol, M.P. Fernandez, G. Comte, Chemical composition of essential oil and headspace-solid microextracts from fruits of *Myrica gale* L. and antifungal activity, Natural Product Research. 22 (2008) 1024–1032. https://doi.org/10.1080/14786410802055568.

[29] K.P. Svoboda, A. Inglis, J. Hampson, B. Galambosi, Y. Asakawa, Biomass production, essential oil yield and composition of Myrica gale L. harvested from wild populations in Scotland and Finland, Flavour and Fragrance Journal. 13 (1998) 6. https://doi.org/10.1002/(SICI)1099-1026(199811/12)13:6<367::AID-FFJ724>3.0.CO;2-M.

[30] R.R. Carlton, P.G. Waterman, A.I. Gray, Variation of leaf gland volatile oil within a population of sweet gale (Myrica gale) (Myricaceae), Chemoecology. 3 (1992) 45–54. https://doi.org/10.1007/BF01261456.

[31] K. Oracz, A. Voegele, D. Tarkowská, D. Jacquemoud, V. Turečková, T. Urbanová, M. Strnad, E. Sliwinska, G. Leubner-Metzger, Myrigalone A Inhibits Lepidium sativum Seed Germination by Interference with Gibberellin Metabolism and Apoplastic Superoxide Production Required for Embryo Extension Growth and Endosperm Rupture, Plant and Cell Physiology. 53 (2012) 81–95. https://doi.org/10.1093/pcp/pcr124.

[32] A. Khaled, M. Sleiman, E. Darras, A. Trivella, C. Bertrand, N. Inguimbert, P. Goupil, C. Richard, Photodegradation of Myrigalone A, an Allelochemical from *Myrica gale* : Photoproducts and Effect of Terpenes, J. Agric. Food Chem. 67 (2019) 7258–7265. https://doi.org/10.1021/acs.jafc.9b01722.

[33] United States Natural Resources Conservation Service - Soil Science Division, Soil Survey Manual, revised, United States Department of Agriculture, 2017. https://books.google.fr/books?id=dieUtAEACAAJ (accessed October 18, 2020).

[34] S. Romdhane, M. Devers-Lamrani, J. Beguet, C. Bertrand, C. Calvayrac, M.-V. Salvia, A.B. Jrad, F.E. Dayan, A. Spor, L. Barthelmebs, F. Martin-Laurent, Assessment of the ecotoxicological impact of natural and synthetic β-triketone herbicides on the diversity and activity of the soil bacterial community using omic approaches, Science of The Total Environment. 651 (2019) 241–249. https://doi.org/10.1016/j.scitotenv.2018.09.159.

[35] Galaxy Workflow4Metabolomics, Galaxy Workflow4Metabolomics. (n.d.). https://galaxy.workflow4metabolomics.org/ (accessed February 28, 2020).

[36] F. Giacomoni, G. Le Corguille, M. Monsoor, M. Landi, P. Pericard, M. Petera, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E.A. Thevenot, C. Caron, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics. 31 (2015) 1493–1495. https://doi.org/10.1093/bioinformatics/btu813.

[37] Y. Guitton, M. Tremblay-Franco, G. Le Corguillé, J.-F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E.A. Thévenot, Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics, The International Journal of Biochemistry & Cell Biology. 93 (2017) 89–101. https://doi.org/10.1016/j.biocel.2017.07.002.

[38] H. Ghosson, D. Raviglione, M.-V. Salvia, C. Bertrand, Online HS-SPME-GC-MS-based Untargeted Volatile Metabolomics for Studying Emerging Complex Biopesticides: a Proof of Concept - Data, MetaboLights, (2020). https://www.ebi.ac.uk/metabolights/MTBLS2026 (accessed August 21, 2020 – Status: In Curation).

[39] K. Haug, K. Cochrane, V.C. Nainala, M. Williams, J. Chang, K.V. Jayaseelan, C. O'Donovan, MetaboLights: a resource evolving in response to the needs of its scientific community, Nucleic Acids Research. 48 (2019) D440–D444. doi:10.1093/nar/gkz1019.

[40] R. Wehrens, G. Weingart, F. Mattivi, metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics, Journal of Chromatography B. 966 (2014) 109–116. https://doi.org/10.1016/j.jchromb.2014.02.051.

[41] H. Ghosson, D. Raviglione, M.-V. Salvia, C. Bertrand, Online HS-SPME-GC-MS-based Untargeted Volatile Metabolomics for Studying Emerging Complex Biopesticides: a Proof of Concept - W4M Workflow and Parameters, Galaxy Workflow4Metabolomics. (2020). https://workflow4metabolomics.usegalaxy.fr/u/hikmatghosson/w/s0003267020308473/ (accessed August 21, 2020).

[42] 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.

[43] MetaboAnalyst, MetaboAnalyst. (n.d.). https://www.metaboanalyst.ca/ (accessed February 28, 2020).

[44] J. Chong, D.S. Wishart, J. Xia, Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis, Current Protocols in Bioinformatics. 68 (2019) e86. https://doi.org/10.1002/cpbi.86.

[45] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Research. 46 (2018) W486–W494. https://doi.org/10.1093/nar/gky310.

[46] J. Fox, The R Commander: A Basic-Statistics Graphical User Interface to R, Journal of Statistical Software. 14 (2005) 1–42. https://doi.org/10.18637/jss.v014.i09.

[47] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets, Anal. Chem. 84 (2012) 283–289. https://doi.org/10.1021/ac202450g.

[48] Y.C. Tai, T.P. Speed, A multivariate empirical Bayes statistic for replicated microarray time course data, Ann. Statist. 34 (2006) 2387–2412. https://doi.org/10.1214/009053606000000759.

[49] M. Lucero, R. Estell, M. Tellez, E. Fredrickson, A retention index calculator simplifies identification of plant volatile organic compounds, Phytochemical Analysis. 20 (2009) 378–384. https://doi.org/10.1002/pca.1137.

[50] 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. 3 (2007) 211–221. https://doi.org/10.1007/s11306-007-0082-2.

[51] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics. 16 (2002) 119–128. https://doi.org/10.1002/cem.695.

[52] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, NY, 1966: pp. 391–420. https://ci.nii.ac.jp/naid/20001378860/en/ (accessed October 18, 2020).

[53] J. Cohen, Things I have learned (so far)., American Psychologist. 45 (1990) 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304.

[54] A. Roux, Analysis of human urinary metabolome by liquid chromatography coupled to high resolution mass spectrometry, Thèse de Doctorat, Université Pierre et Marie Curie - Paris VI, 2011. https://tel.archives-ouvertes.fr/tel-00641529 (accessed October 18, 2020).

[55] S. Wiklund, Multivariate data analysis for Omics, (2008). https://metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf (accessed October 18, 2020).

Conclusions and Perspectives

Conclusions and Perspectives

In the present thesis, diverse issues, problematics and challenging tasks facing the development of the "Environmental Metabolic Footprinting" proxy were addressed. Several advances were achieved through the different studies described in the manuscript. These advances led to the improvement of the novel meta-metabolomics-based approach.

In the first chapter, the extraction of the environmental meta-metabolome has been developed and assessed. The novel solvents mix/dual-step extractions gave promising results towards the need of a broadband meta-metabolome extraction. When compared to previous reference extractions, they were the most performant for extracting the xenometabolomes of the different applied herbicides, and the polar and semi-polar endometabolomes of the different investigated soils. They also showed a good extraction performance for non-polar endometabolites. On the other hand, they showed the highest ability to discriminate between the polluted and the unpolluted soils, with an acceptable reproducibility. These novel extraction protocols are now ready to be applied for laboratory microcosm and field experiments. They can also be tested for other types of pesticides, and transposed to other types of matrices, *e.g.* sediments and plants.

Moreover**, the problematic of methodological development for untargeted meta-metabolomics has been investigated in this first chapter**. This issue was addressed by implementing a novel untargeted metabolomics/chemometrics-based approach that could assess the different analytical requirements for extractions, such as the coverage of wide molecular diversity, the quantitative efficiency, and the reproducibility. The approach could also assess the reliability of the extractions regarding the targeted environmental question, *i.e.* the "resilience time". Hence, it has been shown that the computational and automated preprocessing workflows, combined with chemometrics and statistical analyses can afford a suitable tool in order to help developing extractions and analysis methods. These tools are basically dedicated for the applied untargeted metabolic profiling-based research, but could also be exploited in the present thesis in order to develop analytical methods for an untargeted meta-metabolomics approach. Hence, Analytical Chemistry and Metabolomics could dialectically improve each other's.

On the other hand, **the set-up of (U)HPLC-ESI-HRMS/MS analytical methods dedicated for analyzing complex environmental samples has been achieved in the first and the**

second chapters. These analytical methods proved their ability to generate a large amount of information and high-quality data. The UHPLC and HPLC systems were capable to achieve repeatable, efficient, and short-time separations by 20 min and 26 min of run, respectively. The use of modified C18 columns has been proved as a good compromise to meet the need of covering a wide range of polarity in a single run. The HRMS has been proven as a highly suitable analysis technique to deal with such complex samples that contain a wide variety of known and unknown metabolites. The detections and characterizations were performed with high levels of resolving power, precision, selectivity, sensitivity and repeatability for *m/z* measuring and ion signal detection.

Moreover, the set-up of LC-HRMS data processing workflows has been performed. These computational tools are essential to deal with the large amounts of complex LC-HRMS data that were produced by analyzing complex multi-factorial sample designs consisting of big numbers of samples. Analyzing these big data without such automated tools will not be possible. Thus, this step is mandatory in order to reach the chemometrics and statistical analyses stage.

Besides, the second chapter showed that the in-depth investigation of analytical raw data and the awareness of the instrument fundamentals are essential for avoiding false data interpretation and biased conclusions. In fact, the automated preprocessing workflow and the statistical analyses are powerful tools for information prioritization. However, the raw data present a basis for results validation and quality control. In the present case, the raw data investigation allowed to reveal **the Ion Suppression phenomenon that could lead to the determination of false positives**. This phenomenon was systematically occurring in the spiked samples due to their higher complexity. Thus, a pragmatic dilutions-based strategy was set-up in order to overcome this effect and to prioritize the significant biomarker candidates after filtrating the false positives. This step was added as a validation pipeline that must be included in the EMF workflow.

For the third chapter, a novel "hybrid" EMF concept has been set-up. **It is dedicated for analyzing pesticides' volatile residues. In this framework, the sample design and the use of the online HS-SPME-GC-MS as analytical technique** allowed for the development of a green, non-destructive and cost-effective approach. It needs less number of samples that can be analyzed for several time points during kinetics studies. The automated method could also perform reproducible high throughput analyses without the need of the laborious and time-consuming multi-step sample preparation. On the other hand, a chemometric and statistical

workflow was established for the kinetics study. Different multivariate statistical models were applied and the workflow has been proven suitable for the investigation of the environmental fate. This workflow can now be used for future experiments including LC-MS-based kinetics studies that will be carried out on solid matrices.

Therefore, this approach can now be used for extended studies that can include different types of pesticides and their pure active substances, as well as different types of soil. It can also be combined with other studies that investigate pesticides fate in the solid matrix, in order to provide complementary information on pesticide's degradation. Its transposition and adaption to *in situ* sampling and investigations can also be an option to be exploited in the context of the long-term perspectives.

It is worth mentioning that other suitable methods for Volatile Organic Compounds (VOC) analysis, such as the Dynamic Headspace (DHS) and the In-Tube Extraction (ITEX) can be assessed in order to increase the sensitivity, or to complement the HS-SPME by widening the range of the studied molecules. These methods can also be developed in order to detect soil endogenous VOCs. In fact, the production of these VOCs is influenced by the microbial activity. Thus, such volatile endometabolites can be biomarkers of the environmental impact.

Nevertheless, the EMF improvement still need several questions to be addressed. For instance, the **sensitivity** of the approach is an important indicator to be assessed. This task is however challenging from an analytical point of view. Otherwise, it seems that some of the recommendations in the testing guidelines, particularly those concerning persistence and dissipation[21], along with the **dose-response-based approximation** that was used in Chapter III can afford some precursor ideas to address this issue and develop novel sensitivity concepts and indicators.

On the other hand, **the influence of certain data processing factors on results should be in-depth investigated in order to optimize and validate the processing workflow**. For instance, factors as the change/update of preprocessing platforms and algorithms, the elimination of ion redundancies, the data scaling and normalization, and several other factors can potentially affect the data interpretation and thus the results, particularly when the difference between polluted and unpolluted samples is relatively low, *i.e.*, when the "resilience time" is about to be reached.

[21] European Commission – Directorate General for Agriculture, Guidance Document on Persistence in Soil (9188/VI/97 rev. 8), (2000). https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_guide_fate_soil-persistance.pdf (accessed October 12, 2020).

Reliable conclusions are highly required at this stage. Thus, this question should be evaluated in the next EMF studies.

Besides, another problematics is **the choice between the direct analyses of samples after kinetic time-point extraction, or the stock of samples until the kinetics study is achieved**. In fact, if all the samples are stocked until the end of the study, their analysis can be performed by successive analytical sequences in a short period. This strategy can minimize the analytical drift and assure robust datasets with a high analytical reproducibility. However, if the kinetic study is relatively long, the stock of samples for an extended period may engender certain degradations in their meta-metabolome. The direct sample analysis is thus a suitable strategy to avoid samples stock. However, the analytical response can change by the time due to several causes such as instrument use and maintenance or solvents/reagents batch changing. This analytical variation can alter data interpretation robustness. Therefore, future works must be performed in order to optimize the correction of such analytical drifts, by optimizing an appropriate QC strategy for long period-separated analyses.

Ultimately, the application of the EMF presents an important examination of its capabilities. Thus, new studies as kinetics experiments in laboratory microcosm and in the field, as well as **the transposition of the EMF on different types of environmental matrices as plants and water** constitute an essential path towards the proof of this novel approach as an appropriate proxy to assess the environmental fate and impact of complex (bio)pesticides.

Finally, in the present thesis, a big step towards the development of the novel "Environmental Metabolic Footprinting" proxy has been achieved. These successful advances can also be useful to deal with diverse problematics that might be general in Analytical Chemistry and Mass Spectrometry-based Metabolomics.

Appendices

Appendix I

For Chapter I

Liquid Chromatography-High Resolution Mass Spectrometry-based untargeted profiling as a tool for analytical development: Assessment of novel extraction protocols for herbicide-polluted soil meta-metabolomics

Appendix I

Hikmat Ghosson[1,2,*], Yannick Brunato[1,2], Delphine Raviglione[1,2],
Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,3,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: *Equal Contribution (Last Co-authors)*

A.I.1. Soils physical-biochemical analyses and characterization: in-details

Soils composition analyses and characterization were performed by Galys/Arterris Laboratory (Toulouges, Pyrénées-Orientales, France) accredited by the French Accreditation Committee – Cofrac (Accreditation N° 1-6798). Results are detailed in Table A.I 1.

Table A.I 1: *Soils physical-biochemical properties.*
CEC: Cation Exchange Capacity.
AWC: Available Water Capacity.

Properties	**Soil of Perpignan (SP)**	**Soil of Torreilles (ST)**
Coarse Sand (g Kg^{-1})	323.00	22.00
Fine Sand (g Kg^{-1})	224.00	183.00
Total Sand (%)	*54.70*	*20.50*
Coarse Silt (g Kg^{-1})	118.00	269.00
Fine Silt (g Kg^{-1})	173.00	318.00
Total Silt (%)	*29.10*	*58.70*
Clay (g Kg^{-1})	162.00	208.00
Total Clay (%)	*16.20*	*20.80*
Total Limestone (g Kg^{-1})	72.00	87.20
Organic Matter (g Kg^{-1})	27.50	20.83
Organic Carbon (g Kg^{-1})	15.99	11.78
Total Nitrogen (g Kg^{-1})	1.25	0.96
C/N	12.79	12.27
CEC (meq Kg^{-1})	99.00	91.60
Jorêt-Hébert P_2O_5 (mg Kg^{-1})	308.00	316.00
CaO (mg Kg^{-1})	7218.00	7708.00
MgO (mg Kg^{-1})	385.00	339.00
K_2O (mg Kg^{-1})	250.00	629.00
K_2O/MgO	0.65	1.86
Na_2O (mg Kg^{-1})	59.00	0.00
Exchangeable Mn (mg Kg^{-1})	2.60	15.22
Exchangeable Cu (mg Kg^{-1})	2.80	75.14
Ca/CEC (%)	260.30	7.20
K/CEC (%)	5.36	14.50
Mg/CEC (%)	19.29	18.40
Na/CEC (%)	1.92	0.00
H/CEC (%)	0.00	0.00
pH in Water	8.04	8.40
pH in KCl	7.57	8.00
Capping Index	0.59	1.45
Porosity Index	1.99	2.00
AWC (mm on 30 cm of soil)	26.70	46.00

Table A.I 2: *Summary of the implemented experimental design and sample codes.*
E1: Simple extraction based on EtOAc + HCl (Romdhane et al. [1]).
E2: Simple extraction based on MeOH (Anastassiades et al. [2]).
E3: Simple extraction based on MeOH + FA (Anastassiades et al. [2]).
E4: Double extraction-based protocol (ACN/iPA; H_2O/MeOH).
E5: Double extraction-based protocol (ACN/iPA + FA; H_2O/MeOH + FA).
SP: Soil of Perpignan.
ST: Soil of Torreilles.
Bel: Soil polluted with formulated Nonanoic acid (Beloukha®) herbicide.
Gly: Soil polluted with formulated Glyphosate Tartan Super 360™ herbicide.
Ctr: Unpolluted control soil.

Extraction protocol	Soil type	Environmental condition	Sample group	Replicate	Sample code	"E"-pools
E1	SP	Bel	E1SPBel	R1	E1SPBelR1	E1QC
				R2	E1SPBelR2	
				R3	E1SPBelR3	
				R4	E1SPBelR4	
				R5	E1SPBelR5	
		Ctr	E1SPCtr	R1	E1SPCtrR1	
				R2	E1SPCtrR2	
				R3	E1SPCtrR3	
				R4	E1SPCtrR4	
				R5	E1SPCtrR5	
		Gly	E1SPGly	R1	E1SPGlyR1	
				R2	E1SPGlyR2	
				R3	E1SPGlyR3	
				R4	E1SPGlyR4	
				R5	E1SPGlyR5	
	ST	Bel	E1STBel	R1	E1STBelR1	
				R2	E1STBelR2	
				R3	E1STBelR3	
				R4	E1STBelR4	
				R5	E1STBelR5	
		Ctr	E1STCtr	R1	E1STCtrR1	
				R2	E1STCtrR2	
				R3	E1STCtrR3	
				R4	E1STCtrR4	
				R5	E1STCtrR5	
		Gly	E1STGly	R1	E1STGlyR1	
				R2	E1STGlyR2	
				R3	E1STGlyR3	
				R4	E1STGlyR4	
				R5	E1STGlyR5	
E2	SP	Bel	E2SPBel	R1	E2SPBelR1	E2QC
				R2	E2SPBelR2	
				R3	E2SPBelR3	
				R4	E2SPBelR4	
				R5	E2SPBelR5	
		Ctr	E2SPCtr	R1	E2SPCtrR1	
				R2	E2SPCtrR2	
				R3	E2SPCtrR3	
				R4	E2SPCtrR4	
				R5	E2SPCtrR5	
		Gly	E2SPGly	R1	E2SPGlyR1	
				R2	E2SPGlyR2	
				R3	E2SPGlyR3	
				R4	E2SPGlyR4	
				R5	E2SPGlyR5	
	ST	Bel	E2STBel	R1	E2STBelR1	
				R2	E2STBelR2	
				R3	E2STBelR3	
				R4	E2STBelR4	
				R5	E2STBelR5	
		Ctr	E2STCtr	R1	E2STCtrR1	
				R2	E2STCtrR2	
				R3	E2STCtrR3	
				R4	E2STCtrR4	
				R5	E2STCtrR5	
		Gly	E2STGly	R1	E2STGlyR1	
				R2	E2STGlyR2	
				R3	E2STGlyR3	
				R4	E2STGlyR4	
				R5	E2STGlyR5	
E3	SP	Bel	E3SPBel	R1	E3SPBelR1	E3QC
				R2	E3SPBelR2	
				R3	E3SPBelR3	
				R4	E3SPBelR4	
				R5	E3SPBelR5	
		Ctr	E3SPCtr	R1	E3SPCtrR1	
				R2	E3SPCtrR2	
				R3	E3SPCtrR3	
				R4	E3SPCtrR4	
				R5	E3SPCtrR5	
		Gly	E3SPGly	R1	E3SPGlyR1	
				R2	E3SPGlyR2	
				R3	E3SPGlyR3	
				R4	E3SPGlyR4	
				R5	E3SPGlyR5	

	ST	Bel	E3STBel	R1	E3STBelR1	
				R2	E3STBelR2	
				R3	E3STBelR3	
				R4	E3STBelR4	
				R5	E3STBelR5	
		Ctr	E3STCtr	R1	E3STCtrR1	
				R2	E3STCtrR2	
				R3	E3STCtrR3	
				R4	E3STCtrR4	
				R5	E3STCtrR5	
		Gly	E3STGly	R1	E3STGlyR1	
				R2	E3STGlyR2	
				R3	E3STGlyR3	
				R4	E3STGlyR4	
				R5	E3STGlyR5	
E4	SP	Bel	E4SPBel	R1	E4SPBelR1	E4QC
				R2	E4SPBelR2	
				R3	E4SPBelR3	
				R4	E4SPBelR4	
				R5	E4SPBelR5	
		Ctr	E4SPCtr	R1	E4SPCtrR1	
				R2	E4SPCtrR2	
				R3	E4SPCtrR3	
				R4	E4SPCtrR4	
				R5	E4SPCtrR5	
		Gly	E4SPGly	R1	E4SPGlyR1	
				R2	E4SPGlyR2	
				R3	E4SPGlyR3	
				R4	E4SPGlyR4	
				R5	E4SPGlyR5	
	ST	Bel	E4STBel	R1	E4STBelR1	
				R2	E4STBelR2	
				R3	E4STBelR3	
				R4	E4STBelR4	
				R5	E4STBelR5	
		Ctr	E4STCtr	R1	E4STCtrR1	
				R2	E4STCtrR2	
				R3	E4STCtrR3	
				R4	E4STCtrR4	
				R5	E4STCtrR5	
		Gly	E4STGly	R1	E4STGlyR1	
				R2	E4STGlyR2	
				R3	E4STGlyR3	
				R4	E4STGlyR4	
				R5	E4STGlyR5	
E5	SP	Bel	E5SPBel	R1	E5SPBelR1	E5QC
				R2	E5SPBelR2	
				R3	E5SPBelR3	
				R4	E5SPBelR4	
				R5	E5SPBelR5	
		Ctr	E5SPCtr	R1	E5SPCtrR1	
				R2	E5SPCtrR2	
				R3	E5SPCtrR3	
				R4	E5SPCtrR4	
				R5	E5SPCtrR5	
		Gly	E5SPGly	R1	E5SPGlyR1	
				R2	E5SPGlyR2	
				R3	E5SPGlyR3	
				R4	E5SPGlyR4	
				R5	E5SPGlyR5	
	ST	Bel	E5STBel	R1	E5STBelR1	
				R2	E5STBelR2	
				R3	E5STBelR3	
				R4	E5STBelR4	
				R5	E5STBelR5	
		Ctr	E5STCtr	R1	E5STCtrR1	
				R2	E5STCtrR2	
				R3	E5STCtrR3	
				R4	E5STCtrR4	
				R5	E5STCtrR5	
		Gly	E5STGly	R1	E5STGlyR1	
				R2	E5STGlyR2	
				R3	E5STGlyR3	
				R4	E5STGlyR4	
				R5	E5STGlyR5	

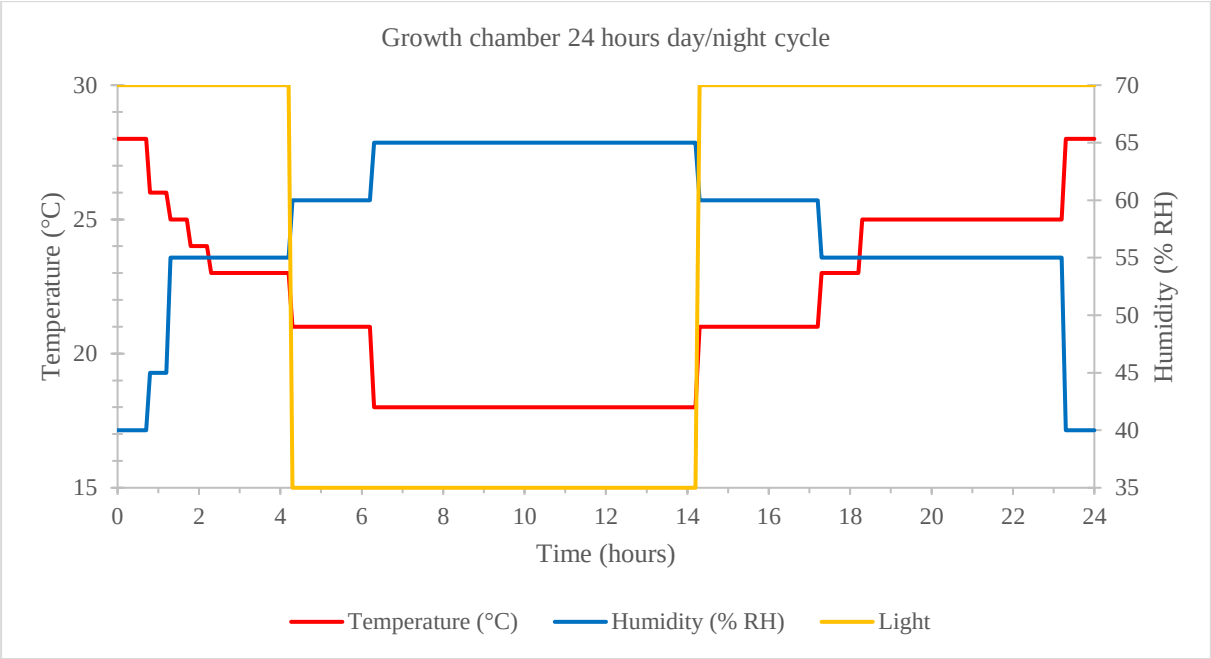


Figure A.I 1: *Growth chamber 24 hours day/night cycle described in Section 2.3. – Chapter I.*
When the orange line is upside: the light is ON (day). When its downside: the light is OFF (night).

