

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'applications et besoins méthodologiques

C. Soulier, N. Huynh, B. Bonnaud, A. Assoumani

Septembre 2024

Document final

En partenariat avec



Avec le soutien de



Contexte de programmation et de réalisation

Ce rapport a été réalisé dans le cadre du programme scientifique et technique AQUAREF pour l'année 2023, au titre de l'action « FG1.2c » « Veille NTS empreintes pour la caractérisation de sources ».

Auteurs :

Coralie Soulier
BRGM
c.soulier@brgm.fr

Nina Huynh
INERIS
Nina.HUYNH@ineris.fr

Bertille Bonnaud
LNE
bertille.bonnaud@lne.fr

Azziz Assoumani
INERIS
Azziz.ASSOUMANI@ineris.fr

Vérification du document :

Cécile Miège
INRAE
cecile.miege@inrae.fr

Les correspondants

OFB : Pierre-François Staub, pierre-francois.staub@ofb.gouv.fr

BRGM : Jean-Philippe Ghestem, jp.ghestim@brgm.fr

Référence du document : C. Soulier, N. Huynh, B. Bonnaud et A. Assoumani- Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'applications et besoins méthodologiques - Rapport AQUAREF 2024 - 47 p.

Droits d'usage :	<i>Accès libre</i>
Couverture géographique :	<i>International</i>
Niveau géographique :	<i>National</i>
Niveau de lecture :	<i>Professionnels, experts</i>
Nature de la ressource :	<i>Document</i>

1	Introduction	17
2	Contexte.....	18
2.1	Définitions	18
2.2	traitement des empreintes HRMS.....	19
3	Quels sont les outils permettant le traitement des empreintes HRMS ?.....	27
3.1	Les outils commerciaux.....	28
3.2	Les outils en libre accès	29
3.3	Avantages / inconvénients.....	34
4	Assurance qualité et contrôle qualité (QA/QC)	34
4.1	Les prérequis analytiques.....	34
4.1.3.1	<i>Blancs</i>	35
4.1.3.2	<i>Seuils d'intensité</i>	35
4.1.3.3	<i>Autres</i>	36
4.1.4.1	<i>Choix des étalons internes</i>	36
4.1.4.2	<i>Utilisation des données des étalons internes</i>	36
4.2	Les formats de données	37
5	Conclusion	38

RESUME

L'analyse non-ciblée par chromatographie couplée à la spectrométrie de masse haute résolution (HRMS) permet l'acquisition d'une grande partie de l'information présente dans un extrait ou un échantillon : la masse d'une molécule associée au temps de rétention chromatographique. L'ensemble de ces couples (masse- temps de rétention) forme ce qu'on appelle une empreinte HRMS. Ces empreintes peuvent être utilisées pour faire de la recherche suspectée dans le but d'identifier des molécules par comparaison avec des bases de données et de la recherche non-ciblée afin de mettre en évidence certains couples par des traitements statistiques.

Parmi les applications d'analyse non-ciblée, l'utilisation des empreintes globales, sans identification préalable des molécules, pour la caractérisation de sources de pollution se développe, parallèlement aux avancées des traitements statistiques de grands volumes de données « big data ». Cette étude a pour objectifs de faire un état des lieux des études par analyse non-ciblée des empreintes HRMS sur des échantillons environnementaux, un inventaire des logiciels fournisseurs et outils en libre accès, et de recenser les contrôles qualités à associer aux analyses pour obtenir des résultats fiables.

Les études impliquant l'analyse non-ciblée via les empreintes HRMS permettent d'obtenir des informations sur les évolutions spatiales, les évolutions temporelles et l'identification de sources de pollution.

Les outils de prétraitement, traitement et visualisation des empreintes HRMS, sont généralement des outils en libre accès. L'opérateur a un large éventail d'outils à sa disposition pour traiter les empreintes HRMS ; son choix doit être guidé par la compatibilité des formats de données entre les données acquises et le format d'entrée dans l'outil, du mode de langage utilisé (R, Python, etc.), du type de traitement des données souhaité (objectifs, prétraitement, visualisation, etc.) et aussi de l'accessibilité des outils en termes de prise en main pour les différents réglages et algorithmes à utiliser.