A.I.2. High Resolution Mass Spectrometry: parameters and conditions for Q/ToF

A.I.2.1. HRMS: Funnels, Quadrupole and Collision Cell tuning

Funnels, Quadrupole and Collision Cell parameters were optimized in order to favor the transfer of ions with *m/z* between 90 and 1000 to the ToF analyzer. The Funnel 1 RF and Multipole RF were set to 400 Vpp. The applied Quadrupole ion energy was fixed to 4 eV. The Quadrupole Low Mass was equal to *m/z* 90. The collision energy was set to 8 eV (no CID fragmentation), with 5 µs of pre-pulse storage. A basic stepping mode was applied for ion transfer between the Collision Cell and the ToF analyzer: the Collision RF was set from 250 to 1000 Vpp, and the transfer time was set from 45 to 90 µs, with 80 % to 20 % of timing. Parameters were experimentally optimized by monitoring the profile of NaF calibration clusters.

A.I.2.2. ToF voltage settings and resolving power

ToF voltage sets are the following: Pulser Push/Pull: +/– 1768.5 V, Corrector Lens: 7904.0 V, Corrector Fill: 62.4 V, Corrector Extract: 593.9 V, Flight Tube: 12000.0 V, Decelerator: 586.4 V, Reflector: 3465.1, Detector Source: 0.0 V, Detector ToF: 1675.4 V, Detector ToF Delta: 0.0 V. The resolving power was experimentally assessed at the Full Width at Half Maximum (FWHM) of *m/z* peaks of all NaF clusters. Values are summarized in Table A.I 3.

***Table A.I 3**: Q/ToF resolving power at the FWHM of m/z peaks.*
‡: Means of Resolution and S/N calculated from all QC injections (along analytical sequences).

Cluster ion	***m/z***	**Resolution‡**	**S/N‡**
$Na(HCOONa)1^+$	90.9766	22683.57	1916.68
$Na(HCOONa)2^+$	158.9641	25756.33	2086.76
$Na(HCOONa)3^+$	226.9515	34347.90	10443.73
$Na(HCOONa)4^+$	294.9389	25372.44	798.82
$Na(HCOONa)5^+$	362.9263	32274.18	6185.83
$Na(HCOONa)6^+$	430.9138	33452.43	6669.00
$Na(HCOONa)7^+$	498.9012	29481.97	3175.87
$Na(HCOONa)8^+$	566.8886	30183.26	3655.60
$Na(HCOONa)9^+$	634.8760	28902.60	2367.15
$Na(HCOONa)10^+$	702.8635	28330.47	2289.43
$Na(HCOONa)11^+$	770.8509	26111.96	1133.65
$Na(HCOONa)12^+$	838.8383	27197.31	1158.60
$Na(HCOONa)13^+$	906.8257	26164.43	464.65
$Na(HCOONa)14^+$	974.8132	26243.74	381.80

***Table A.I 4**: Collision energy ramp for CID.*

m/z	**Width (*m/z*)**	**Collision Energy (eV)**	**Charge State**
100	± 0.50	30	1
250	± 0.50	40	1
500	± 0.50	50	1
750	± 0.50	60	1
1000	± 0.50	70	1

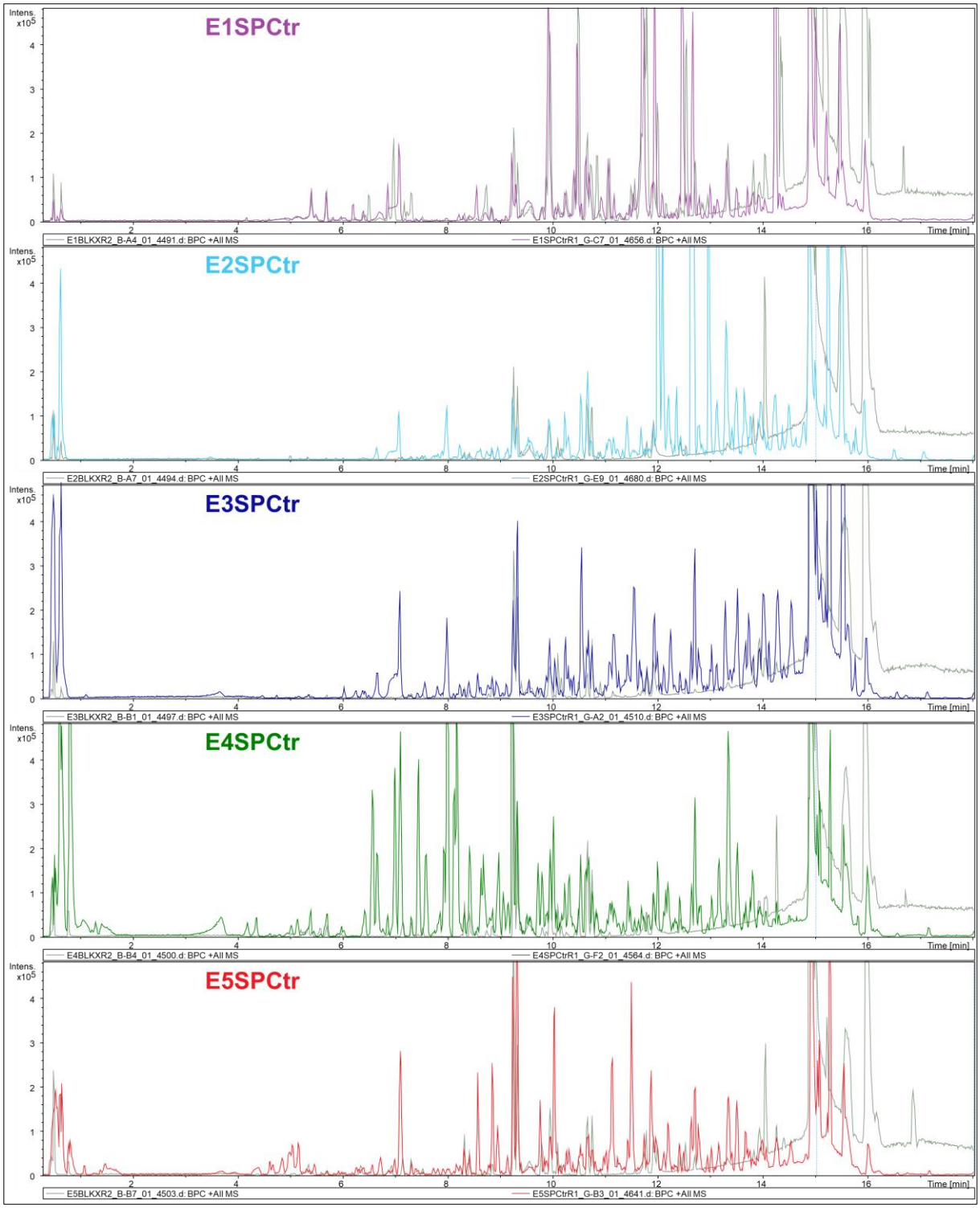


***Figure A.I 2**: Chromatograms for Unpolluted Control Soil – Perpignan (SPCtr).*
The grey chromatogram in background is for the blank extraction. Intensities of all BPCs (Figure A.I 2 to Figure A.I 7) are standardized to a fixed scale of 4.80E5. BPCs generated using DataAnalysis 4.3.

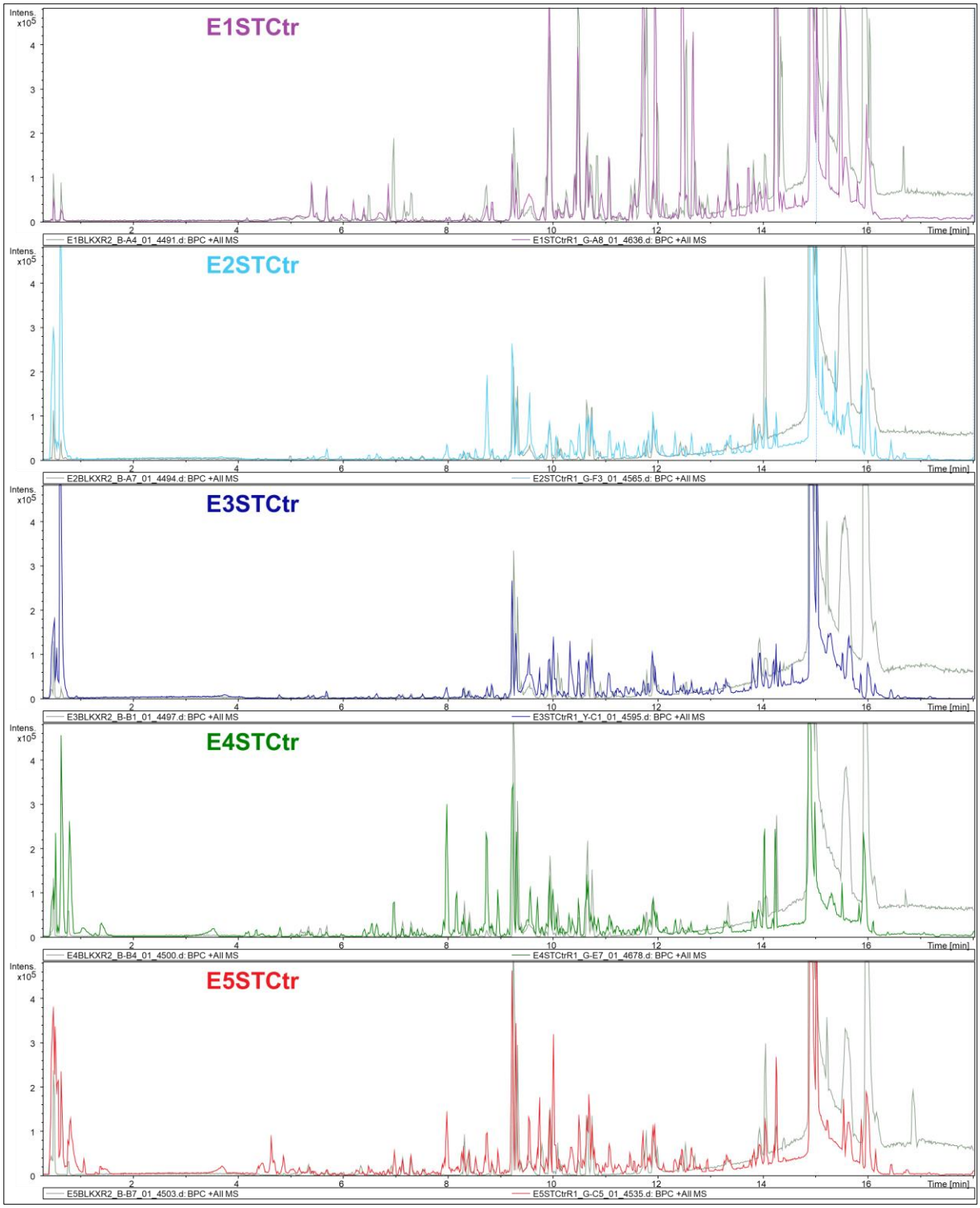


***Figure A.I 3**: Chromatograms for Unpolluted Control Soil – Torreilles (STCtr).*
The grey chromatogram in background is for the blank extraction. Intensities of all BPCs (Figure A.I 2 to Figure A.I 7) are standardized to a fixed scale of 4.80E5. BPCs generated using DataAnalysis 4.3.

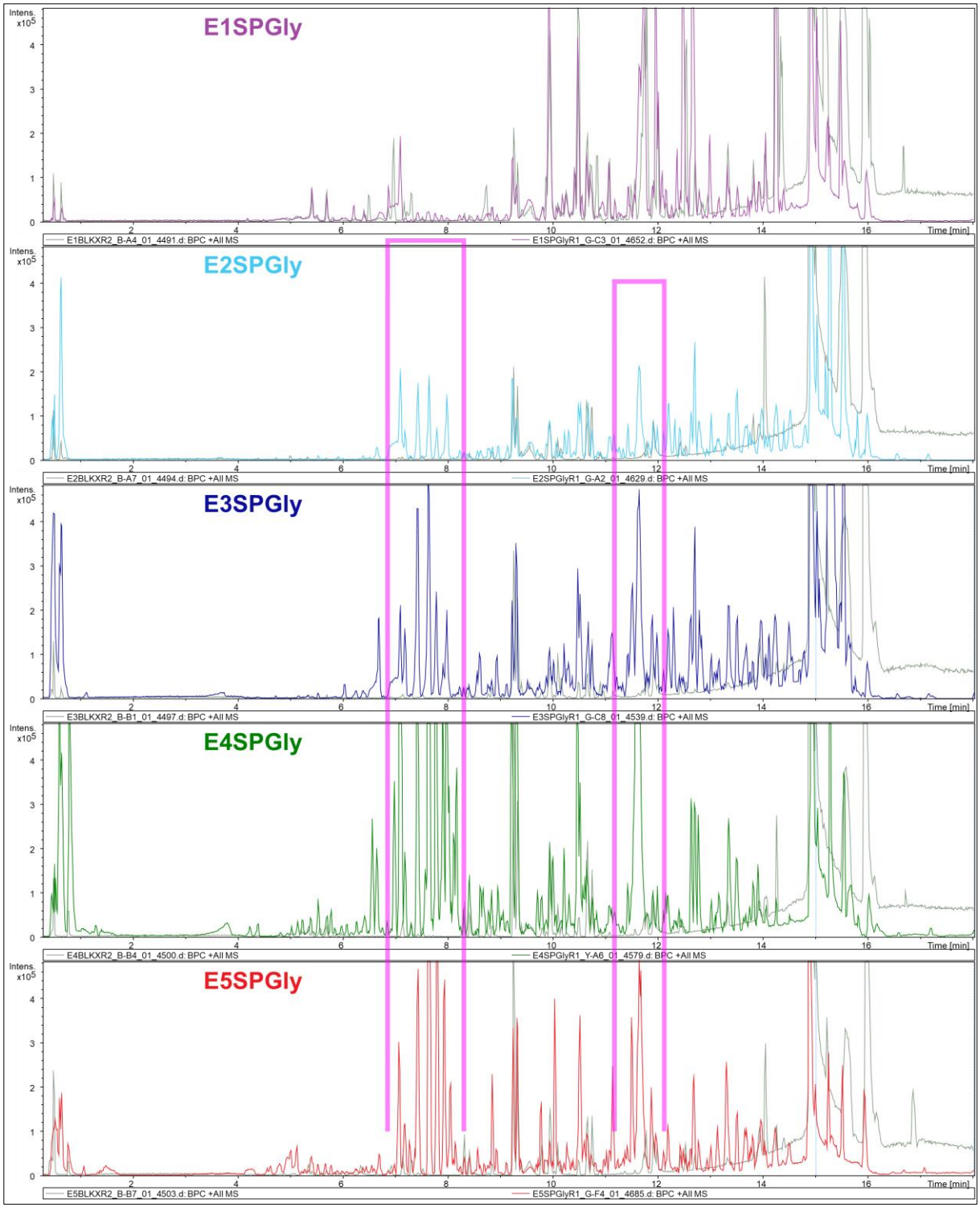


***Figure A.I 4**: Chromatograms for Formulated Glyphosate-Polluted Soil – Perpignan (SPGly).*
The grey chromatogram in background is for the blank extraction. Intensities of all BPCs (Figure A.I 2 to Figure A.I 7) are standardized to a fixed scale of 4.80E5. BPCs generated using DataAnalysis 4.3.

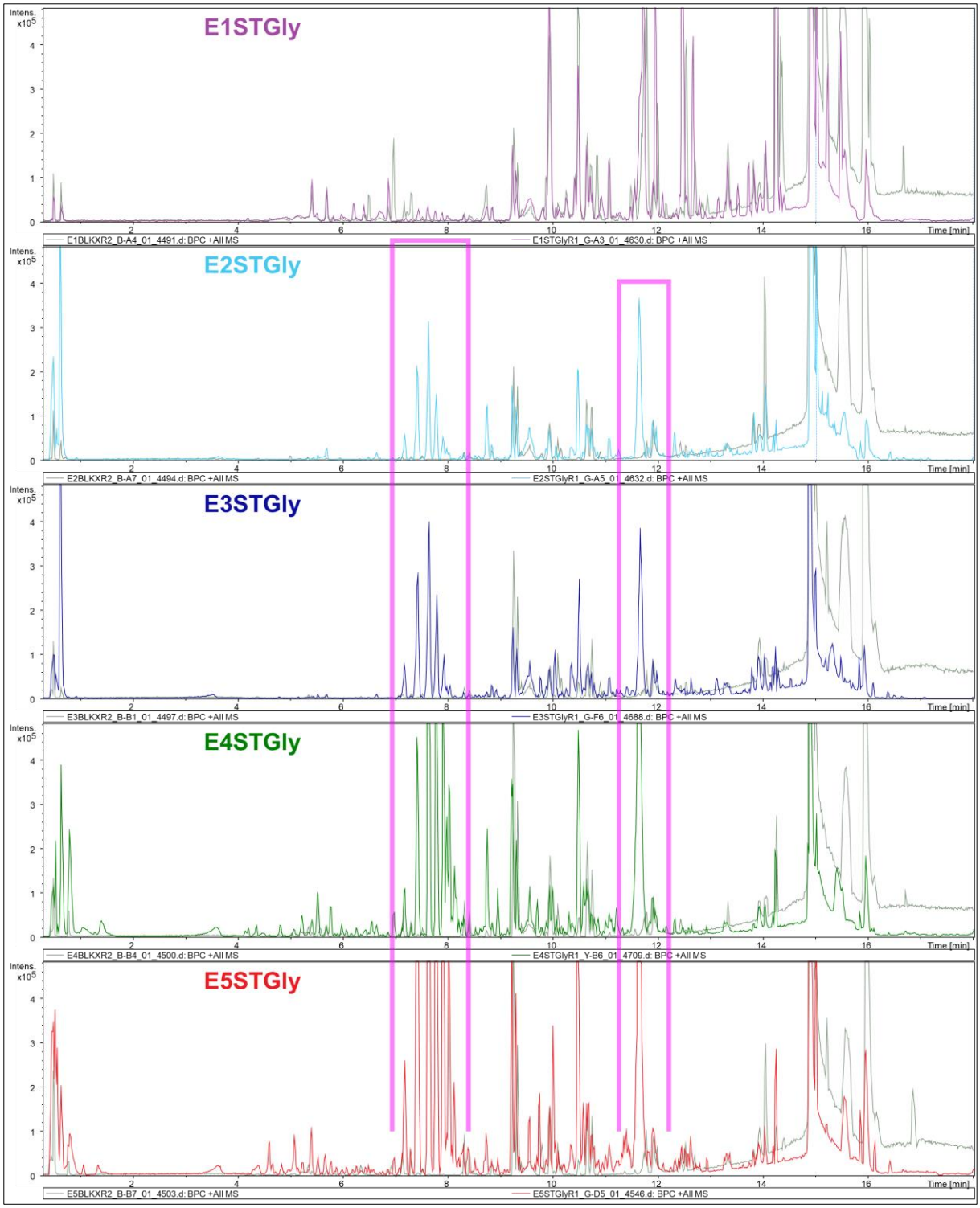


***Figure A.I 5**: Chromatograms for Formulated Glyphosate-Polluted Soil – Torreilles (STGly).*
The grey chromatogram in background is for the blank extraction. Intensities of all BPCs (Figure A.I 2 to Figure A.I 7) are standardized to a fixed scale of 4.80E5. BPCs generated using DataAnalysis 4.3.

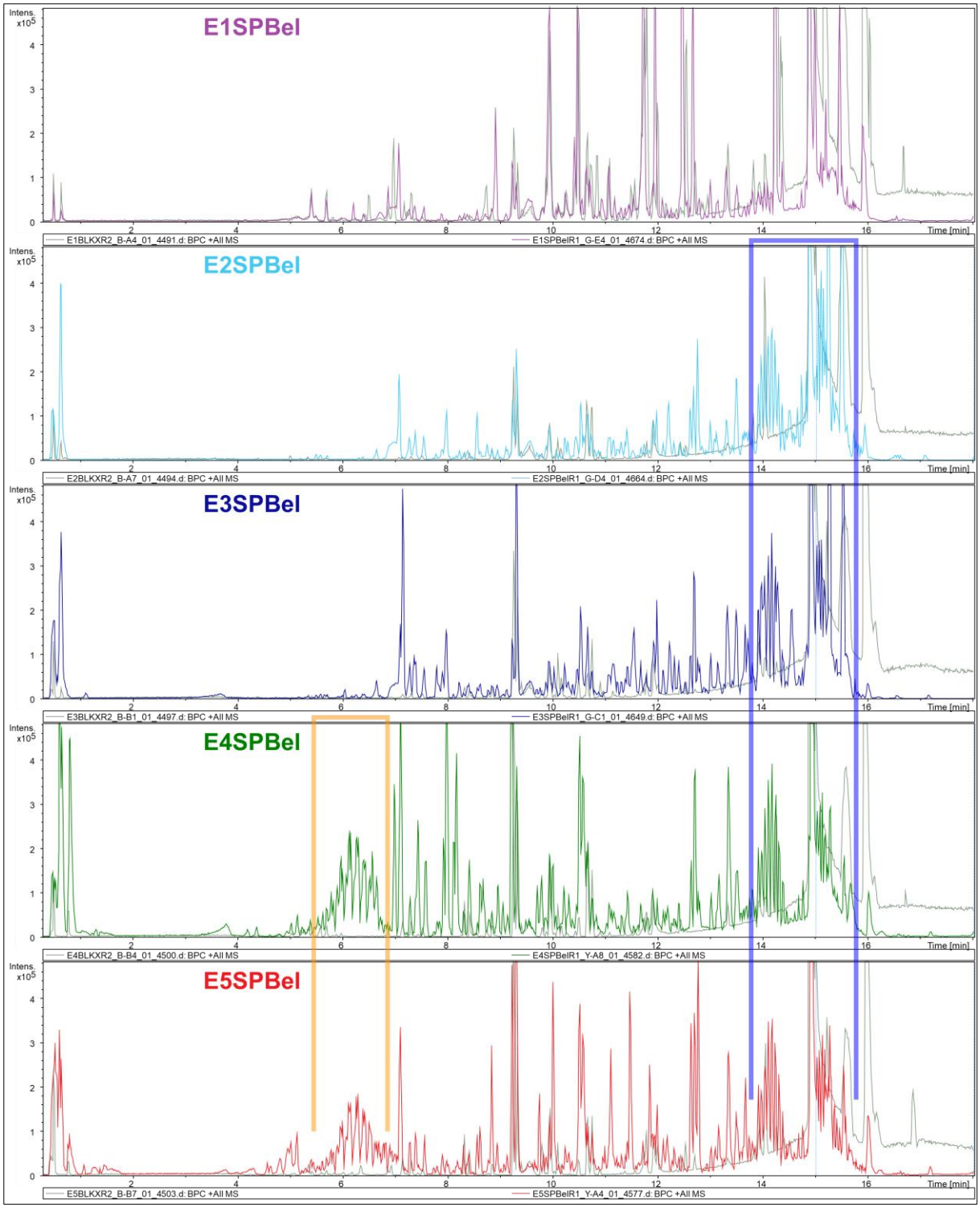


Figure A.I 6: *Chromatograms for Formulated Nonanoic acid-Polluted Soil – Perpignan (SPBel).*

The grey chromatogram in background is for the blank extraction. Intensities of all BPCs (Figure A.I 2 to Figure A.I 7) are standardized to a fixed scale of 4.80E5. BPCs generated using DataAnalysis 4.3.

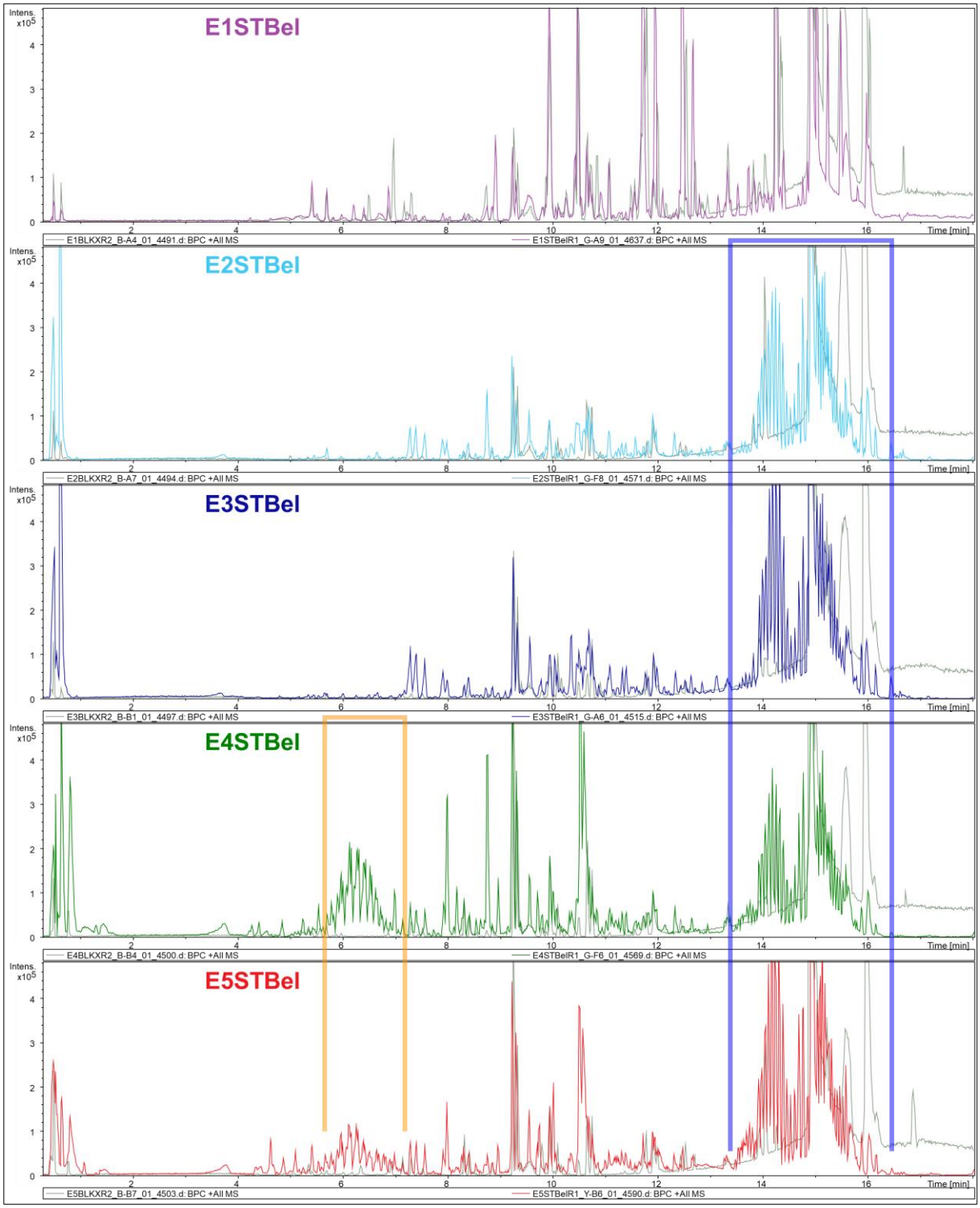


***Figure A.I 7**: Chromatograms for Formulated Nonanoic acid-Polluted Soil – Torreilles (STBel).*

The grey chromatogram in background is for the blank extraction. Intensities of all BPCs (Figure A.I 2 to Figure A.I 7) are standardized to a fixed scale of 4.80E5. BPCs generated using DataAnalysis 4.3.

***Table A.I 5**: Xenometabolome MF calculation after subtraction of endometabolome ($MF_{(Ctr)}$).*

MF calculation	Soil	E1	E2	E3	E4	E5
$MF_{(Gly)}$-$MF_{(Ctr)}$	***SP***	136	58	187	450	541
	ST	159	149	278	452	429
$MF_{(Bel)}$-$MF_{(Ctr)}$	***SP***	586	576	691	1537	1687
	ST	695	1115	982	1715	1992

***Table A.I 6**: OPLS-DA Cross-Validation results. Performed on the light dataset after glog transformation and Pareto scaling.*
All data are in %.

	SP						**ST**					
	p1			***o1***			***p1***			***o1***		
	R2X	*R2Y*	*Q2*	*R2X*	*R2Y*	*Q2*	*R2X*	*R2Y*	*Q2*	*R2X*	*R2Y*	*Q2*
Gly												
E1	15.80	93.90	46.20	12.20	05.90	03.44	15.60	93.30	48.20	14.70	06.36	04.89
E2	17.70	97.30	63.10	18.50	02.40	07.36	17.70	97.80	59.60	14.70	01.84	05.27
E3	17.50	98.80	61.10	08.66	01.13	08.62	17.10	97.60	56.40	12.70	02.19	06.57
E4	**24.60**	97.70	**80.80**	14.30	01.95	03.01	**19.60**	87.60	**61.30**	23.50	11.80	11.80
E5	**18.10**	94.50	**60.50**	11.80	05.11	10.90	**18.00**	98.40	**62.30**	11.30	01.48	06.76
Bel												
E1	24.60	99.00	82.90	15.10	00.82	03.84	22.30	99.10	78.00	09.90	00.91	03.68
E2	29.80	99.30	88.70	13.10	00.65	01.90	34.90	99.40	91.50	08.57	00.62	00.79
E3	28.80	97.90	85.70	14.00	01.96	02.89	28.00	98.70	86.50	12.50	01.15	02.13
E4	**37.20**	99.60	**94.50**	13.00	00.31	01.31	**36.00**	97.50	**93.30**	19.90	02.46	02.60
E5	**36.50**	99.10	**91.90**	10.80	00.81	00.81	**34.30**	99.50	**92.40**	09.82	00.41	01.27

A.I.3. Principal Component Analysis

For a general overview of the different datasets generated by the automated pre-processing, PCA was performed on both original and light data matrices. PCA was performed after glog and Pareto scaling were applied to features intensities. Results are shown in Figure A.I 8.

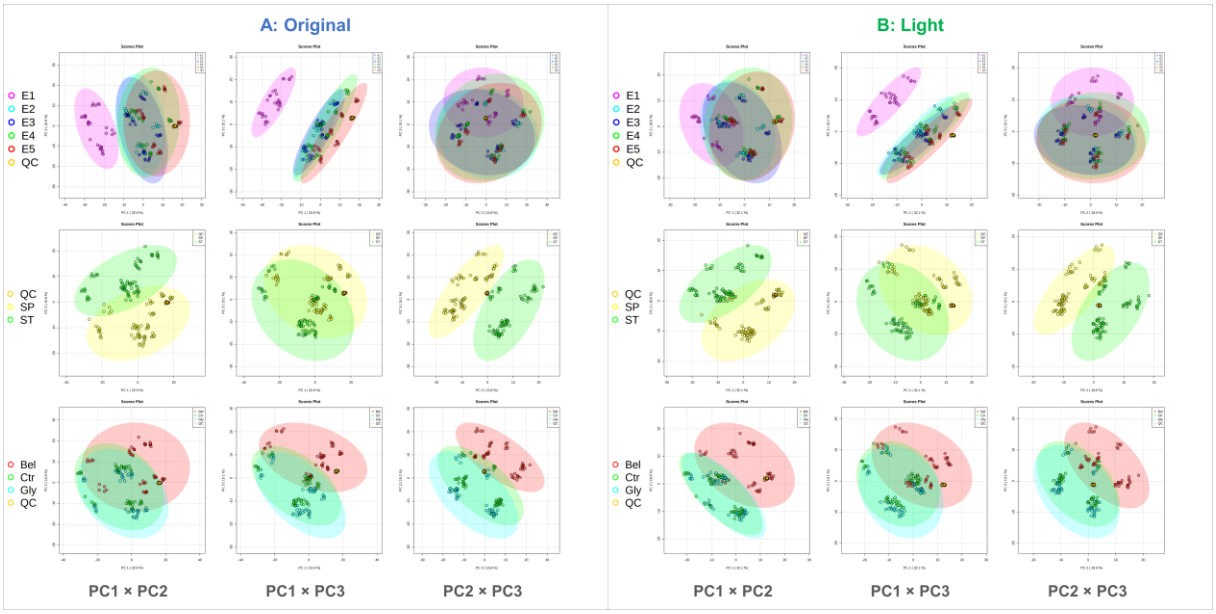


Figure A.I 8: *Principal Component Analysis.*
A: *PCA performed on the original data matrix (411 features).* ***B:*** *PCA performed on the light data matrix (234 features). Contributions (in %) of Principal Components are detailed in Table A.I 7.*
Plots generated using MetaboAnalyst.

Three aspects of discriminations between the metabolic profiles can be observed according to PCA results. The first is the distinction of all samples originating from the E1 extraction protocol (purple cluster in the first row). This difference is mainly observed according to the PC1. It can be explained by the poor metabolic profile generated by E1 protocol. The second observed discrimination is between the two types of soils mainly revealed by the PC2 (green and yellow clusters in the second row). The poor metabolic profile of the ST can explain the significance of this difference. The third discrimination to be noticed (notably according to PC3) is for all soil samples polluted with formulated Nonanoic acid herbicide (red cluster in the third row). The significance of this discrimination can be explained by the heavy xenometabolome originating from the herbicide.

For both original and light datasets, PCAs were capable to reveal the same aspects of metabolic profile discriminations with a slight difference in term of percentage of PC contributions. This is shown by Figure A.I 8 and by Table A.I 7. Such results demonstrate that in general, the elimination of ion redundancies does not significantly affect the metabolic information held by datasets and explained by statistical data analyses.

Table A.I 7: *Contributions (in %) of the Principal Components in the two different data matrices.*

Principal Component	Original	Light
PC1 (%)	32.90	32.10
PC2 (%)	15.80	18.60
PC3 (%)	15.10	13.10

A.I.4. R command lines for Euclidean Distance and Euclidean Distance SDs calculation

The detailed step-by-step description of the calculation process is expanded below.

First, a sub-matrix proper for each extraction-soil-environmental group (and the QC) has been created from the global matrix ("DataMatrix.csv"), through the following script:

```
MX<-read.table("DataMatrix.csv", h=T, sep=";")
fix(MX)

E1SPBel<-MX[1:5,]
row.names(E1SPBel)<-c()
E1SPCtr<-MX[6:10,]
row.names(E1SPCtr)<-c()
E1SPGly<-MX[11:15,]
row.names(E1SPGly)<-c()
E1STBel<-MX[16:20,]
row.names(E1STBel)<-c()
E1STCtr<-MX[21:25,]
row.names(E1STCtr)<-c()
E1STGly<-MX[26:30,]
row.names(E1STGly)<-c()
E2SPBel<-MX[31:35,]
row.names(E2SPBel)<-c()
E2SPCtr<-MX[36:40,]
row.names(E2SPCtr)<-c()
E2SPGly<-MX[41:45,]
row.names(E2SPGly)<-c()
E2STBel<-MX[46:50,]
row.names(E2STBel)<-c()
E2STCtr<-MX[51:55,]
row.names(E2STCtr)<-c()
E2STGly<-MX[56:60,]
row.names(E2STGly)<-c()
E3SPBel<-MX[61:65,]
row.names(E3SPBel)<-c()
E3SPCtr<-MX[66:70,]
row.names(E3SPCtr)<-c()
E3SPGly<-MX[71:75,]
row.names(E3SPGly)<-c()
E3STBel<-MX[76:80,]
row.names(E3STBel)<-c()
E3STCtr<-MX[81:85,]
row.names(E3STCtr)<-c()
E3STGly<-MX[86:90,]
row.names(E3STGly)<-c()
E4SPBel<-MX[91:95,]
row.names(E4SPBel)<-c()
E4SPCtr<-MX[96:100,]
row.names(E4SPCtr)<-c()
```

```
E4SPGly<-MX[101:105,]
row.names(E4SPGly)<-c()
E4STBel<-MX[106:110,]
row.names(E4STBel)<-c()
E4STCtr<-MX[111:115,]
row.names(E4STCtr)<-c()
E4STGly<-MX[116:120,]
row.names(E4STGly)<-c()
E5SPBel<-MX[121:125,]
row.names(E5SPBel)<-c()
E5SPCtr<-MX[126:130,]
row.names(E5SPCtr)<-c()
E5SPGly<-MX[131:135,]
row.names(E5SPGly)<-c()
E5STBel<-MX[136:140,]
row.names(E5STBel)<-c()
E5STCtr<-MX[141:145,]
row.names(E5STCtr)<-c()
E5STGly<-MX[146:150,]
row.names(E5STGly)<-c()
QC<-MX[151:180,]
row.names(QC)<-c()
```

Then, Euclidean Distance calculation was performed. A Euclidean Distance matrix proper for each group was generated with its summary (*i.e.* its Minimum, its 1st Quartile, its Median, its Mean, its 3rd Quartile, and its Maximum), through the script below:

```
DistEucl_E1SPBel<-dist(E1SPBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E1SPBel)

DistEucl_E1SPCtr<-dist(E1SPCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E1SPCtr)

DistEucl_E1SPGly<-dist(E1SPGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E1SPGly)

DistEucl_E1STBel<-dist(E1STBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E1STBel)

DistEucl_E1STCtr<-dist(E1STCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E1STCtr)

DistEucl_E1STGly<-dist(E1STGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E1STGly)

DistEucl_E2SPBel<-dist(E2SPBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E2SPBel)

DistEucl_E2SPCtr<-dist(E2SPCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E2SPCtr)

DistEucl_E2SPGly<-dist(E2SPGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E2SPGly)

DistEucl_E2STBel<-dist(E2STBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E2STBel)

DistEucl_E2STCtr<-dist(E2STCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E2STCtr)
```

```
DistEucl_E2STGly<-dist(E2STGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E2STGly)

DistEucl_E3SPBel<-dist(E3SPBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E3SPBel)

DistEucl_E3SPCtr<-dist(E3SPCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E3SPCtr)

DistEucl_E3SPGly<-dist(E3SPGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E3SPGly)

DistEucl_E3STBel<-dist(E3STBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E3STBel)

DistEucl_E3STCtr<-dist(E3STCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E3STCtr)

DistEucl_E3STGly<-dist(E3STGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E3STGly)

DistEucl_E4SPBel<-dist(E4SPBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E4SPBel)

DistEucl_E4SPCtr<-dist(E4SPCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E4SPCtr)

DistEucl_E4SPGly<-dist(E4SPGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E4SPGly)

DistEucl_E4STBel<-dist(E4STBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E4STBel)

DistEucl_E4STCtr<-dist(E4STCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E4STCtr)

DistEucl_E4STGly<-dist(E4STGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E4STGly)

DistEucl_E5SPBel<-dist(E5SPBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E5SPBel)

DistEucl_E5SPCtr<-dist(E5SPCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E5SPCtr)

DistEucl_E5SPGly<-dist(E5SPGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E5SPGly)

DistEucl_E5STBel<-dist(E5STBel, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E5STBel)

DistEucl_E5STCtr<-dist(E5STCtr, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E5STCtr)

DistEucl_E5STGly<-dist(E5STGly, method = "euclidean", diag = T, upper = T)
summary(DistEucl_E5STGly)

DistEucl_QC<-dist(QC, method = "euclidean", diag = T, upper = T)
summary(DistEucl_QC)
```

After, Euclidean Distance SDs belonging to each group were calculated from their proper Euclidean Distance matrices (generated in the previous step), by applying the following script:

```
sd(DistEucl_E1SPBel, na.rm = FALSE)
sd(DistEucl_E1SPCtr, na.rm = FALSE)
sd(DistEucl_E1SPGly, na.rm = FALSE)
sd(DistEucl_E1STBel, na.rm = FALSE)
sd(DistEucl_E1STCtr, na.rm = FALSE)
sd(DistEucl_E1STGly, na.rm = FALSE)
sd(DistEucl_E2SPBel, na.rm = FALSE)
sd(DistEucl_E2SPCtr, na.rm = FALSE)
sd(DistEucl_E2SPGly, na.rm = FALSE)
sd(DistEucl_E2STBel, na.rm = FALSE)
sd(DistEucl_E2STCtr, na.rm = FALSE)
sd(DistEucl_E2STGly, na.rm = FALSE)
sd(DistEucl_E3SPBel, na.rm = FALSE)
sd(DistEucl_E3SPCtr, na.rm = FALSE)
sd(DistEucl_E3SPGly, na.rm = FALSE)
sd(DistEucl_E3STBel, na.rm = FALSE)
sd(DistEucl_E3STCtr, na.rm = FALSE)
sd(DistEucl_E3STGly, na.rm = FALSE)
sd(DistEucl_E4SPBel, na.rm = FALSE)
sd(DistEucl_E4SPCtr, na.rm = FALSE)
sd(DistEucl_E4SPGly, na.rm = FALSE)
sd(DistEucl_E4STBel, na.rm = FALSE)
sd(DistEucl_E4STCtr, na.rm = FALSE)
sd(DistEucl_E4STGly, na.rm = FALSE)
sd(DistEucl_E5SPBel, na.rm = FALSE)
sd(DistEucl_E5SPCtr, na.rm = FALSE)
sd(DistEucl_E5SPGly, na.rm = FALSE)
sd(DistEucl_E5STBel, na.rm = FALSE)
sd(DistEucl_E5STCtr, na.rm = FALSE)
sd(DistEucl_E5STGly, na.rm = FALSE)
sd(DistEucl_QC, na.rm = FALSE)
```

Finally, for each group, the Euclidean Distance RSD was calculated through multiplying the Euclidean Distance SD by 100 and then dividing it on the Euclidean Distance Mean (taken from the summary of the Euclidean Distance matrix).

References

[1] S. Romdhane, M. Devers-Lamrani, L. Barthelmebs, C. Calvayrac, C. Bertrand, J.-F. Cooper, F.E. Dayan, F. Martin-Laurent, Ecotoxicological Impact of the Bioherbicide Leptospermone on the Microbial Community of Two Arable Soils, Front. Microbiol. 7 (2016). https://doi.org/10.3389/fmicb.2016.00775.

[2] M. Anastassiades, D.I. Kolberg, A. Benkenstein, E. Eichhorn, S. Zechmann, D. Mack, C. Wildgrube, I. Sigalov, D. Dörk, A. Barth, Quick Method for the Analysis of numerous Highly Polar Pesticides in Foods of Plant Origin via LC-MS/MS involving Simultaneous Extraction with Methanol (QuPPe-Method), EU Reference Laboratories for Residues of Pesticides, Fellbach, Germany, 2016.
https://www.eurl-pesticides.eu/userfiles/file/EurlSRM/meth_QuPPe_Version9_1.pdf
(accessed June 15, 2020).

Appendix II-A

For Chapter II

Electrospray Ionization and heterogeneous matrix effects in Liquid Chromatography-Mass Spectrometry-based meta-metabolomics: A biomarker or a suppressed ion?

Appendix II-A

Hikmat Ghosson[1,2,*], Yann Guitton[3], Amani Ben Jrad[1,2], Chandrashekhar Patil[1], Delphine Raviglione[1,2], Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,4,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), Oniris, INRAE, Nantes, F-44307, France

4: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: Equal Contribution (Last Co-authors)

A.II-A.1. Mass Spectrometry conditions – Supplementary Information

A.II-A.1.1. Funnels, Quadrupole and Collision Cell tuning

Funnels, Quadrupole and Collision Cell parameters were optimized in order to favor the transfer of ions with *m/z* between 80 and 1600 to the ToF analyzer. The Funnel 1 RF and Multipole RF were set to 400 Vpp. The in-source CID (isCID) energy was equal to 0.0 eV. The applied Quadrupole ion energy was fixed to 4.0 eV. The Quadrupole Low Mass was equal to *m/z* 50. For the collision cell, the collision energy was set to 8.0 eV (no CID fragmentation), with 7 µs of pre-pulse storage. A basic stepping mode was applied for ion transfer between the Collision Cell and the ToF analyzer: the Collision RF was set from 300 to 1200 Vpp, and the transfer time was set from 80 to 140 µs, with 50 % to 50 % of timing. Parameters were experimentally optimized by monitoring NaF calibration clusters profile.

A.II-A.1.2. ToF voltage settings and resolution

ToF voltage sets are the following: Pulser Push/Pull: +/– 1768.5 V, Corrector Lens: 7904.0 V, Corrector Fill: 62.4 V, Corrector Extract: 593.9 V, Flight Tube: 12000.0 V, Decelerator: 586.4 V, Reflector: 3465.1, Detector Source: 0.0 V, Detector ToF: 1675.4 V, Detector ToF Delta: 0.0 V. ToF resolution was experimentally assessed at the FWHM of *m/z* peaks of all NaF clusters. Values are summarized in Table A.II-A 1 for the Full HRMS acquisitions and in Table A.II-A 2 for the MS/MS acquisitions.

The maximum resolution for the maXis Q/ToF is 80000 at the FWHM of *m/z* 1522. This maximum resolution cannot be reached at *m/z* 200, as the resolution of the ToF decreases following the decrease of *m/*z ratio [1]. In addition, unlike for the FT-MS, the FWHM is constant and does not depend on scan rate in Q/ToF. The resolution is calculated following the "M/ΔM" formula, where the ΔM is measured at the FWHM. Therefore, as the peaks between *m/z* 1246.7629 and *m/z* 1518.7125 are detected at S/N ratios between 205.8 and 45.1 (relatively low S/N if compared to those obtained for smaller *m/*z), the FWHM ΔM is thus larger and the resolution at ~1522 seems to be degraded. This explanation can be supported when comparing the resolution values presented in Table A.II-A 1 and Table A.II-A 2. In fact, the applied conditions for the MS/MS acquisitions (Section 2.7.1. – Chapter II) such as the scan range (*m/z* 40-650) and the ion transfer tunes led to the increase of the S/N of the clusters detected between *m/z* 90.9766 and *m/z* 634.8760. An increase of resolution associated with the increase of S/N can be noticed when comparing the values belonging to the same cluster ions detected in both the Full HRMS and the MS/MS acquisitions.

Table A.II-A 1: *Resolution values at the FWHM of m/z peaks for the Full HRMS acquisitions.*

Cluster ion	***m/z***	**Resolution**	**S/N**
$Na(HCOONa)1^+$	90.9766	22046	1116.3
$Na(HCOONa)2^+$	158.9641	24907	1385.4
$Na(HCOONa)3^+$	226.9515	33449	6710.3
$Na(HCOONa)4^+$	294.9389	25814	552.7
$Na(HCOONa)5^+$	362.9263	29046	4114.0
$Na(HCOONa)6^+$	430.9138	28946	4167.2
$Na(HCOONa)7^+$	498.9012	27787	1909.9
$Na(HCOONa)8^+$	566.8886	28151	2315.3
$Na(HCOONa)9^+$	634.8760	27026	1610.5
$Na(HCOONa)10^+$	702.8635	28438	2498.1
$Na(HCOONa)11^+$	770.8509	28304	1909.1
$Na(HCOONa)12^+$	838.8383	27576	2233.9
$Na(HCOONa)13^+$	906.8257	26846	1018.1
$Na(HCOONa)14^+$	974.8132	26071	750.1
$Na(HCOONa)15^+$	1042.8006	25926	580.9
$Na(HCOONa)16^+$	1110.7880	24739	392.6
$Na(HCOONa)17^+$	1178.7754	28792	305.7
$Na(HCOONa)18^+$	1246.7629	32394	205.8
$Na(HCOONa)19^+$	1314.7503	21538	147.5
$Na(HCOONa)20^+$	1382.7377	29554	104.7
$Na(HCOONa)21^+$	1450.7251	23803	68.0
$Na(HCOONa)22^+$	1518.7125	31862	45.1

Table A.II-A 2: *Resolution values at the FWHM of m/z peaks for the MS/MS acquisitions.*

Cluster ion	***m/z***	**Resolution**	**S/N**
$Na(HCOONa)1^+$	90.9766	23495	5271.5
$Na(HCOONa)2^+$	158.9641	28782	6810.8
$Na(HCOONa)3^+$	226.9515	37535	24773.9
$Na(HCOONa)4^+$	294.9389	30164	1870.6
$Na(HCOONa)5^+$	362.9263	36968	11119.6
$Na(HCOONa)6^+$	430.9138	37395	10116.8
$Na(HCOONa)7^+$	498.9012	34391	4716.4
$Na(HCOONa)8^+$	566.8886	34926	5147.1
$Na(HCOONa)9^+$	634.8760	34247	3348.8

Table A.II-A 3: *Scheduled precursors list applied for Q/ToF Auto MS/MS acquisitions (ESI+).*
Ion isolation is achieved at a low resolution in the Quadrupole mass analyzer.
For all ions, the charge state (z) was equal to 1.
The slight shifts in RT are due to LC capillary maintenance, which changed the void (dead) volume.

Feature	***m/z* (± 0.50)**	**RT (min)**	**RT window (min)**	**CID Energy (eV)**			
				Low	***Intermediate***	***High***	***Relative***
M274T527	274.27	8.85	0.07	20.0	35.0	50.0	27.4
M578T743	578.48	12.50	0.05	30.0	45.0	60.0	57.8
M622T896	621.52	15.19	0.08	30.0	45.0	60.0	62.2
M624T939	623.54	15.88	0.12	30.0	45.0	60.0	62.4

Table A.II-A 4: *Inclusion lists applied for Q/Orbitrap PRM MS/MS acquisitions (ESI+ and ESI–).*
In ESI– mode, only 3 of the targeted metabolites were detected (i.e. M578T743, M622T896, and M624T939).
Ion isolation is achieved at a low resolution in the Quadrupole mass analyzer.
For all ions, the charge state (z) was equal to 1.
The Q/Orbitrap is coupled to a different LC instrument with a different void (dead) volume, which explains the slight shifts in RT.

Feature	***m/z* (± 0.50)**	**Ion Species**	**Start RT (min)**	**End RT (min)**	**HCD Energy – "CE" (eV)**		
					Low	***Intermediate***	***High***
ESI+							
M274T527	274.27	$[M+H]^+$	9.13	9.29	25	40	55
M578T743	578.48	$[M+H]^+$	12.65	12.87	35	50	65
M622T896	621.52	$[M+H]^+$	15.32	15.53	35	50	65
M624T939	623.53	$[M+H]^+$	16.06	16.27	35	50	65
ESI–							
M578T743	576.46	$[M-H]^-$	12.71	12.89	35	50	65
M622T896	619.51	$[M-H]^-$	15.40	15.61	35	50	65
M624T939	621.52	$[M-H]^-$	16.14	16.40	35	50	65

A.II-A.2. Mass Spectrometry data: essential indicators for the quality of analyses

Mass Spectrometry data, consisting of MS scans for ions spectra, are the foundation of the chromatographic data. They are not only essential for metabolites detection, quantification and identification, but also represent a critical indicator for the quality of the acquired data. Indeed, they can reveal several fundamental/instrumental-related artefacts and problems that can influence the quality of the results, and can be related to different experimental issues. For instance, matrix effect and ion suppression can be related to high sample concentration, high sample injection volume, non-optimal source parameters or source fouling. Ion degradation and fragmentation can be results of high-energy application in source or ion transfer compartments. Signal saturation and shifts in *m/z* measuring precision can also be related to non-adapted sample concentrations. Random shifts in *m/z* measuring precision during analytical sequences can be the result of unstable ambient temperature (particularly for ToF analyzers). Imprecise *m/z* measuring, degraded resolution and insufficient MS resolving power are mainly issued from non-optimized HRMS analyzer tunings or non-optimal calibration or "Lock-Mass" methods. Noise artefacts can be related to non-optimal source parameters, source fouling or the application of unsuitable solvents or additives. It is obvious that the relations between the mentioned issues (as well for several other non-mentioned) and their related causes are highly complex and interlaced, hence the importance of the critical check of these MS data, with an awareness of the fundamental and technical aspects of the MS technique, toward unbiased interpretations of results and reliable conclusions.

A.II-A.3. Assessment of potential MS-related aberrations

Before examining dilution profiles, EIC area integration bias related to shifts in *m/z* measuring through dilutions was assessed. Two main reasons that can lead to this bias were examined: i) the effect of charge quantity-in-space, and ii) the saturation of the detector of the mass spectrometer (the Q/ToF MS).

The charge quantity-in-space effect can lead to shifts in *m/z* measuring and a loss of precision and/or resolving power in the ToF analyzer. This is due to repulsions between ions when the quantity of charge-in-space is relatively high. This phenomenon was ruled out after assessing *m/z* precision variations though all original and diluted samples. For all the examined features, no systematic shifts were observed for *m/z* values following dilutions (some examples are shown in Figure A.II-A 7). Moreover, shift magnitudes were assessed by calculating SDs of *m/z* ratios though all samples. For a given *m/z*, one SD is calculated from all injections including different conditions and dilution levels (*m/z* measured at the apex of EIC peaks). The highest SD was 0.0016 (for the feature M625T939), which is below the *m/z* window applied for performing EIC (*i.e.* ± 0.0050). Therefore, the shift in *m/z* measuring does not affect EIC areas integration.

Detector saturation risking significant shifts in *m/z* measuring at the apex of the EIC peaks was assessed for each feature. Differences between *m/z* measured at the apex and *m/z* measured at the tail of EIC were calculated in the original concentrated samples (F01). They were below 0.0050 for all features. The highest shift observed was 0.0030 for M461T812, with an intensity fold of 11.92 between the apex and the tail. For M624T939 showing the highest intensity fold between the apex and the tail (457.33), *m/z* shift was 0.0009. Therefore, no signal saturation was observed and thus all features were detected within the dynamic range of the detector.

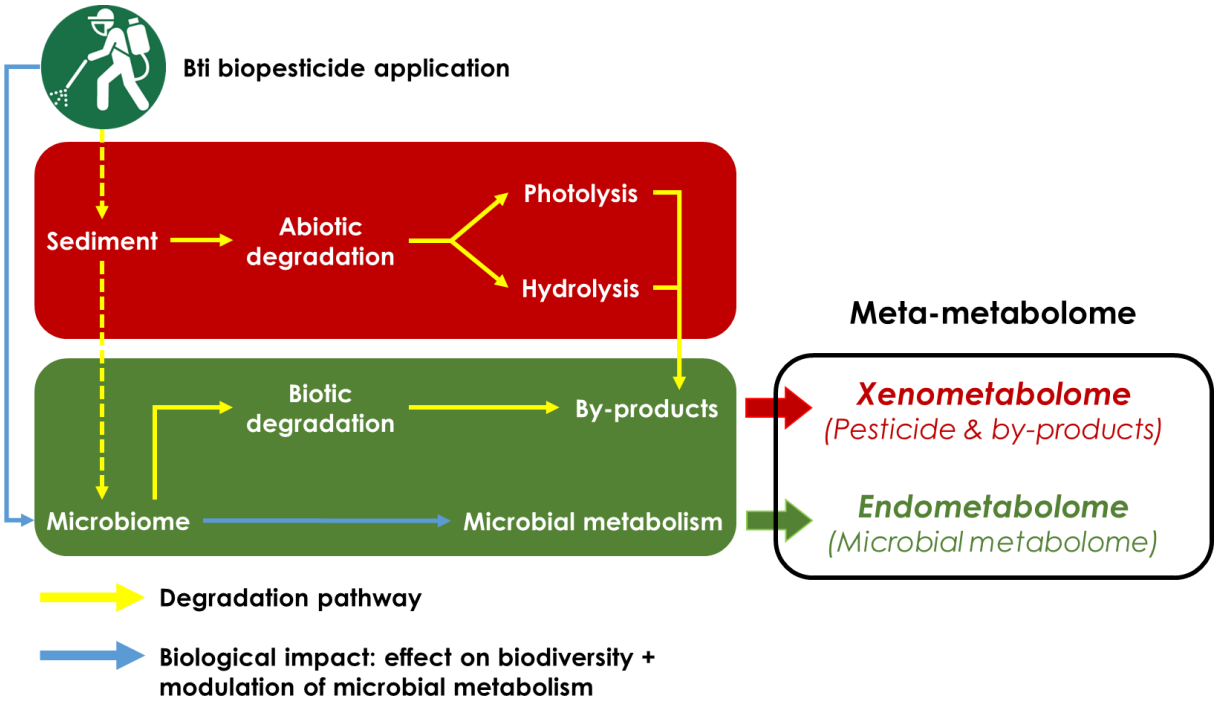


***Figure A.II-A 1**: The Environmental Metabolic Footprinting (EMF) concept.*
Diagram adapted with modifications from Patil et al. [2], with permission from authors.

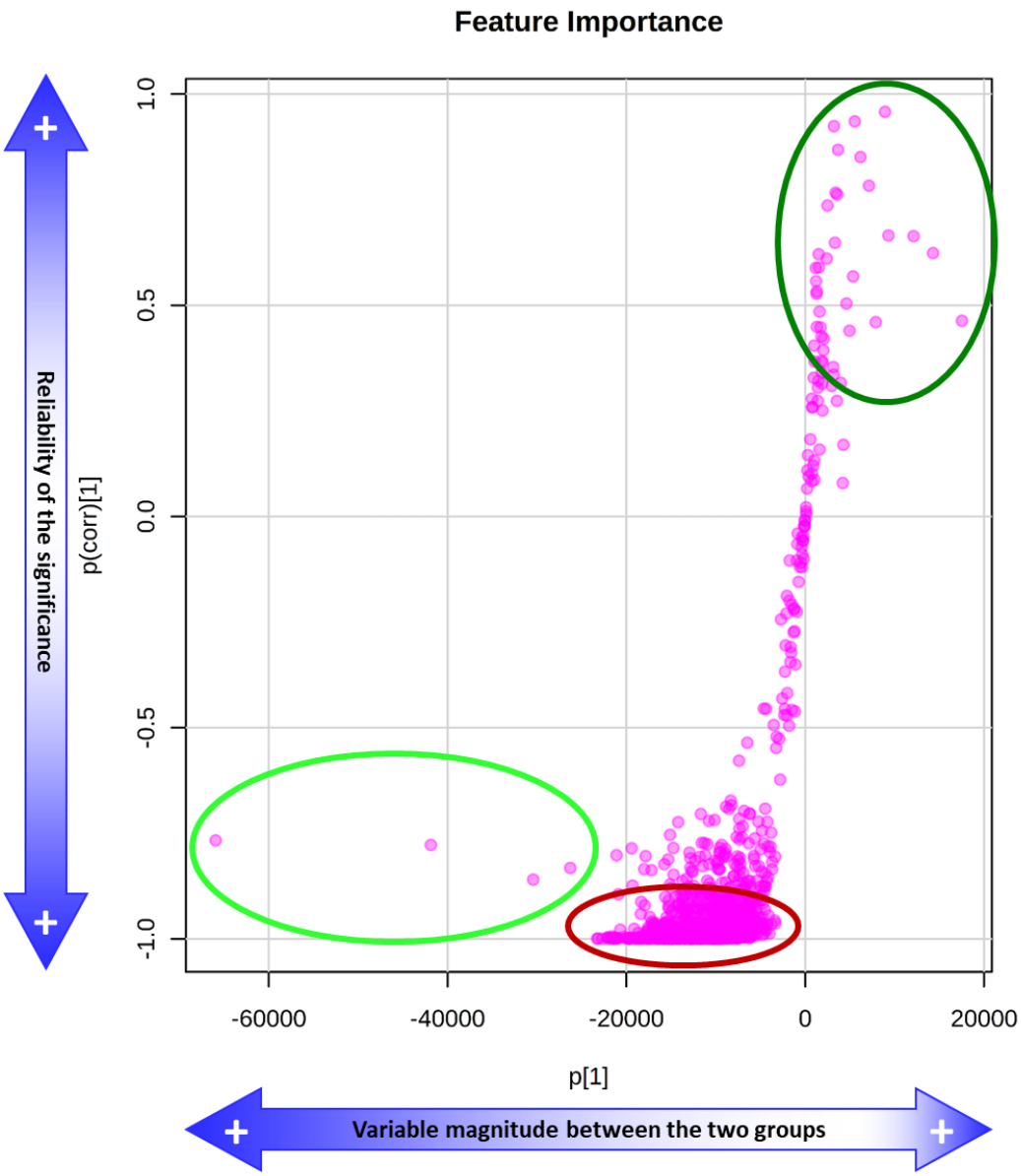


Figure A.II-A 2**:** *OPLS-DA S-Plot facilitates the mining of features of interest, since it correlates the variations of features to the studied factor [3], and introduces the two following dimensions: i) variables' magnitudes between the two conditions, and ii) the reliability/significance of their magnitudes, by assessing their "intra-condition" variations.*
Further is the feature from the 0 of the p[1] axis, higher is the magnitude of its variation between the two groups.
Further is the feature from the 0 of the p(corr)[1] axis, lower is its "intra-group" variation, thus the reliability of its variation significance is higher [4–6].
Plot generated using MetaboAnalyst.

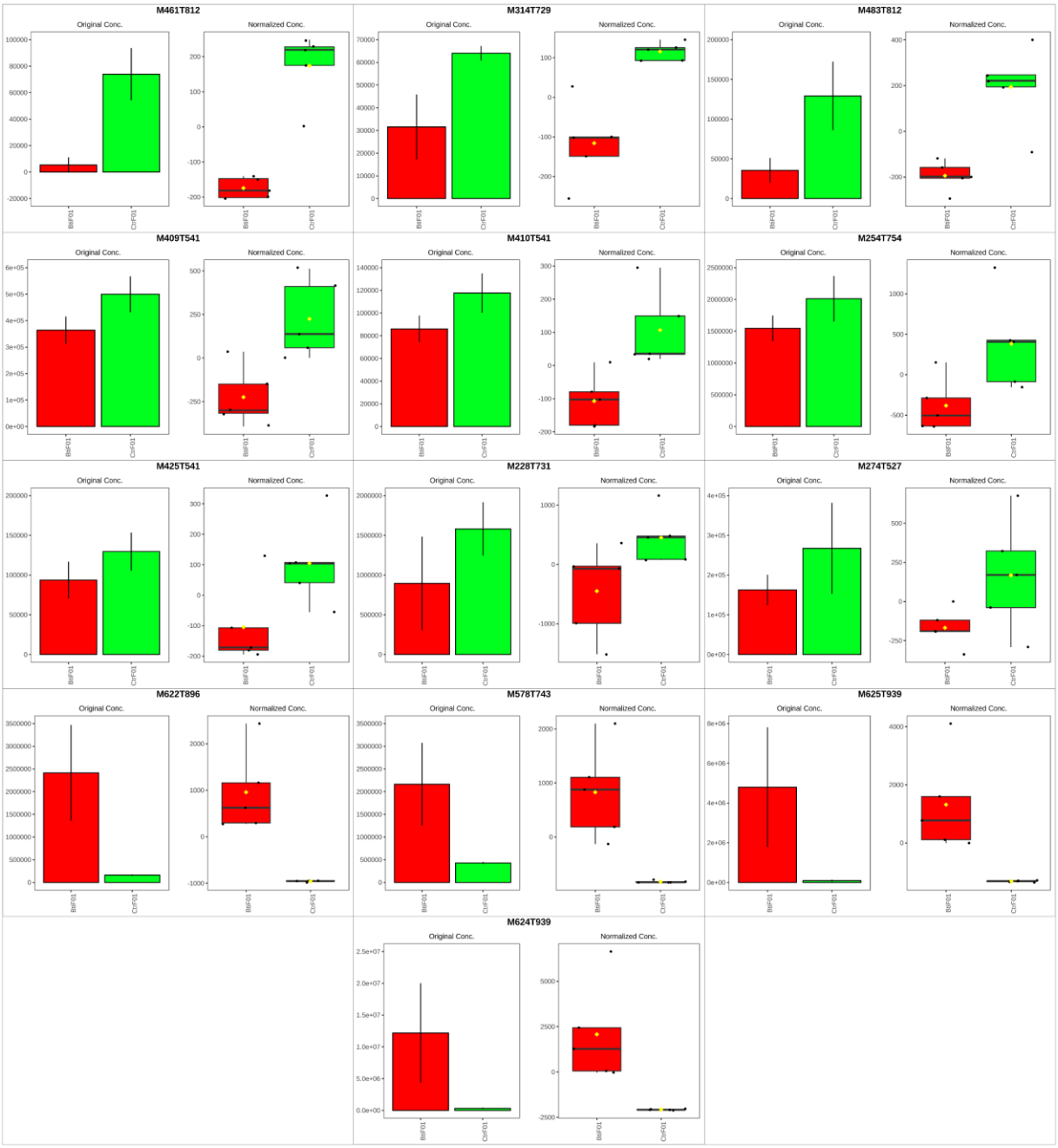


***Figure A.II-A 3**: Histograms and boxplots showing the relative abundances of the selected biomarker candidates, generated by the automated processing workflow.*

Histograms show the original peak area values. Boxplots show the normalized peak area values after Pareto scaling. The green refers to abundances in control samples, red is for abundances in spiked samples.

Plots are sorted according to the descending order of features' scores of significance ($-Log_{10}$[p-Val] of t-Test applied on automatically processed data – Table C.II 1), from the left to the right, and then from the top to the bottom. The first three rows represent plots of features overexpressed in control samples, last two rows are plots of features overexpressed in spiked samples.

Plots generated and t-Test performed using MetaboAnalyst.

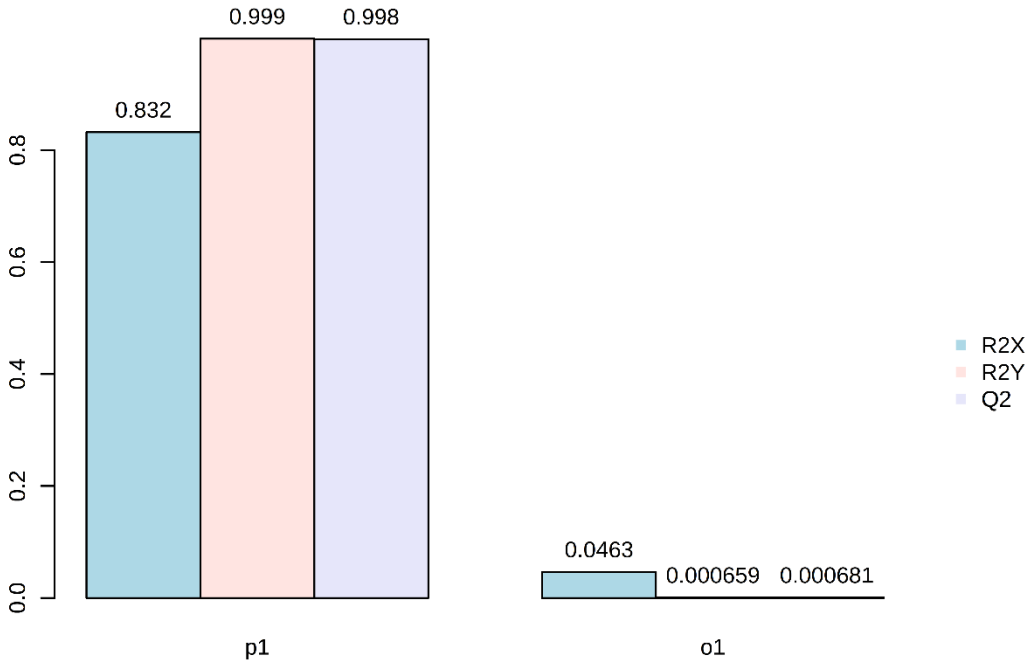


Figure A.II-A 4: *OPLS-DA Cross-Validation test showing that the model is tending to an overfitting due to the high redundancies generated by the multi-charged isotope patterns of the detected macromolecules: R2Y and Q2 of the p1 are almost equal, and the difference between R2X and R2Y is relatively low (< 0.30) [6–8].*
Plot generated using MetaboAnalyst.

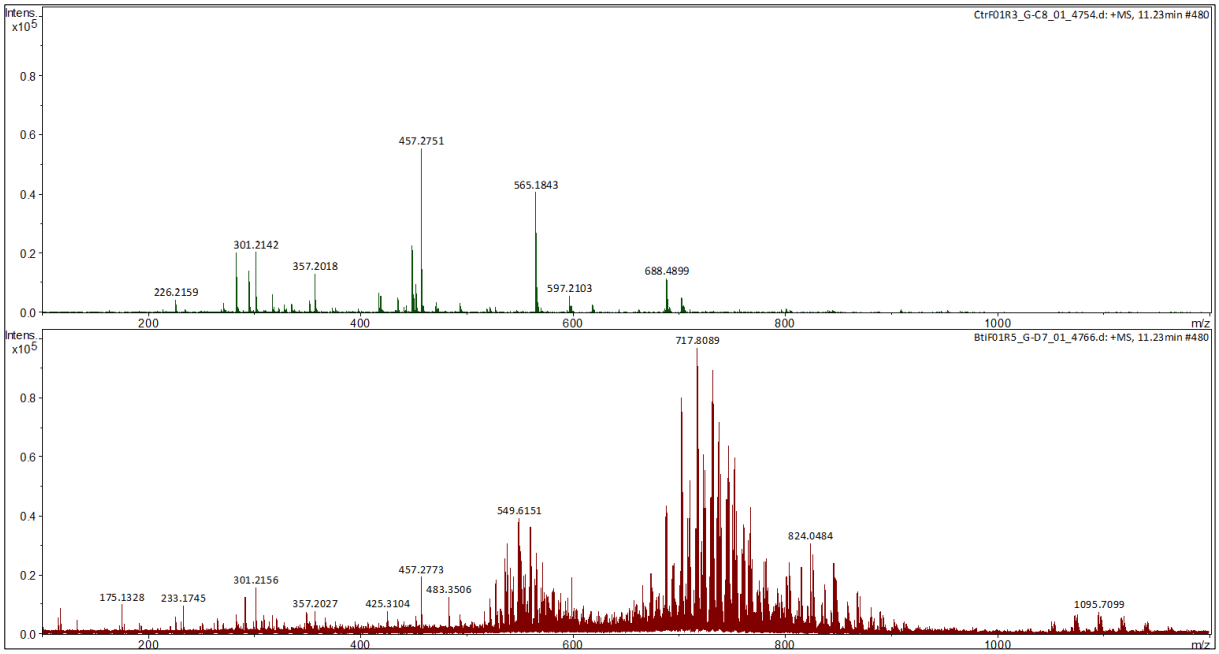


***Figure A.II-A 5**: Mass spectra of the contaminant M457T675 in the two studied conditions (control in green vs. spiked in red).*
Analytical data about the compound:
Experimental m/z mean: 457.2766 – RT: 11.25 min
Suggested M formula: $C_{23}H_{38}N_4O_4$ – $[M+Na]^+$ adduct – m/z error: -4.21 ppm
Spectra range: 100 m/z to 1200 m/z. Intensity scale is fixed to the same value in both conditions (1.05E5).
Spectra generated using Compass DataAnalysis 4.3 software.

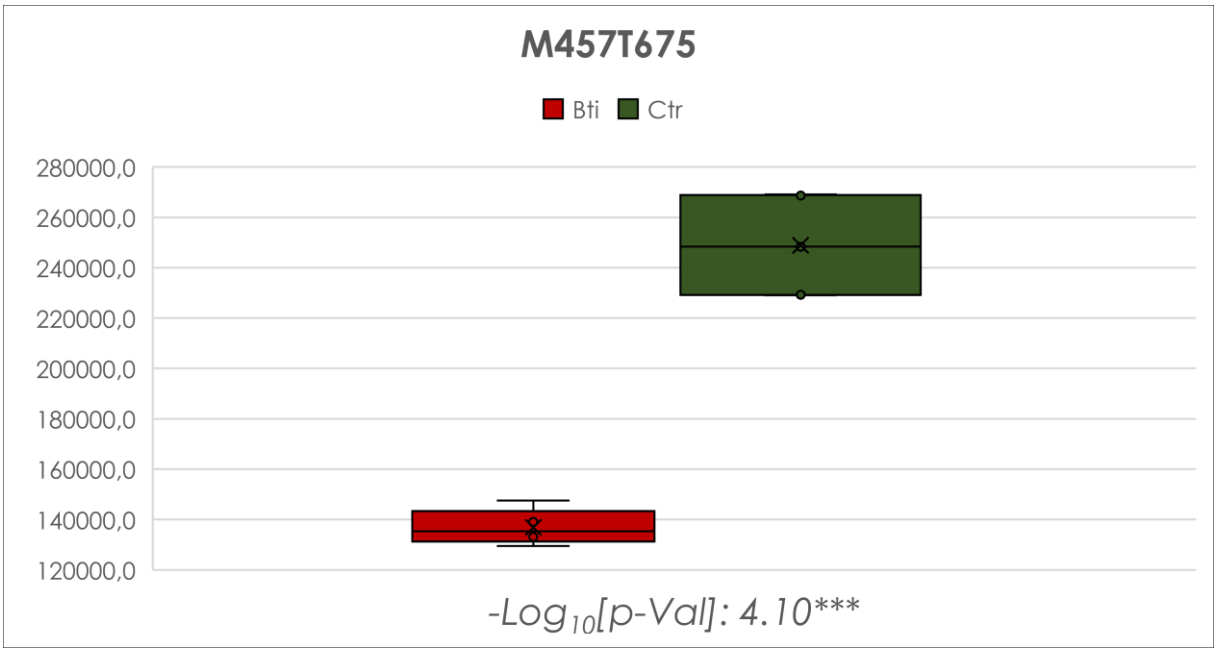


***Figure A.II-A 6**: Boxplot showing the relative abundances of the contaminant M457T675 in the two groups of samples.*
Vertical axis represents the EIC peak area of 457.2766 m/z ion.
The green refers to abundances in control samples, red is for abundances in spiked samples.

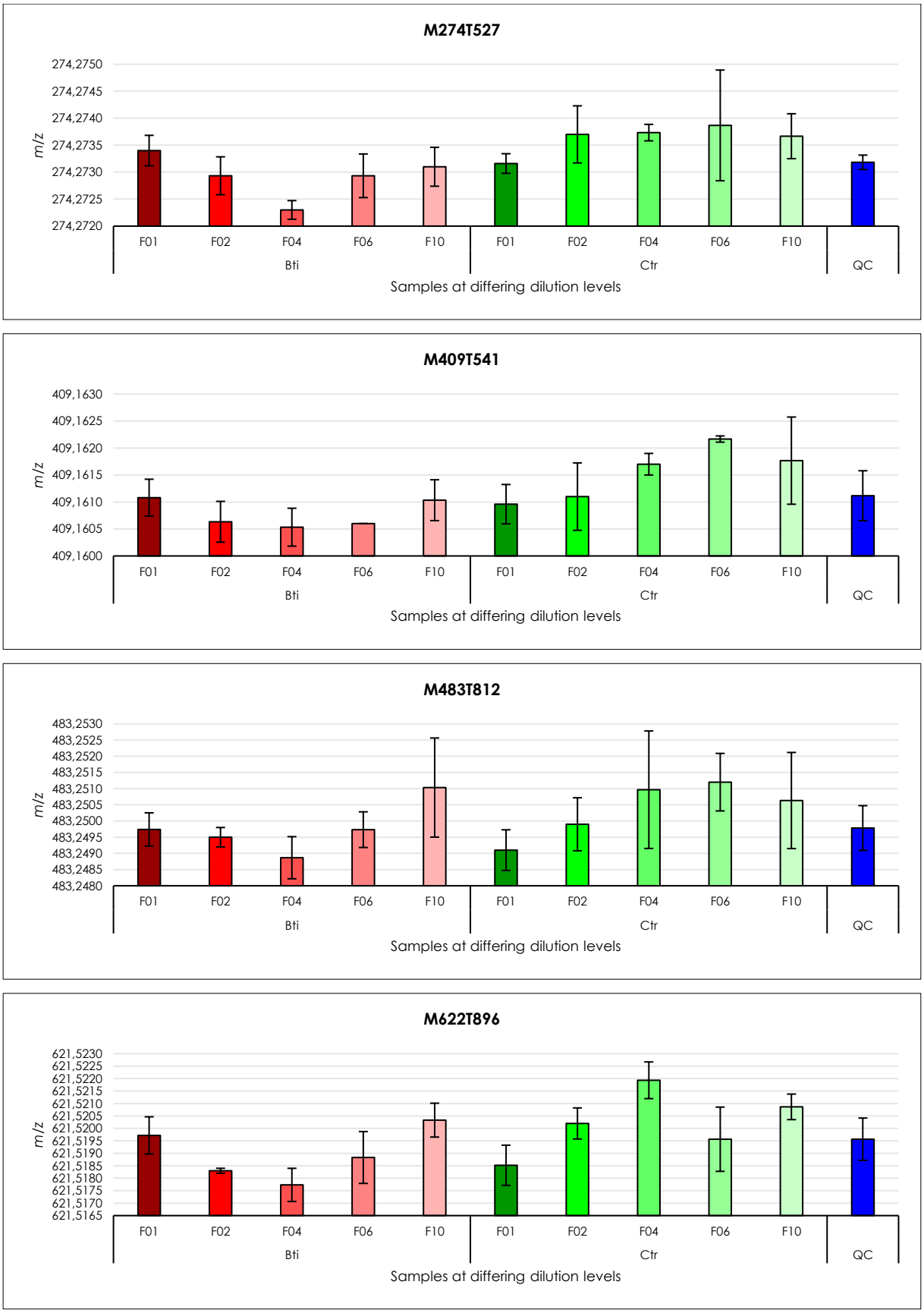


***Figure A.II-A 7**: Evolution of exact m/z measures through dilution levels: some examples.*
Histograms show m/z means and SDs calculated inter-replicates of each group (intra-group).
Legend: Bti: Spiked, Ctr: Control, F01: original samples, F02: 1/2 dilution, F04: 1/4 dilution, F06: 1/6 dilution, F10: 1/10 dilution.

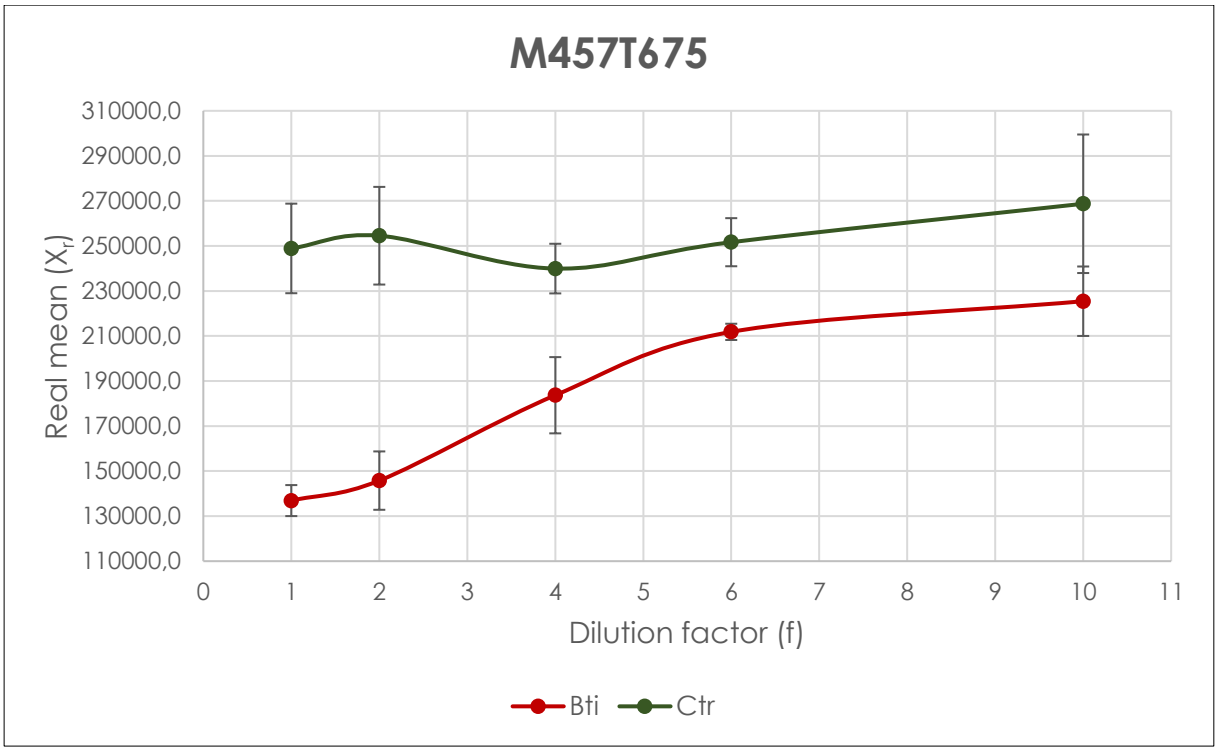


***Figure A.II-A 8**: EIC peak area real means evolution versus dilution factor ($\overline{X_r}$) for the contaminant M457T675.*
Green curves represent the evolution of means in control samples (Ctr). Red curves are for spiked samples (Bti).

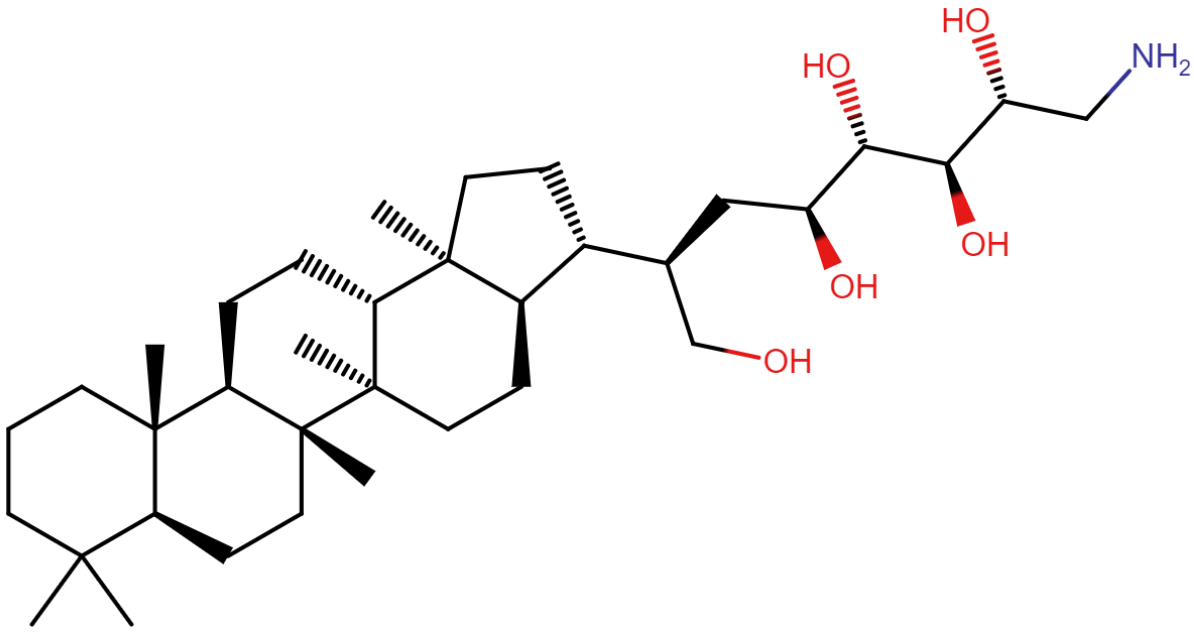


***Figure A.II-A 9**: The molecular structure proposed as a putative annotation for the feature M578T743.*

A.II-A.4. R scripts and command lines

A.II-A.4.1. Command lines for data processing and statistical analyses using the MetaboAnalyst Platform [9–11].

```
library("MetaboAnalystR")

# Dataset Upload

[1] mSet<-InitDataObjects("pktable", "stat", FALSE)
[2] mSet<-Read.TextData(mSet, "Replacing_with_your_file_path", "rowu",
"disc");

# Data Integrity Check

[3] mSet<-SanityCheckData(mSet)

# Data Filtering: no filtering was performed

[4] mSet<-ReplaceMin(mSet);
[5] mSet<-PreparePrenormData(mSet)

# Data Normalization: Pareto Scaling

[6] mSet<-Normalization(mSet, "NULL", "NULL", "ParetoNorm", ratio=FALSE,
ratioNum=20)
[7] mSet<-PlotNormSummary(mSet, "norm_0_", "png", 72, width=NA)
[8] mSet<-PlotSampleNormSummary(mSet, "snorm_0_", "png", 72, width=NA)

# Setting colors and shapes to samples for plots

[9] colVec<-c("#ff0000","#00ff00","#0000ff")
[10] shapeVec<-c(20,20,20)
[11] mSet<-UpdateGraphSettings(mSet, colVec, shapeVec)

# Principal Component Analysis (PCA)

[12] mSet<-PCA.Anal(mSet)
[13] mSet<-PlotPCA2DScore(mSet, "pca_score2d_0_", "png", 600, width=NA,
1,2,0.95,0,0)
[14] mSet<-UpdatePCA.Loading(mSet, "none");
[15] mSet<-PlotPCALoading(mSet, "pca_loading_1_", "png", 600, width=NA,
1,2);

# Elimination of QC group from the dataset

[16] feature.nm.vec <- c("")
[17] smpl.nm.vec <- c("")
[18] grp.nm.vec <- c("QC")
[19] mSet<-UpdateData(mSet)
[20] mSet<-PreparePrenormData(mSet)

# Data Normalization: Pareto Scaling (mandatory step after dataset editing)

[21] mSet<-Normalization(mSet, "NULL", "NULL", "ParetoNorm", ratio=FALSE,
ratioNum=20)
[22] mSet<-PlotNormSummary(mSet, "norm_1_", "png", 72, width=NA)
[23] mSet<-PlotSampleNormSummary(mSet, "snorm_1_", "png", 72, width=NA)
```

```
# Orthogonal Partial Least Squares - Discriminant Analysis (orthoPLS-DA)

[24] mSet<-OPLSR.Anal(mSet, reg=TRUE)
[25] mSet<-PlotOPLS2DScore(mSet, "opls_score2d_0_", "png", 600, width=NA,
1,2,0.95,0,0)
[26] mSet<-UpdateOPLS.Splot(mSet, "none");
[27] mSet<-PlotOPLS.Splot(mSet, "opls_splot_1_", "none", "png", 600,
width=NA);
[28] mSet<-PlotOPLS.MDL(mSet, "opls_mdl_0_", "png", 600, width=NA)

# Histograms and Boxplots generation for the relevant features

[29] mSet<-PlotCmpdSummary(mSet, "M624T939", "png", 600, width=NA)
[30] mSet<-PlotCmpdSummary(mSet, "M625T939", "png", 600, width=NA)
[31] mSet<-PlotCmpdSummary(mSet, "M622T896", "png", 600, width=NA)
[32] mSet<-PlotCmpdSummary(mSet, "M578T743", "png", 600, width=NA)
[33] mSet<-PlotCmpdSummary(mSet, "M461T812", "png", 600, width=NA)
[34] mSet<-PlotCmpdSummary(mSet, "M483T812", "png", 600, width=NA)
[35] mSet<-PlotCmpdSummary(mSet, "M314T729", "png", 600, width=NA)
[36] mSet<-PlotCmpdSummary(mSet, "M409T541", "png", 600, width=NA)
[37] mSet<-PlotCmpdSummary(mSet, "M410T541", "png", 600, width=NA)
[38] mSet<-PlotCmpdSummary(mSet, "M254T754", "png", 600, width=NA)
[39] mSet<-PlotCmpdSummary(mSet, "M425T541", "png", 600, width=NA)
[40] mSet<-PlotCmpdSummary(mSet, "M228T731", "png", 600, width=NA)
[41] mSet<-PlotCmpdSummary(mSet, "M274T527", "png", 600, width=NA)

# T-test: Group variance: Unequal

[42] mSet<-Ttests.Anal(mSet, F, 1.0, FALSE, FALSE, FALSE)
```

A.II-A.4.2. Command lines for the Welch Two Sample t-Test (R Commander 2.4-2 package) [12].

```
library(Rcmdr)

t.test(Area~Class, alternative='two.sided', conf.level=.95,
var.equal=FALSE, data=dataset.name)
```

References

[1] S. Beck, A. Michalski, O. Raether, M. Lubeck, S. Kaspar, N. Goedecke, C. Baessmann, D. Hornburg, F. Meier, I. Paron, N.A. Kulak, J. Cox, M. Mann, The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics, Mol Cell Proteomics. 14 (2015) 2014–2029. https://doi.org/10.1074/mcp.M114.047407.

[2] C. Patil, C. Calvayrac, Y. Zhou, S. Romdhane, M.-V. Salvia, J.-F. Cooper, F.E. Dayan, C. Bertrand, Environmental Metabolic Footprinting: A novel application to study the impact of a natural and a synthetic β-triketone herbicide in soil, Science of The Total Environment. 566–567 (2016) 552–558. https://doi.org/10.1016/j.scitotenv.2016.05.071.

[3] J. Boccard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Analytica Chimica Acta. 769 (2013) 30–39. https://doi.org/10.1016/j.aca.2013.01.022.

[4] J. Cohen, Things I have learned (so far)., American Psychologist. 45 (1990) 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304.

[5] A. Roux, Analysis of human urinary metabolome by liquid chromatography coupled to high resolution mass spectrometry, Thèse de Doctorat, Université Pierre et Marie Curie - Paris VI, 2011. https://tel.archives-ouvertes.fr/tel-00641529 (accessed October 18, 2020).

[6] S. Wiklund, Multivariate data analysis for Omics, (2008). https://metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf (accessed October 18, 2020).

[7] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, NY, 1966: pp. 391–420. https://ci.nii.ac.jp/naid/20001378860/en/ (accessed October 18, 2020).

[8] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics. 16 (2002) 119–128. https://doi.org/10.1002/cem.695.

[9] MetaboAnalyst, MetaboAnalyst. (n.d.). https://www.metaboanalyst.ca/ (accessed February 28, 2020).

[10] J. Chong, D.S. Wishart, J. Xia, Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis, Current Protocols in Bioinformatics. 68 (2019) e86. https://doi.org/10.1002/cpbi.86.

[11] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Research. 46 (2018) W486–W494. https://doi.org/10.1093/nar/gky310.

[12] J. Fox, The R Commander: A Basic-Statistics Graphical User Interface to R, Journal of Statistical Software. 14 (2005) 1–42. https://doi.org/10.18637/jss.v014.i09.

Appendix II-B

For Chapter II

Electrospray Ionization and heterogeneous matrix effects in Liquid Chromatography-Mass Spectrometry-based meta-metabolomics: A biomarker or a suppressed ion?

Appendix II-B

Hikmat Ghosson[1,2,*], Yann Guitton[3], Amani Ben Jrad[1,2], Chandrashekhar Patil[1], Delphine Raviglione[1,2], Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,4,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), Oniris, INRAE, Nantes, F-44307, France

4: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: Equal Contribution (Last Co-authors)

Putative identification of the relevant biomarker candidates

A.II-B.1. Full HRMS and MS/MS acquisitions

After prioritizing the relevant features and considering them as validated biomarker candidates (as described in Section 3.4. – Chapter II), putative identifications of these features were conducted. They are based on two main information: i) the elemental composition of the ions, determined by the exact *m/z* measures, the adduct types and the isotope patterns, and ii) the fragmentation patterns acquired after MS/MS experiments.

After screening sediment extracts profiles of both the control and the spiked samples by LC-ESI-Q/ToF analyses (ESI+), RT ranges belonging to the features of interest were defined for the Tandem Mass Spectrometry experiments. To acquire reliable fragmentation patterns, the most intense ions belonging to the "metabolite" were selected for MS/MS. The MS/MS experiments using the LC-Q/ToF system are described in Section 2.7.1. – Chapter II. Results were as the following: MS/MS experiments succeeded for only 3 of the selected features, *i.e.*

for M578T743, M622T896 and M624T939. For each of the 3 features, 4 MS/MS spectra were separately acquired at 4 different levels of CID energy. They are shown with the Full HRMS spectra in Figures A.II-B 3-A.II-B 17.

For the remaining feature (*e.g.* M274T527), no fragments could be detected. Thus, as the LC-Q/ToF MS/MS experiments have failed to provide sufficient information, acquisitions using the LC-Q/Orbitrap were performed. As described in Section 2.7.2. – Chapter II. an ESI+ screening was first achieved to acquire new Full HRMS data at a higher resolution (theatrically 140000 at *m/z* 200, and then decays following the function $R \propto \sqrt{(m/z)^{-1}}$[22]). These Full HRMS acquisitions provided precise measures for the exact *m/z* and the isotope patterns of the ions of interest. They were in concordance with the Q/ToF results for the 4 validated biomarker candidates (as well as for the suppressed ions). The Full HRMS screening also allowed defining the RT ranges of the features for MS/MS experiments. For each of the 4 metabolites, the most intense ion was fragmented at 3 different levels of HCD energy (methods described in Section 2.7.2. – Chapter II). All MS/MS experiments succeeded to provide fragmentation patterns. The acquired Full HRMS and MS/MS spectra are presented in Figures A.II-B 18-A.II-B 33.

For further investigations, a Full HRMS ESI– screening was performed. Only 3 features were detected in the negative mode, *i.e.* M578T743, M622T896 and M624T939. Their detected ions (Table A.II-A 4 – Appendix II-A) were fragmented and their acquired Full HRMS and MS/MS spectra are shown in Figures A.II-B 34-A.II-B 45.

It should be mentioned that ESI– Q/ToF acquisitions were performed and no ions belonging to the features of interest could be detected.

A.II-B.2. Computational putative identification

The acquired Full HRMS and MS/MS data were manually transformed to ".ms" data as explained in Figure A.II-B 1. In brief, a .ms file was created from each acquisition type (*i.e.* ESI+ Q/ToF, ESI+ Q/Orbitrap, or ESI– Q/Orbitrap). The file included the name of the feature, the exact *m/z* belonging to the "parent" ion (the precursor), the adduct type, the isotope pattern (the exact *m/z* and the relative intensity of each detected isotope), the applied collision energies, and the MS/MS spectra acquired at each energy level (*i.e.* the exact *m/z* ratios and the intensities of the detected fragments). After creating the .ms files, they were imported into the SIRIUS 4.4.29 software[23] in order to perform a computational spectral data processing. The processing

[22] Zubarev, R. A. & Makarov, A. *Anal. Chem.* **(2013)**, 85(11):5288–5296. doi:10.1021/ac4001223
[23] Dührkop, K. *et al. Nat. Methods* **(2019)**, 16(4):299–302. doi:10.1038/s41592-019-0344-8

applied for each .ms file (*i.e.* for each acquisition type) is described in Figure A.II-B 2 and Section 2.8. – Chapter II. It aims first to determine the elemental composition of the given ion. Then, a database search is performed in order to putatively annotate the metabolite, based on the suggested elemental composition of its ion. After, a computational process is performed in order to assess the probability of the annotated candidate to fragment according to the fragments-based proposed fragmentation tree. For each feature, fragmentations trees are shown in Figures A.II-B 46-A.II-B 55. The annotations are shown in Table C.II 2 of the main manuscript.

Note: *Only Figure A.II-B 1 and Figure A.II-B 2 are presented in the current manuscript. The remaining Figures could not be included due to hardcopy formatting issues. Nonetheless, they are published as Supporting Information for the article by Ghosson et al.*[24]*, and accessible online via the following link:*

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Frcm.8977&file=rcm8977-sup-0002-AppendixB.pdf.

Or on publisher's website via the Digital Object Identifier below:

https://doi.org/10.1002/rcm.8977.

[24] Ghosson, H. *et al. Rapid Commun. Mass Spectrom.* **(2021)**, 35(2):e8977. doi:10.1002/rcm.8977

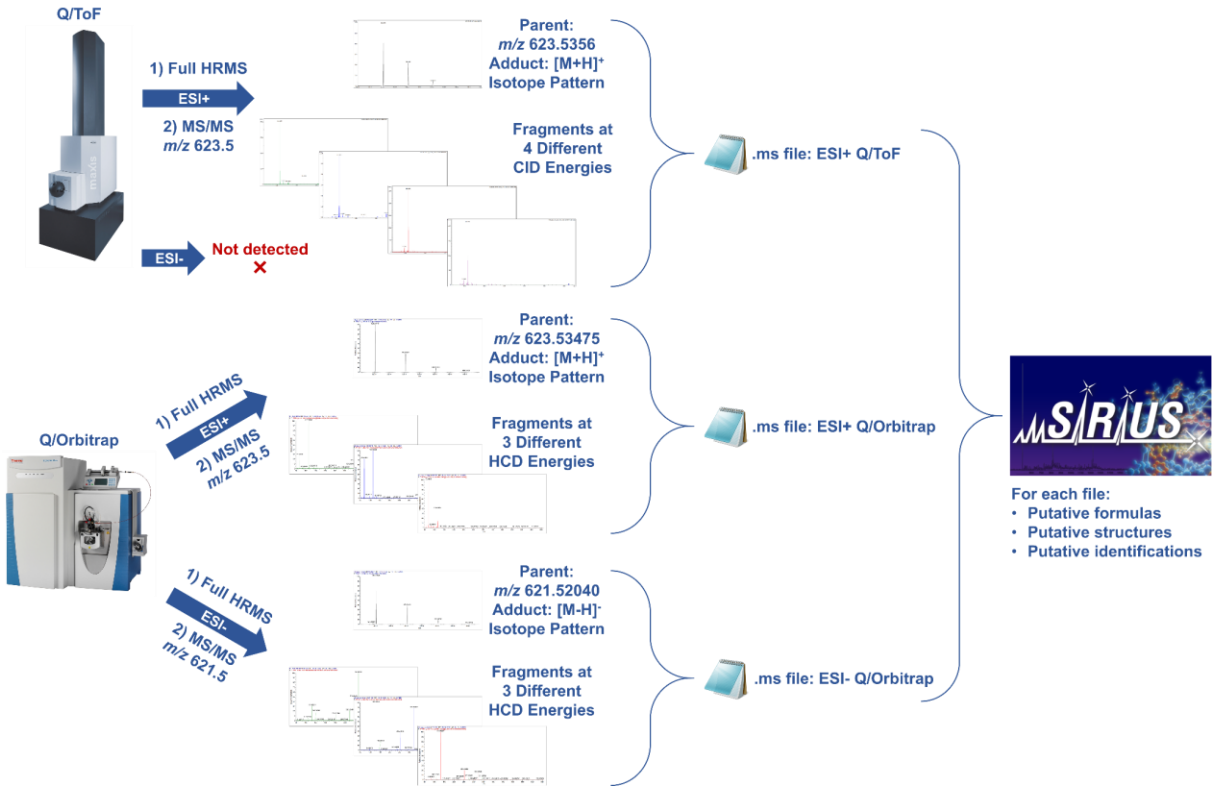


***Figure A.II-B 1**: Creation of .ms files from the different acquisitions.*
The given example is for the feature M624T939.

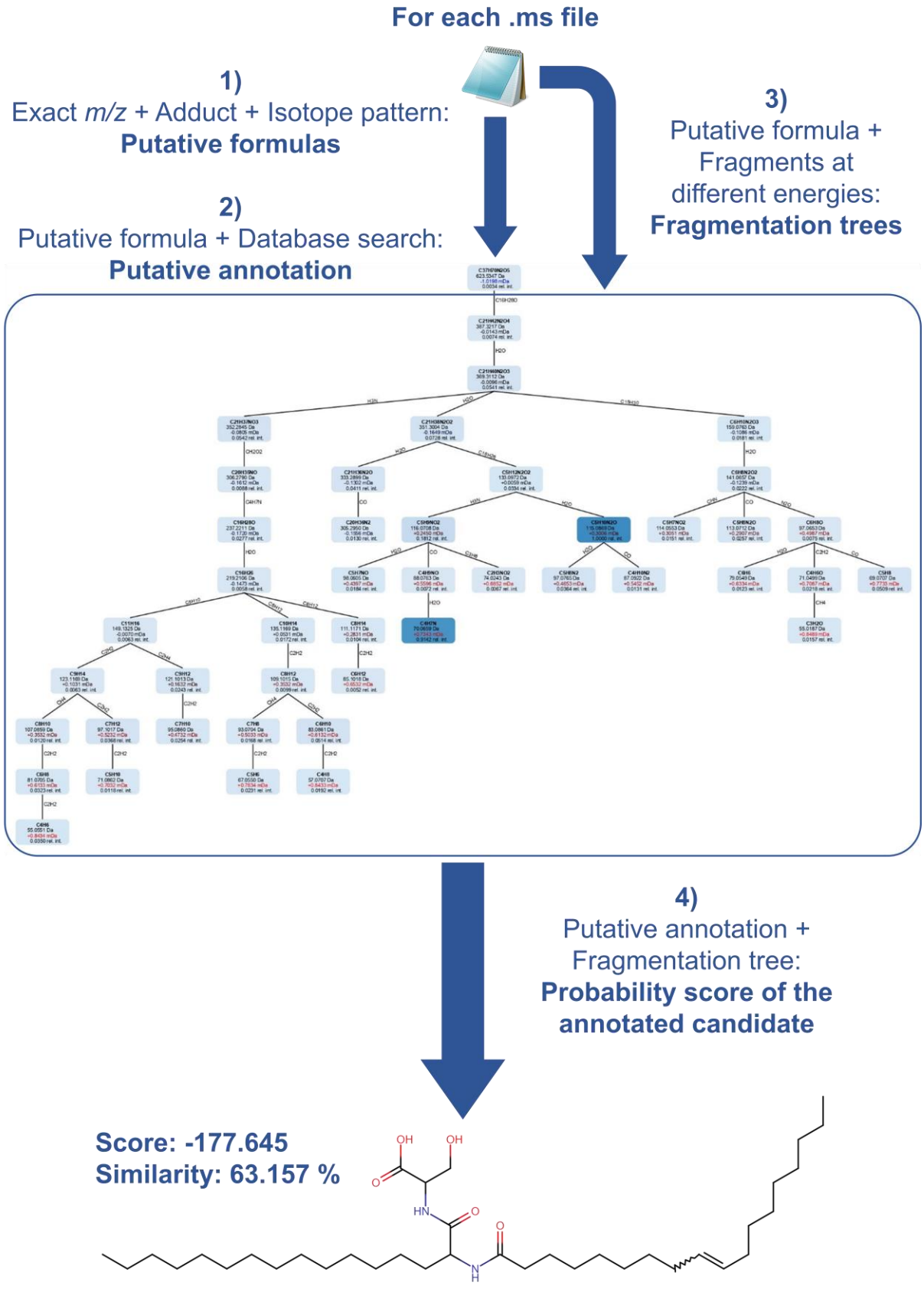


***Figure A.II-B 2**: From a .ms file to a putative annotation.*
The given example is for the feature M624T939.

Appendix III-A

For Chapter III

Online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted volatile metabolomics for studying emerging complex biopesticides: A proof of concept

Appendix III-A

Hikmat Ghosson[1,2,*], Delphine Raviglione[1,2], Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,3,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: *Equal Contribution (Last Co-authors)*

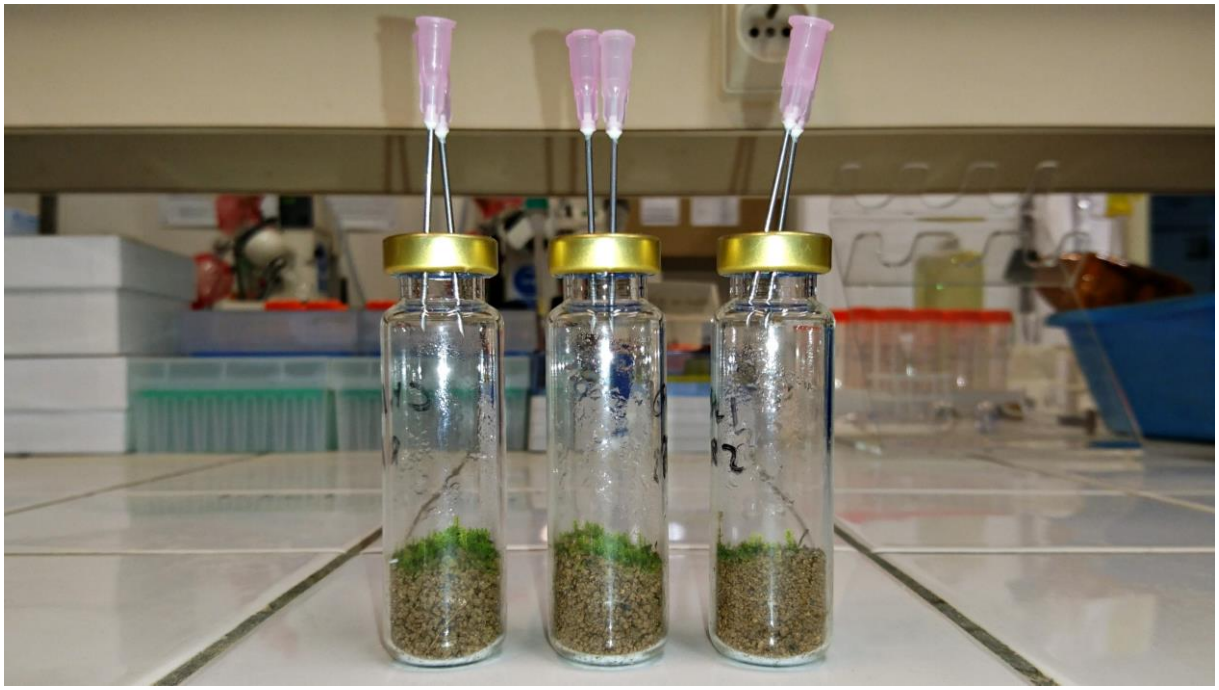


Figure A.III-A 1: *Soil samples set-up: a living microcosm in a 20 mL vial.*

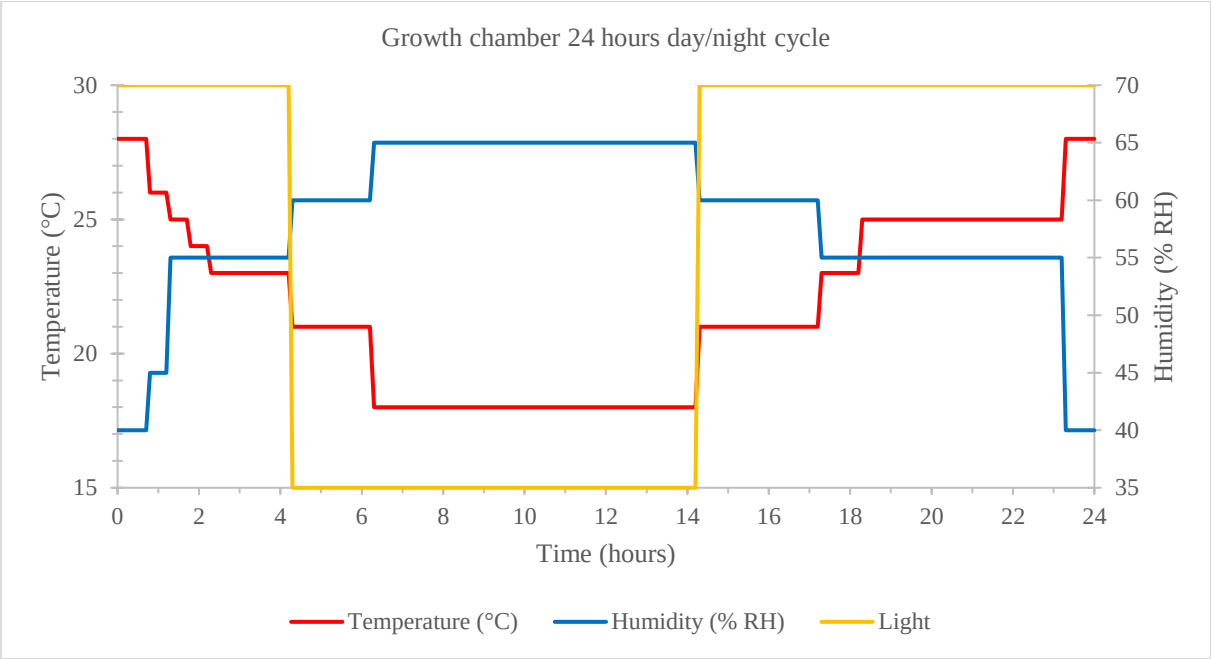


***Figure A.III-A 2**: Growth chamber 24 hours day/night cycle described in Section 2.3. – Chapter III (Soil samples set-up).*
When the orange line is upside: the light is ON (day). When its downside: the light is OFF (night).

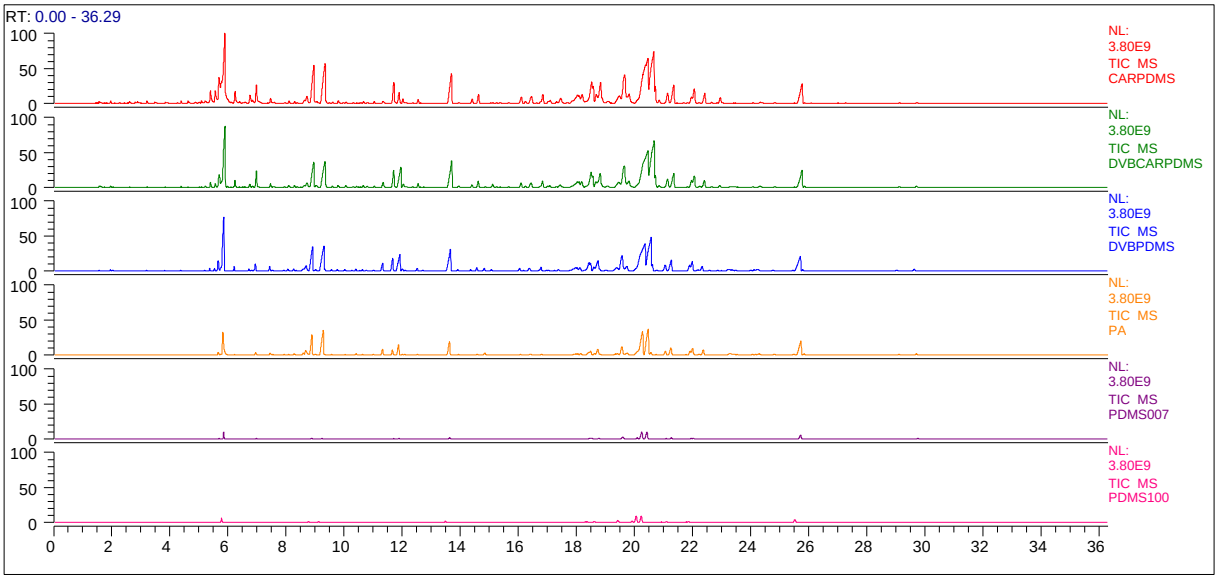


***Figure A.III-A 3**: Tests of different SPME fiber coatings.*
m/z 40-400 Total Ion Chromatograms (TICs) are represented for each fiber type with a fixed scale set to 3.80E9.
Colors code:
Red: 85 µm CAR/PDMS
Green: 50/30 µm DVB/CAR/PDMS
Blue: 65 µm PDMS/DVB
Orange: 85 µm PA
Purple: 7 µm PDMS
Pink: 100 µm PDMS
Chromatograms performed using Xcalibur 4.1.31.9 software.

Table A.III-A 1: *Total TIC area integrations and numbers of molecular features for the different SPME coatings tests.*

Fiber Type	Total TIC area	Number of Molecular Features
85 µm CAR/PDMS	2.35×10^{11}	1095
50/30 µm DVB/CAR/PDMS	1.82×10^{11}	937
65 µm PDMS/DVB	1.27×10^{11}	856
85 µm PA	7.46×10^{10}	731
7 µm PDMS	1.16×10^{10}	249
100 µm PDMS	9.96×10^{09}	224

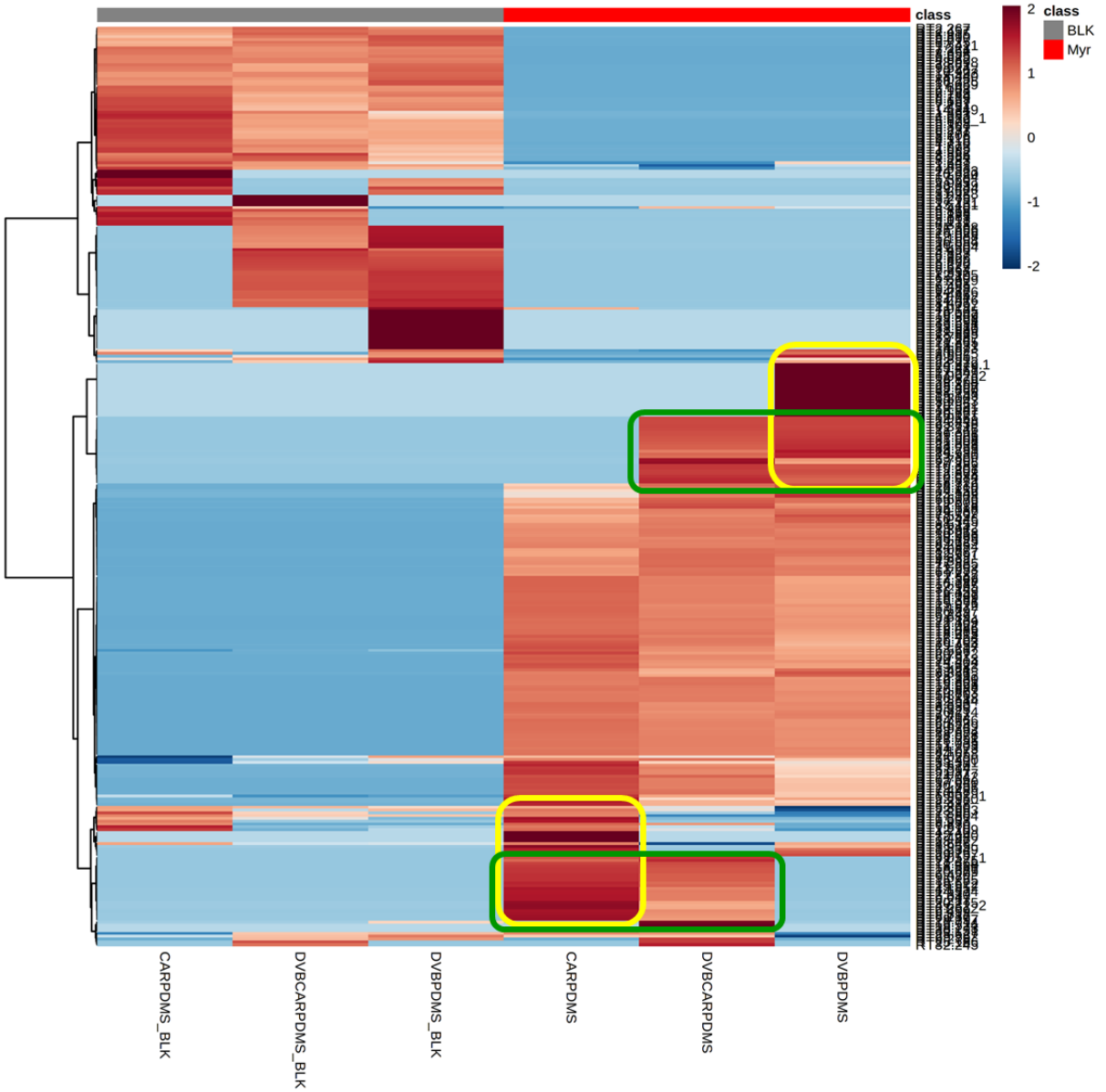


Figure A.III-A 4: *Heatmap analysis applied for in-depth assessment of performances of PDMS/DVB, DVB/CAR/PDMS and CAR/PDMS coatings.*

The grey bloc on the left of the figure represents blank injections (empty vials). The red bloc on the right represents HS-SPME-GC-MS analyses applied on spiked soil samples.

Intensities of features are standardized/auto-scaled (each feature's intensity is divided by the sum of its intensities in all samples). Clustering algorithm: Ward, distance measure: Euclidean. Only features are clustered.

Plot generated using MetaboAnalyst.

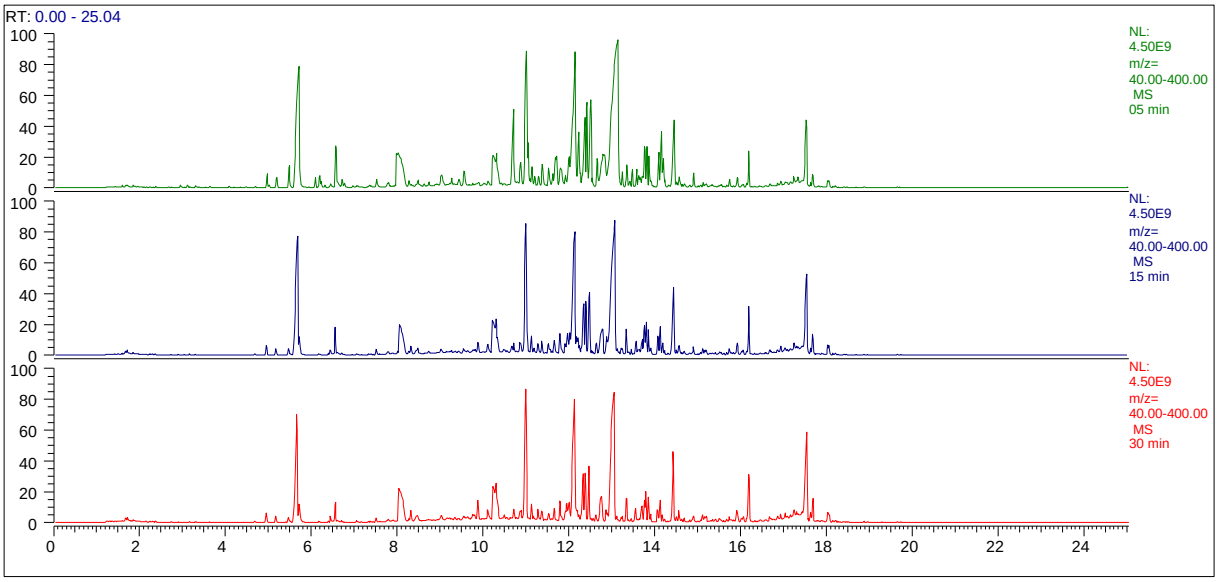


***Figure A.III-A 5**: Chromatograms of the different incubation time tests.*
Colors code: Green chromatogram is for an incubation time of 5 min. Blue is for 15 min. Red is for 30 min.
Chromatograms consist of an m/z 40-400 TIC. A fixed intensity scale is applied for all chromatograms (4.50E9).
A classical GC-MS screening method was applied using an Agilent J&W DB-FFAP GC column (30 m × 0.250 mm × 0.25 µm).
Initial temperature of 40 °C was held for 2 min before the application of a single oven temperature ramp of 10 °C min^{-1} until reaching a temperature of 230 °C that was held for 4 min before the end of the run (previous unpublished work).
Chromatograms performed using Xcalibur 4.1.31.9 software.

***Table A.III-A 2**: Total TIC area integrations and numbers of molecular features for the different incubation time tests.*

Incubation time (min)	Total TIC area	Number of Molecular Features
05	2.59×10^{11}	786
15	1.93×10^{11}	640
30	1.90×10^{11}	666

***Table A.III-A 3**: EICs area integrations for L-α-bornyl acetate, epi-γ-Eudesmol and α-Terpineol through the different incubation times.*
Compounds were putatively identified by EI spectral library search (NIST).

Incubation time (min)	Area integrations of EICs of major ions in each compound pseudo-spectra		
	L-α-bornyl acetate	***epi-γ-Eudesmol***	***α-Terpineol***
05	3.64×10^{09}	9.02×10^{08}	1.05×10^{10}
15	3.25×10^{08}	5.86×10^{08}	9.42×10^{09}
30	1.40×10^{08}	5.64×10^{08}	8.47×10^{09}

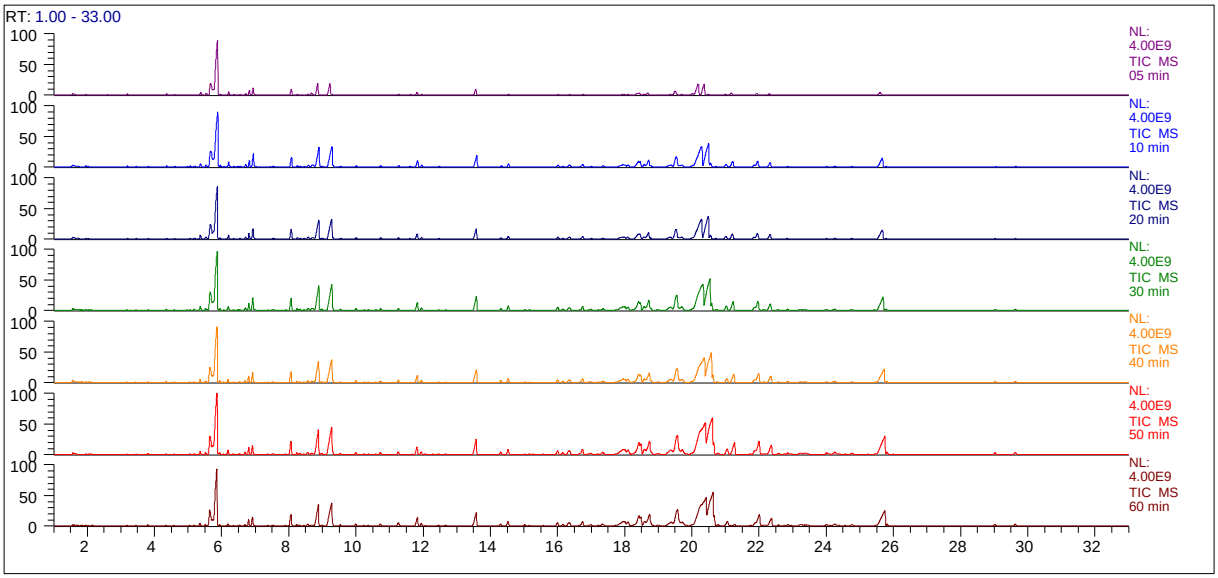


Figure A.III-A 6: *Chromatograms of the different extraction time tests.*
Colors code:
Purple: 5 min
Blue: 10 min
Dark blue: 20 min
Green: 30 min
Orange: 40 min
Red: 50 min
Brown: 60 min
m/z 40-400 Total Ion Chromatograms (TICs) are represented for each fiber type with a fixed scale set to 3.80E9 and a zoom on 1-33 min of RT. The used GC-MS method is the same described in Section 2.5. – Chapter III.
Chromatograms performed using Xcalibur 4.1.31.9 software.

Table A.III-A 4: *Total TIC area integrations and numbers of molecular features for the different extraction time tests.*

Extraction time (min)	Total TIC area	Number of Molecular Features
05	5.11×10^{10}	552
10	1.17×10^{11}	866
20	1.07×10^{11}	843
30	1.66×10^{11}	1042
40	1.51×10^{11}	969
50	2.03×10^{11}	1090
60	1.74×10^{11}	1040

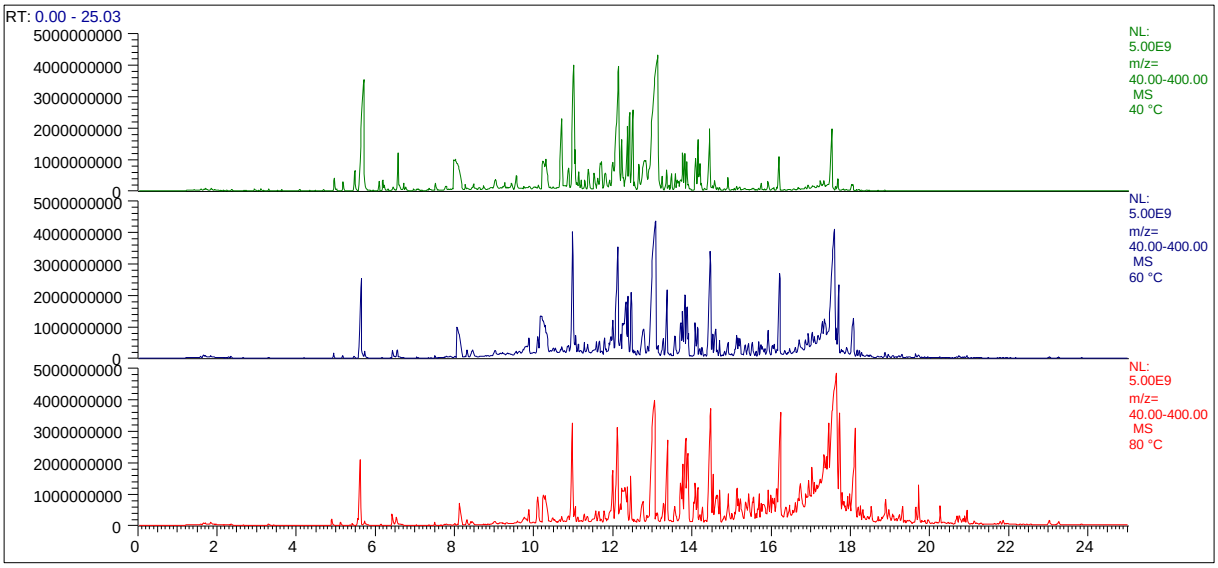


***Figure A.III-A 7**: Tests of different HS-SPME extraction temperatures.*
Colors code: Green chromatogram is for an extraction temperature of 40 °C. Blue is for 60 °C. Red is for 80 °C. Chromatograms consist of an m/z 40-400 TIC. A fixed intensity scale is applied for all chromatograms (5.00E9). A classical GC-MS screening method was applied using an Agilent J&W DB-FFAP GC column (30 m × 0.250 mm × 0.25 µm). Initial temperature of 40 °C was held for 2 min before the application of a single oven temperature ramp of 10 °C min^{-1} until reaching a temperature of 230 °C that was held for 4 min before the end of the run (previous unpublished work). Chromatograms performed using Xcalibur 4.1.31.9 software.

***Table A.III-A 5**: TIC area integrations for extraction temperature tests.*

Extraction temperature (°C)	**TIC area between 40-130 °C**	**TIC area between 130-230 °C**
40	9.04×10^{10}	1.72×10^{11}
60	7.57×10^{10}	2.64×10^{11}
80	5.16×10^{10}	4.10×10^{11}

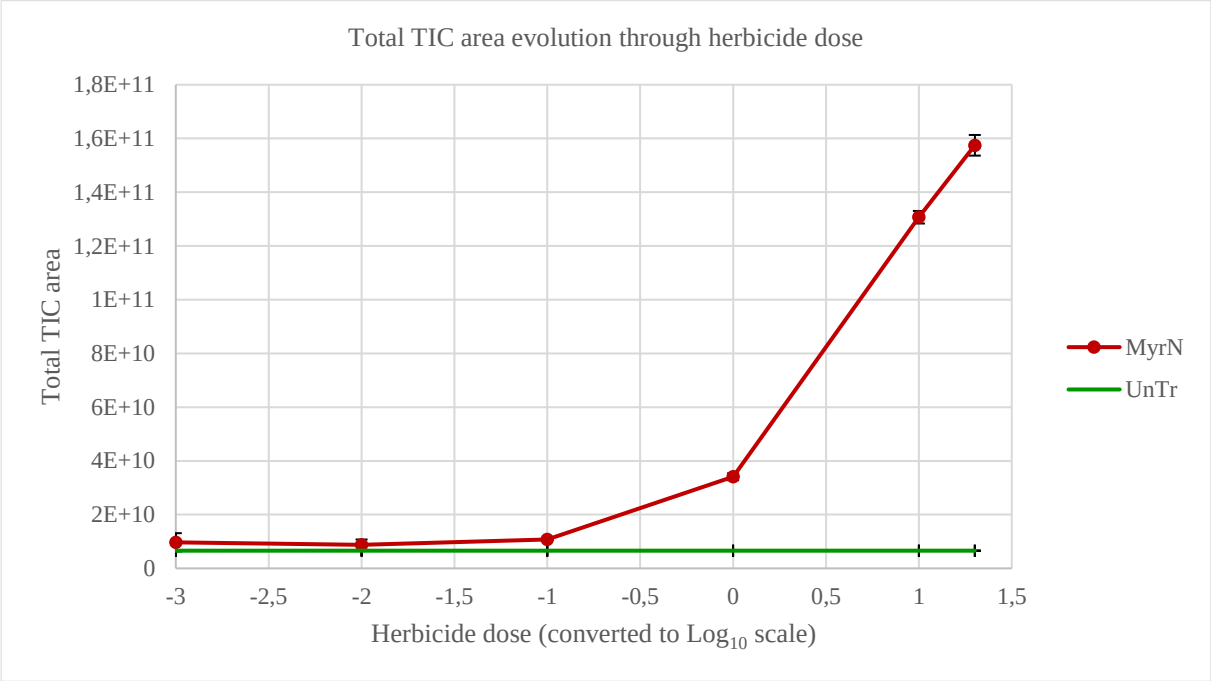


***Figure A.III-A 8**: Total TIC area evolution through the applied herbicide dose.*
Points of the applied herbicide doses on the graph (after Log_{10} scaling):
*10^{-3}-time: **-3**; 10^{-2}-time: **-2**; 10^{-1}-time: **-1**; 1-time: **0**; 10-times: **1**; 20-times: **1.30***

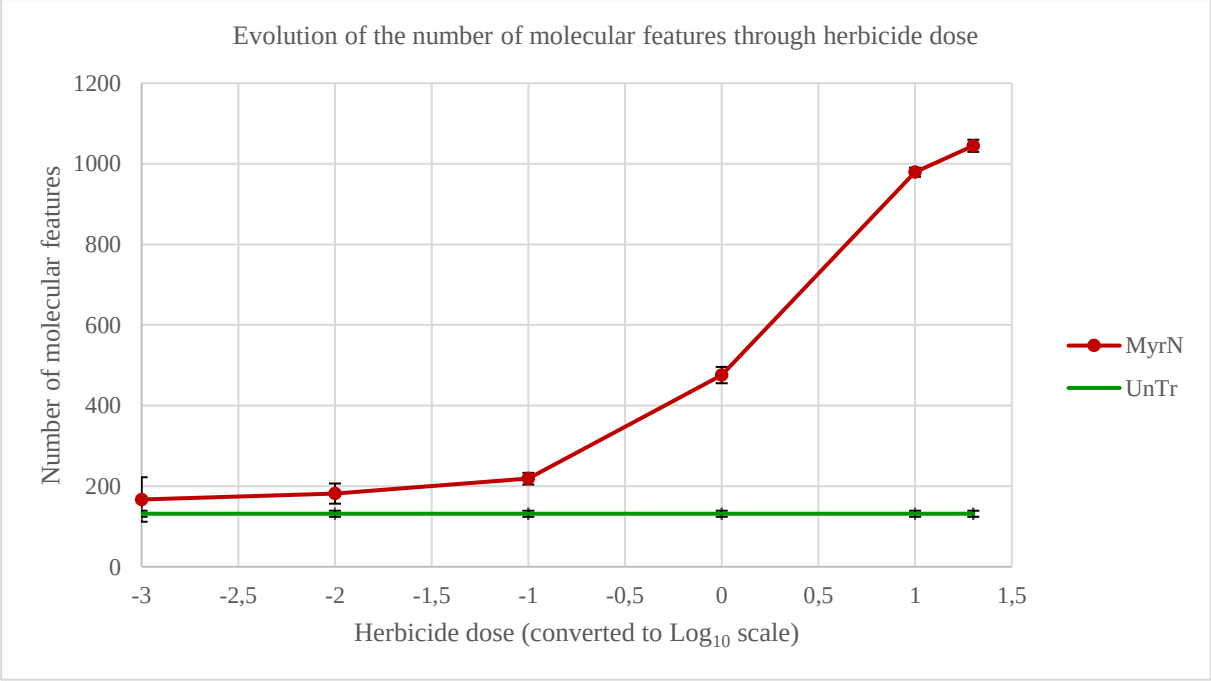


***Figure A.III-A 9**: The evolution of the number of molecular features through the applied herbicide dose.*
Points of the applied herbicide doses on the graph (after Log_{10} scaling):
*10^{-3}-time: **-3**; 10^{-2}-time: **-2**; 10^{-1}-time: **-1**; 1-time: **0**; 10-times: **1**; 20-times: **1.30***

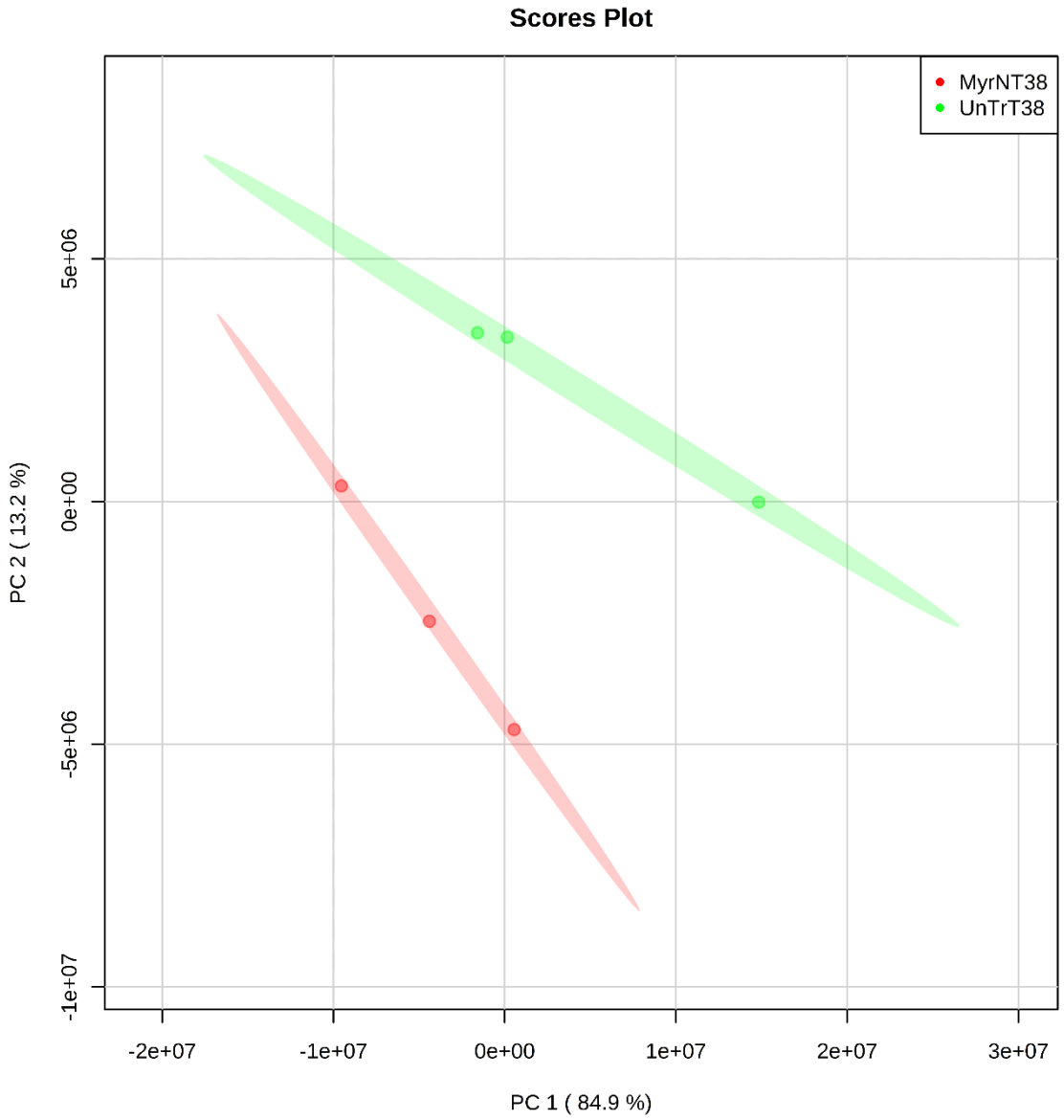


***Figure A.III-A 10*:** *PCA showing the projection of the volatile metabolic profiles of untreated control soil (green) and herbicide spiked soil (red) samples at day 38 after spiking.*
Plot generated using MetaboAnalyst.

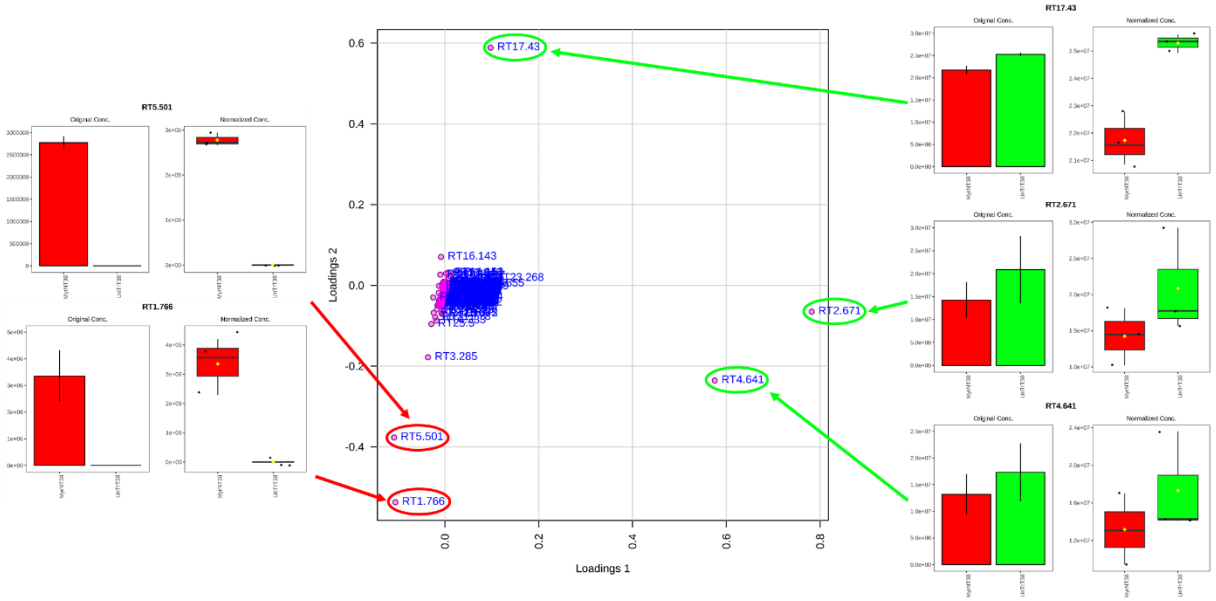


Figure A.III-A 11: *Loading plot of the PCA at day 38 after spiking (Figure A.III-A 10) showing the most relevant features constituting the two principal components PC1 and PC2.*
The most relevant 2 features of the PC1, i.e. RT2.671 and RT4.641, respectively, are showing a high contribution in variation despite their high "intra-group" variation.
For the PC2, RT17.43 is the most relevant feature with the highest score of variation. It was identified as a bleeding silicon derivate. However, the second 2 most relevant features, i.e. RT1.766 and RT5.501 respectively, were previously detected and considered as herbicide xenometabolites.
Green color refers to untreated control samples. Red color refers to spiked samples.
Plots generated using MetaboAnalyst.

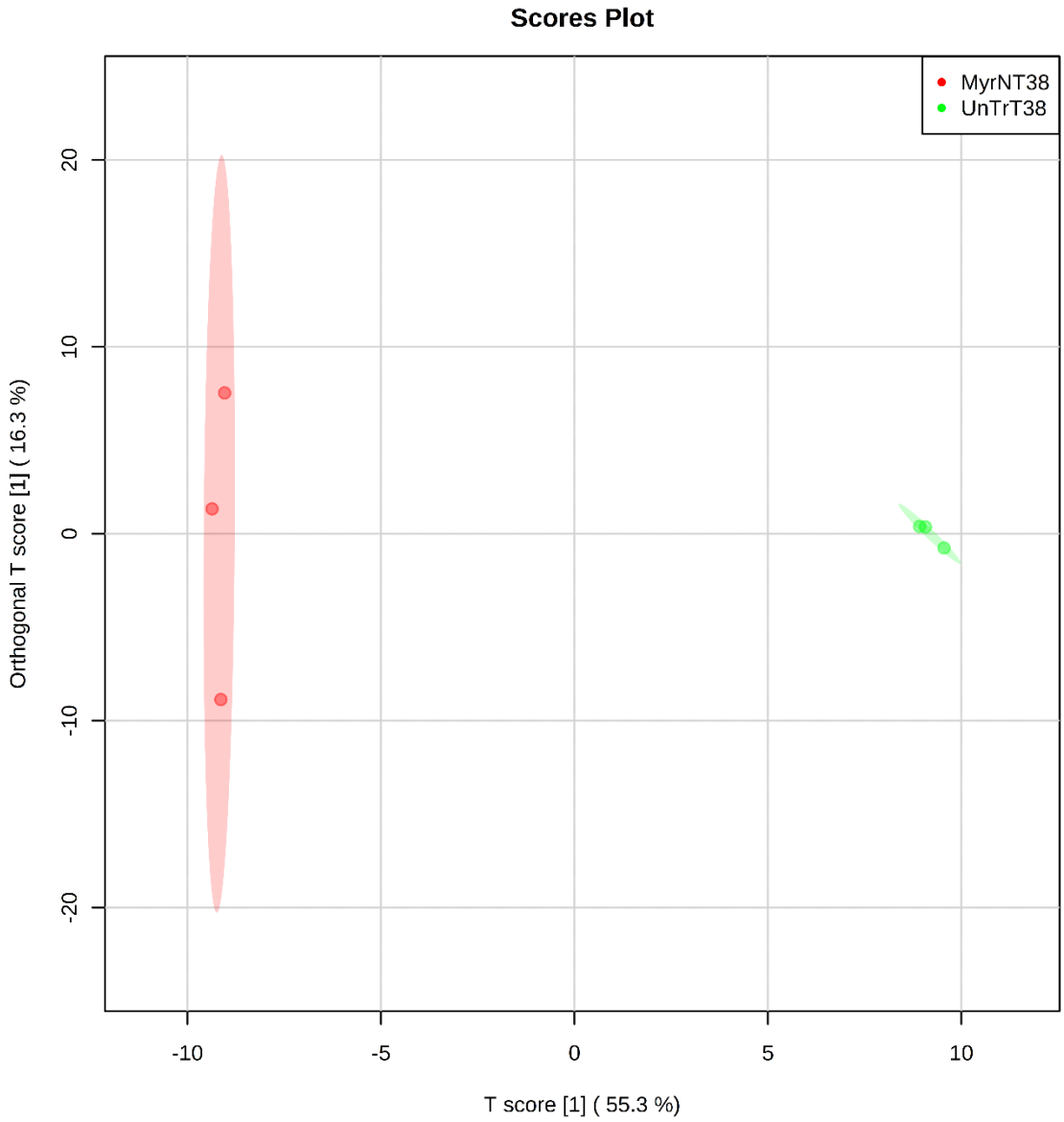


Figure A.III-A 12**:** *OPLS-DA plot, showing the discrimination between the 2 defined groups of samples: untreated control soil (green), and herbicide-spiked soil (red) at day 38 after spiking.*
Plot generated using MetaboAnalyst.

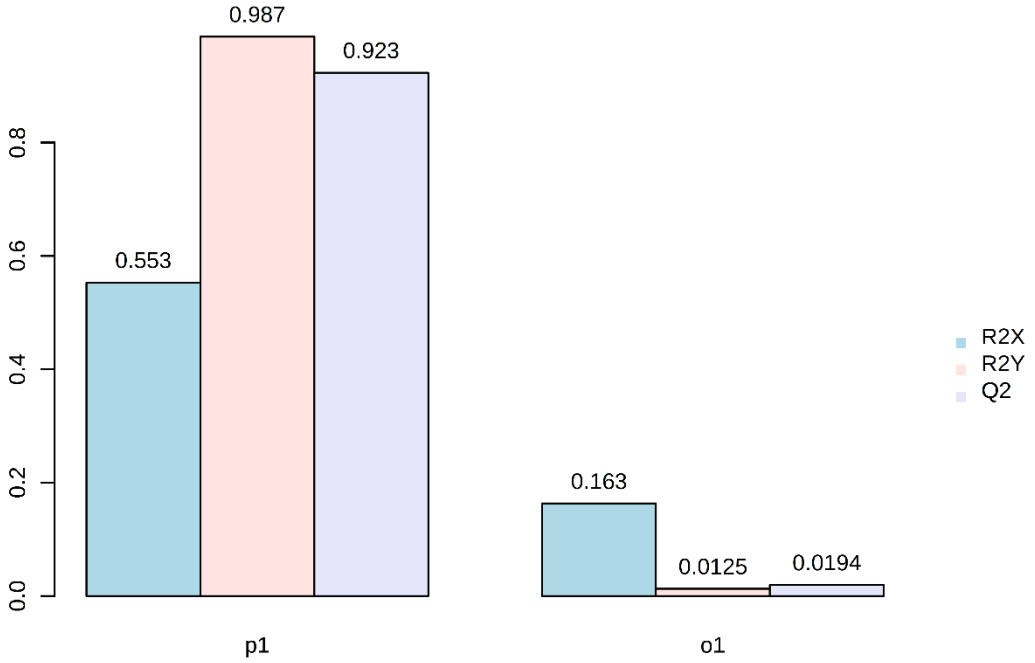


***Figure A.III-A 13*:** *Results of the OPLS-DA model Cross-Validation (CV) test (at day 38 after spiking).*
*The R2X represents the percentage of variations that are explained by the components of the OPLS-DA. i.e., the p1 R2X is the T score [1], and the o1 R2X is the Orthogonal T score [1] (**Figure A.III-A 12**). The R2Y represents the correlation of the two groups of samples to the explained variation. The Q2 value assesses the prediction of the components.*
For p1, if R2Y and Q2 are above 90 %, with R2Y higher than R2X and Q2, the model prediction is considered as reliable with a high correlation between the discriminated samples. Thus, the discrimination is significant. For the o1, however, a good prediction and a relatively high correlation between samples (R2Y and Q2 > 50 %) means that a systematic "groups-independent" discrimination has occurred due to a causation factor (e.g. analytical/instrumental drift, sampling error). This leads to decrease the confidence in significance of discrimination between the defined groups, as this discrimination can be considered caused or influenced by the systematic error [1–3], which is not the case for the results presented above (R2Y = 1.25 %, Q2 = 1.94 %).
Plot generated using MetaboAnalyst.

Table A.III-A 6: *Summary of results for all prioritized features.*

†: The given-code represents the retention time of the compound (in minutes) preceded by the Retention Time "RT" abbreviation.
‡: If MF ≥ 700, ***and*** *Δ between experimental and NIST RI ≤ 10, the considered level of identification confidence is* ***2****. If MF < 700* ***or*** *Δ between experimental and NIST RI > 10, the considered level of identification confidence is* ***3*** *(levels defined by Sumner et al. 2007 [4]).*
⸸: The percentage of the "sum of major fragments EICs area/total TIC area" ratio, calculated at day 1 after spiking.
N/A: Not Available. N/C: Not Calculated. The relative intensity was not calculated for degradation by-products that were not detected at day 1.
References: Popovici et al. 2008 [5], Svoboda et al. 1998 [6], and Carlton et al. 1992 [7].

Compound given-code†	**Putative identity‡** **(level 2 or 3 of identification confidence)**	**CAS#** **(*: PubChem CID if CAS is N/A)**	**MF**	**RI** **(Experimental)**	**RI** **(NIST)**	**Relative intensity (%)⸸**	**Reference**
Myrica gale methanolic extract components							
RT1.282	Unknown	N/A	N/A	N/C	N/A	0.01	N/A
RT1.683	Benzenehexanenitrile, β,β-dimethyl-ε-oxo-	62623-62-5	817	N/C	N/A	0.04	N/A
RT1.766	Methyl benzyl sulfoxide	824-86-2	801	N/C	N/A	0.21	N/A
RT5.043	psi-Cumene	95-63-6	853	995	990 ± 6	0.04	N/A
RT5.207	α-Phellandrene	99-83-2	892	1006	1005 ± 2	0.58	[5,7]
RT5.492	p-Cymene	99-87-6	881	1026	1025 ± 2	1.56	[5–7]
RT5.501	o-Cymene	527-84-4	902	1029	1022 ± 2	1.46	N/A
RT5.599	Eucalyptol	470-82-6	910	1036	1032 ± 2	11.09	[5–7]
RT6.035	γ-Terpinene	99-85-4	863	1059	1060 ± 3	0.08	[6,7]
RT6.624	Fenchone	1195-79-5	918	1089	1096 ± N/A	0.85	N/A
RT6.765	Linalool	78-70-6	887	1096	1099 ± 2	0.37	[6,7]
RT6.972	Acetic acid, methoxyphenyl-, methyl ester	56143-21-6	758	1106	N/A	0.07	N/A
RT7.358	2-p-Menthen-1-ol	619-62-5	800	1124	1126 ± N/A	0.23	[5]
RT7.490	cis-2-Norbornanol	17974-51-5	834	1130	N/A	0.06	N/A
RT7.739	trans-2-Menthenol	29803-81-4	890	1141	1140 ± 4	0.11	N/A
RT7.857	cis-4-methoxy thujane	N/A	767	1146	N/A	0.75	N/A
RT8.152	Ether, p-menth-6-en-2-yl methyl	121209-92-5	690	1159	N/A	0.60	N/A
RT8.267	trans-4-methoxy thujane	115562-89-5	820	1164	N/A	0.57	N/A
RT8.374	δ-Terpineol	7299-42-5	873	1168	1166 ± 3	0.42	[5]
RT8.421	Borneol	464-45-9	903	1172	1166 ± 7	1.25	[7]
RT8.688	L-terpinen-4-ol	20126-76-5	934	1179	1182 ± 0	7.09	[6,7]
RT8.823	p-tert-Butylbenzyl alcohol	877-65-6	750	1184	1336 ± N/A	0.21	N/A
RT8.844	Thymol methyl ether	1076-56-8	709	1186	1235 ± 2	0.03	N/A
RT9.090	α-Terpineol	98-55-5	913	1193	1189 ± 2	7.94	[6,7]
RT9.095	5-Caranol, (1S,3R,5S,6R)-(-)-	6909-21-3	779	1196	N/A	0.01	N/A
RT9.257	trans-Dihydrocarvone	5948-04-9	802	1198	1201 ± 2	0.02	N/A
RT9.336	Levoverbenone	1196-01-6	669	1205	1204 ± N/A	0.13	N/A
RT9.483	2-Oxabicyclo[2.2.2]octan-6-one, 1,3,3-trimethyl-	107598-08-3	731	1211	1217 ± N/A	0.06	N/A
RT9.534	Anisole, 4-sec-butyl	4917-90-2	746	1213	1236 ± N/A	0.04	N/A
RT9.599	4-(2-Methoxypropan-2-yl)-1-methylcyclohex-1-ene	14576-08-0	808	1215	N/A	0.14	N/A
RT9.816	Geosmin	19700-21-1	742	1223	1384 ± 0	0.20	N/A
RT10.073	Anisole, 2-isopropyl-4-methyl-	31574-44-4	886	1232	1230 ± 15	0.28	N/A
RT10.564	(S)-(-)-Citronellic acid, methyl ester	2270-60-2	875	1250	1261 ± N/A	0.04	N/A
RT10.591	Piperitone	89-81-6	712	1252	1253 ± 3	0.28	[7]
RT11.233	Methyl hydrocinnamate	103-25-3	928	1267	1279 ± 2	1.28	N/A
RT11.398	L-α-bornyl acetate	5655-61-8	885	1278	1284 ± 2	0.89	[7]
RT11.751	2-Undecanone	112-12-9	922	1285	1294 ± 2	1.60	[5]
RT13.338	α-Terpineol acetate	80-26-2	934	1342	1350 ± 3	4.75	[7]
RT16.159	γ-Elemene	29873-99-2	859	1424	1433 ± 3	0.39	[7]
RT16.195	4,4-Dimethyl-3-(3-methylbut-3-enylidene)-2-methylenebicyclo[4.1.0]heptane	79718-83-5	762	1427	N/A	0.05	N/A
RT16.862	Cedrene	11028-42-5	839	1442	1422 ± 24	0.21	N/A
RT17.170	α-Gurjunene	489-40-7	874	1454	1409 ± 2	0.37	[5]
RT17.570	Aromadendrene, dehydro-	589433*	784	1466	1464 ± 1	5.74	N/A
RT17.948	γ-Muurolene	30021-74-0	768	1475	1477 ± 3	0.67	[7]
RT18.195	Eremophila-1(10),11-diene	10219-75-7	861	1484	1499 ± 8	0.99	[7]
RT18.448	(+)-β-Selinene	17066-67-0	853	1488	1486 ± 3	2.49	[7]
RT18.573	α-Selinene	473-13-2	904	1492	1494 ± 3	1.70	[7]

RT18.605	α-Muurolene	31983-22-9	913	1493	1499 ± 3	0.49	[7]
RT19.175	γ-Cadinene	39029-41-9	904	1511	1513 ± 2	0.87	[7]
RT19.299	β-Cadinene	523-47-7	850	1516	1518 ± 10	3.90	[6]
RT19.447	Calamenene	483-77-2	846	1518	1523 ± 5	0.61	[6,7]
RT19.496	δ-Guaiene	3691-11-0	814	1521	1505 ± 3	0.03	N/A
RT19.821	δ-Selinene	28624-23-9	866	1530	1509	0.64	[7]
RT19.889	Unknown	N/A	N/A	1538	N/A	0.04	N/A
RT20.047	γ-Selinene	58893-88-2	907	1541	1544 ± N/A	8.30	N/A
RT20.170	3,7(11)-Selinadiene	6813-21-4	913	1547	1542 ± 3	5.96	[5]
RT20.525	Patchoulane	3724-42-3	703	1548	1552 ± N/A	0.03	N/A
RT20.823	4,4-Dimethyl-3-(3-methylbut-3-enylidene)-2-methylenebicyclo[4.1.0]heptane	79718-83-5	820	1559	N/A	0.17	N/A
RT21.001	β-Vatirenene	27840-40-0	904	1565	1563	0.91	N/A
RT21.549	α-Cedrene epoxide	29597-36-2	768	1578	1585 ± 0	0.17	N/A
RT21.717	1(10),11-Eremophiladien-9-ol	61847-19-6	797	1582	1553 ± N/A	0.11	N/A
RT21.778	Aristolene epoxide	535269*	817	1585	N/A	2.84	N/A
RT22.148	cis-β-Elemenone	32663-57-3	905	1595	1593 ± 3	1.42	[6,7]
RT22.417	β-Ionone	14901-07-6	755	1601	1491 ± 2	0.34	N/A
RT22.687	3-Cyclohexene-1-propanol, 4-methyl-γ-methylene-α-(2-methyl-1-propen-1-yl)-	38142-56-2	702	1609	1608 ± 4	0.09	N/A
RT23.095	1,4-Benzenedipropanol, α,α',γ,γ,γ',γ'-hexamethyl-	54964-98-6	710	1622	N/A	1.81	N/A
RT23.271	Epicubenol	19912-67-5	776	1624	1627±2	0.08	N/A
RT24.100	Aristol-1(10)-en-9-ol	1372763-27-3	742	1651	1692 ± 12	0.50	N/A
RT24.248	Pogostole	21698-41-9	731	1655	1655 ± N/A	0.24	N/A
RT24.428	10-Isopropenyl-3,7-cyclodecadien-1-one	55521-11-4	821	1660	N/A	0.17	N/A
RT24.642	epi-γ-Eudesmol	117066-77-0	842	1665	1662 ± 2	0.11	N/A
RT24.854	Cadalene	483-78-3	748	1669	1674±3	0.17	N/A
RT25.500	Germacrone	6902-91-6	930	1691	1693 ± 3	5.81	[5–7]
RT25.513	Juniper camphor	473-04-1	924	1694	1692 ± 8	0.24	[7]
RT28.197	1-[1-Methoxy-3,3-dimethyl-2-(3-methylbuta-1,3-dienyl)cyclopentyl]ethanone	606240*	735	1771	N/A	0.07	N/A
RT28.907	α-Phellandrene, dimer	7350-11-0	906	1789	1801 ± N/A	0.29	N/A
RT29.140	Unknown	N/A	N/A	1797	N/A	0.13	N/A
RT29.488	cis-Valerenyl acetate	101527-78-0	767	1807	1817 ± 12	0.03	N/A
RT29.523	Isovalencenyl formate	352457-47-7	719	1808	1800 ± N/A	0.02	N/A
Degradation by-products							
RT2.425	Methyl isovalerate	556-24-1	831	777	773 ± 5	0.04	N/A
RT3.050	Tyranton	123-42-2	834	843	838 ± 8	<0.01	N/A
RT4.259	Butanoic acid, 2,2-dimethyl-3-oxo-, methyl ester	38923-57-8	800	946	936	N/C	N/A
RT4.333	Methyl 2-methylhexanoate	2177-81-3	770	952	953 ± 2	N/C	N/A
RT4.441	Camphene	5794-04-7	932	957	952 ± 2	0.02	[6,7]
RT4.459	β-Pinene	127-91-3	678	960	979 ± 2	<0.01	[6,7]
RT4.953	2,3-Dehydro-1,8-cineole	92760-25-3	819	991	991 ± 2	<0.01	N/A
RT6.690	Methyl 2-propylheptanoate	56247-53-1	720	1096	1155 ± N/A	N/C	N/A
RT6.885	3-Acetyl-2,5-dimethylfuran	10599-70-9	590	1101	1099 ± 4	0.01	N/A
RT7.536	Methyl octanoate	111-11-5	584	1133	1126 ± 2	N/C	N/A
RT7.837	Unknown	N/A	N/A	1146	N/A	<0.01	N/A
RT7.877	(+)-Camphor	464-49-3	932	1148	1143 ± 9	2.80	[5]
RT8.073	Camphene hydrate	465-31-6	870	1155	1148 ± 2	0.31	[5]
RT8.140	3-Isopropyl-2-methylcyclopentanone	54549-81-4	715	1157	1174 ± N/A	N/C	N/A
RT8.267_2	cis-p-Menthan-3-one	491-07-6	809	1164	1164 ± 6	0.35	N/A
RT8.677	2(3H)-Benzofuranone, hexahydro-3a,7a-dimethyl-, cis-	38110-72-4	729	1182	N/A	N/C	N/A
RT9.209	Tetrahydrocarvone	499-70-7	856	1200	1208 ± N/A	0.01	N/A
RT11.495	8,9-Dehydrothymol methyl ether	39701-08-1	733	1281	1247 ± N/A	0.02	N/A
RT12.486	5-Methoxy-4,4,6-trimethyl-7-oxabicyclo[4.1.0]heptan-2-one	567112*	645	1314	N/A	N/C	N/A
RT13.266	Unknown	N/A	N/A	1337	N/A	N/C	N/A
RT16.983	Selinan	30824-81-8	621	1450	1476 ± 12	0.01	N/A
RT20.371	3,5,11-Eudesmatriene	193615-07-5	859	1547	1495 ± N/A	1.92	N/A

References

[1] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics. 16 (2002) 119–128. https://doi.org/10.1002/cem.695.

[2] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, NY, 1966: pp. 391–420. https://ci.nii.ac.jp/naid/20001378860/en/ (accessed October 18, 2020).

[3] S. Wiklund, Multivariate data analysis for Omics, (2008). https://metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf (accessed October 18, 2020).

[4] 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. 3 (2007) 211–221. https://doi.org/10.1007/s11306-007-0082-2.

[5] J. Popovici, C. Bertrand, E. Bagnarol, M.P. Fernandez, G. Comte, Chemical composition of essential oil and headspace-solid microextracts from fruits of *Myrica gale* L. and antifungal activity, Natural Product Research. 22 (2008) 1024–1032. https://doi.org/10.1080/14786410802055568.

[6] K.P. Svoboda, A. Inglis, J. Hampson, B. Galambosi, Y. Asakawa, Biomass production, essential oil yield and composition of Myrica gale L. harvested from wild populations in Scotland and Finland, Flavour and Fragrance Journal. 13 (1998) 6. https://doi.org/10.1002/(SICI)1099-1026(199811/12)13:6<367::AID-FFJ724>3.0.CO;2-M.

[7] R.R. Carlton, P.G. Waterman, A.I. Gray, Variation of leaf gland volatile oil within a population of sweet gale (Myrica gale) (Myricaceae), Chemoecology. 3 (1992) 45–54. https://doi.org/10.1007/BF01261456.

Appendix III-B

For Chapter III

Online Headspace-Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry-based untargeted volatile metabolomics for studying emerging complex biopesticides: A proof of concept

Appendix III-B

Hikmat Ghosson[1,2,*], Delphine Raviglione[1,2], Marie-Virginie Salvia[1,2,†], Cédric Bertrand[1,2,3,†]

1: PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

2: UFR Sciences Exactes et Expérimentales, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

3: S.A.S. AkiNaO, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

*: hikmat.ghosson@univ-perp.fr / hikmatghosson@gmail.com

†: *Equal Contribution (Last Co-authors)*

Summary

Information shown below represent the kinetics tracking of the 101 detected compounds that were considered after the prioritization by time-series Heatmap, MEBA-based filtering and manual peak tracking crosscheck.

For each compound, three different plots are shown with their related data in the tables:

1. The first graph presents the kinetics profile of the compound in each single sample, where its pseudo-spectrum peak area is integrated over time.
2. The second plot presents the evolution of means and standard deviations of compound peak areas over time. These means and standard deviations are calculated in the three replicates of each group of samples.
3. The third figure consists of boxplots of compound peak areas and their evolution over time.

MEBA plots are also shown in the annex documents.

Notes:

- ✓ *Peak Areas represent the sum of EICs of compound's major ions (in its pseudo-spectrum).*
- ✓ *The time scale is reliable for the first two plots, not for the third, however.*
- ✓ *The Figures and Tables could not be included in the present manuscript due to hardcopy formatting issues. They are published as an Appendix for the article by Ghosson et al.*[25]*, and can be retrieved online via the following link:* *https://ars.els-cdn.com/content/image/1-s2.0-S0003267020308473-mmc2.pdf*. *Or on publisher's website via the Digital Object Identifier below:* *https://doi.org/10.1016/j.aca.2020.08.016*.
- ✓ *The MEBA plots can be retrieved online via the following link:* *https://hal.archives-ouvertes.fr/hal-03139690*.

[25] Ghosson, H. *et al. Anal. Chim. Acta* **(2020)**, 1134:58–74. doi:10.1016/j.aca.2020.08.016

Abstract

Despite the ecological and sanitary awareness, worldwide consumption of pesticides is increasing. As these chemical products represent several adverse effects on human health and environment, measures should be taken in order to limit their impacts. Biocontrol products are proposed as an alternative solution of the synthetic products. In fact, these "biopesticides" are presumed to be less harmful and relatively less persistent. However, this a priori must be examined and strict risk assessment of those new substances should be considered.

The development of biocontrol solutions proceeds first of all through the proposed protocols to study their activity and their environmental fate and impact. Currently, half-life (DT50) is used in order to evaluate the environmental fate of synthetic pesticides. However, DT50 approach gives only information about pesticides' persistence in the environment, but no indications concerning the formation of degradation products or its impact on biodiversity are provided. Furthermore, biocontrol products are complex (bio)chemical mixes. The DT50 is not applicable for such complex products. Therefore, novel analytical approaches should be considered in order to overcome these difficulties.

A novel approach based on meta-metabolomics and Mass Spectrometry; the "Environmental Metabolic Footprinting" (EMF), was recently introduced. It affords a novel universal and integrative proxy; the "resilience time", dedicated to assess the environmental fate and impact of complex (bio)pesticides in environmental matrices (e.g. soil, sediment). Nonetheless, the development of such Mass Spectrometry-based untargeted meta-metabolomics approach needs to be in-depth studied. Several tasks should be addressed: 1) performant extraction protocols and GC/LC-(HR)MS-based analytical methods should be set up, 2) suitable data processing and chemometric tools should be developed to deal with the complexity of the generated datasets, 3) the impact of xenometabolome complexity on MS-based analyses should be assessed, and 4) the study of the volatile residues should be considered and thus needs new analytical methodologies to be developed.

The work was carried out following 3 main axes. The first axis addressed 1) the development of extraction protocols and LC-HRMS methods to analyze both pesticides xenometabolites and soil endometabolites, and 2) the development of a novel chemometric approach to assess the extraction performance. Novel extraction protocols have been proven optimal for the EMF, and the chemometric approach was thus validated. The second axis assessed the impact of xenometabolome complexity on the determination of environmental biomarkers. Ion suppression was revealed and thus a pragmatic strategy has been developed to overcome its influence. The third axis aimed to set-up a novel methodology in order to analyze the volatile residues of complex pesticides. HS-SPME-GC-MS analyses were coupled to chemometrics in order to perform kinetics studies and to follow the transformation of the volatile residues. The chemometric workflow proved its reliability to explain pesticide's transformation and novel xenometabolites and by-products were identified.

In conclusion, significant advances were carried to the EMF. It has been consolidated for laboratory and field applications that must be investigated in order to improve the proxy and to validate it as a reliable approach for pesticides risk evaluation.

Keywords: Metabolomics; Mass Spectrometry; Chemometrics; Pesticides; Environmental Impact; Biocontrol Products

Résumé

Malgré la prise de conscience écologique et sanitaire, la consommation mondiale de pesticides est en augmentation. Étant donné que ces produits chimiques présentent de nombreux effets néfastes sur la santé humaine et l'environnement, des mesures doivent être prises afin de limiter leurs effets. Les produits de biocontrôle sont proposés comme une solution alternative aux produits synthétiques. En effet, ces « biopesticides » sont présumés être moins nocifs et relativement moins persistants. Toutefois, cet a priori doit être examiné et une évaluation stricte des risques de ces nouvelles substances doit être envisagée.

Le développement de solutions de biocontrôle passe d'abord par les protocoles proposés pour étudier leur activité, leur devenir et leur impact environnemental. Actuellement, le temps de demi-vie (t½) est utilisé pour évaluer le devenir environnemental des pesticides synthétiques. Cependant, l'approche t½ ne donne que des informations sur la persistance des pesticides dans l'environnement, mais aucune indication concernant la formation de produits de dégradation ou son impact sur la biodiversité n'est apportée. De plus, les produits de biocontrôle sont des mélanges (bio)chimiques complexes. La t½ n'est pas applicable pour ces produits complexes. Par conséquent, de nouvelles approches analytiques doivent être envisagées afin de surmonter ces difficultés.

Une nouvelle approche basée sur la méta-métabolomique et la Spectrométrie de Masse; « Empreinte Métabolique Environnementale » (EMF), a été récemment introduite. Elle offre un nouveau proxy universel et intégratif; le « temps de résilience », dédié à l'évaluation du devenir environnemental et de l'impact des (bio)pesticides complexes dans des matrices environnementales (ex. sol, sédiments). Néanmoins, le développement d'une telle approche de méta-métabolomique non ciblée basée sur la Spectrométrie de Masse doit être étudié en profondeur. Plusieurs tâches doivent alors être abordées: 1) les protocoles d'extraction performants et les méthodes analytiques basées sur la GC/LC-(HR)MS doivent être mis en place, 2) le traitement de données et les outils chimiométriques appropriés doivent être développés pour maitriser la complexité des ensembles des données générées, 3) l'impact de la complexité du xénométabolome sur les analyses basées sur la MS doit être évalué, et 4) l'étude des résidus volatiles doit être envisagée et nécessite donc le développement de nouvelles méthodologies analytiques.

Le travail a été mené sur 3 axes principaux. Le premier axe portait sur 1) le développement de protocoles d'extraction et des méthodes LC-HRMS pour analyser à la fois les xénométabolites des pesticides et les endométabolites du sol, et 2) le développement d'une nouvelle approche chimiométrique pour évaluer les performances d'extraction. De nouveaux protocoles d'extraction se sont avérés optimaux pour l'EMF, et l'approche chimiométrique a donc été validée. Le deuxième axe a évalué l'impact de la complexité du xénométabolome sur la détermination des biomarqueurs environnementaux. La suppression d'ion a été révélée et une stratégie pragmatique a donc été élaborée pour surmonter son influence. Le troisième axe visait à mettre en place une nouvelle méthodologie pour analyser les résidus volatils de pesticides complexes. Des analyses HS-SPME-GC-MS ont été couplées à la chimiométrie afin de réaliser des études cinétiques et de suivre la transformation des résidus volatils. Le workflow chimiométrique a prouvé sa fiabilité pour expliquer la transformation du pesticide et de nouveaux xénométabolites et sous-produits ont été identifiés.

En conclusion, une avancée significative a été apportée à l'EMF. Elle a été consolidée pour les applications en laboratoire et sur le terrain qui doivent être étudiées afin d'améliorer le proxy et de le valider comme une approche fiable pour l'évaluation des risques des pesticides.

Mots-clés : Métabolomique ; Spectrométrie de Masse ; Chimiométrie ; Pesticides ; Impact Environnemental ; Produits de Biocontrôle