Les prérequis analytiques à appliquer sont impératifs pour obtenir des résultats fiables. Pour l'analyse non-ciblée, les différents guides relatifs aux critères d'assurance qualité recommandent pour la série d'analyse :

- Une randomisation des échantillons dans la séquence d'analyse pour répartir aléatoirement le biais analytique ;
- L'analyse de plusieurs réplicats d'un même échantillon pour minimiser les taux de faux positifs et faux négatifs ; le nombre de 2 à 3 réplicats est recommandé.
- La réalisation de blancs d'analyse à mesurer en début et fin de série, pour prendre en compte d'éventuelles contaminations au laboratoire ;
- La mise en œuvre de blancs terrain qui peuvent permettre de prendre en compte des contaminations pendant l'échantillonnage ;
- L'emploi d'étalons internes pour maîtriser et corriger la fluctuation des signaux (dus par exemple à l'encrassement de la source de l'équipement). Une réflexion sur l'emploi d'une liste minimale d'étalons internes communs est en cours à différents niveaux : européens mais aussi entre laboratoires qui harmonisent leurs travaux (Aquaref, GE SANA de la CIPR).
- L'injection d'échantillons de contrôle pour vérifier l'absence de biais, en vérifiant les intensités des pics obtenus.

Par rapport au traitement du signal, on peut souligner le besoin de disposer :

- d'un critère pour le traitement des blancs ; la suppression dans les échantillons, des signaux retrouvés dans les blancs, se fait avec un seuil différent selon les études.
- d'un critère pour ne prendre en compte que les pics d'une intensité supérieure à un seuil à définir.
- d'une méthode de prise en compte de la réponse des étalons internes en vue de normaliser la réponse des molécules. Celle-ci n'est pas encore aboutie à ce jour.

Au vu des prérequis analytiques décrits dans ce rapport et bien que de nombreux contrôles qualité aient été mis en place, les empreintes HRMS acquises dans les études du réseau de surveillance prospective

[1], [2] ne permettraient pas d'obtenir des résultats fiables. En effet les analyses n'ont pas été réalisées en réplicats ce qui permettrait de minimiser le taux de faux positifs et de faux négatifs.

Mots clés (thématique et géographique) :

Analyse non-ciblée, empreintes HRMS, contrôles qualité, sources de pollution

USE OF HRMS (HIGH-RESOLUTION MASS SPECTROMETRY) FINGERPRINTS: EXAMPLES OF APPLICATIONS AND METHODOLOGICAL REQUIREMENTS

C. Soulier, N. Huynh, B. Bonnaud and A. Assoumani

ABSTRACT

Non-targeted analysis using chromatography coupled with high-resolution mass spectrometry (HRMS) makes it possible to acquire a large proportion of the information present in an extract or sample: the mass of a molecule combined with the chromatographic retention time. These couples (mass - retention time) form what is known as an HRMS fingerprint. These fingerprints can be used to carry out suspected research in order to identify molecules by comparison with databases, and non-targeted research in order to highlight certain couples by statistical analysis.

Among the applications of non-targeted analysis, the use of global fingerprints, without prior identification of the molecules, to characterise sources of pollution is developing, in parallel with advances in the statistical processing of large volumes of 'big data'.

This study includes a review of studies using non-targeted analysis on environmental samples, an inventory of software suppliers and open-access tools, and identification of the quality controls that need to be associated with analyses in order to obtain reliable results.

Studies involving the non-targeted analysis of HRMS provide information on spatial and temporal changes and the identification of pollution sources.

The data pre-processing, processing and visualisation tools used for the non-targeted analysis of HRMS fingerprints are generally freely available. The operator has a wide range of tools at his disposal for processing HRMS fingerprints; his choice should be guided by the compatibility of the data formats, the language mode used (R, Python, etc.), the type of data processing desired (targeting, pre-processing, visualisation, etc.) and also the accessibility of the tools in terms of getting to grips with the various settings and algorithms to be used.

The analytical prerequisites to be applied are imperative for obtaining reliable results. For non-targeted analysis, the various guides relating to quality assurance criteria recommend the following for the analysis series:

- Randomisation of samples in the analysis sequence to distribute the analytical bias randomly;
- The analysis of several replicas of the same sample to minimise the false positive and false negative rates; the number of 2 to 3 replicas is mentioned.
- Analysis blanks to be measured at the beginning and end of the series, to take account of any contamination in the laboratory;
- The use of field blanks to take account of contamination during sampling;
- The use of internal standards to control and correct signal fluctuation (due, for example, to soiling of the equipment source). Consideration is being given to the use of a minimum list of common internal standards at various levels: European, but also between laboratories harmonising their work (Aquaref, ICRP's GE SANA).
- The injection of control samples to check the absence of bias, by verifying the intensities of the peaks obtained.

In terms of signal processing, the need for :

- a criterion for processing blanks; signals found in blanks are removed from samples at different thresholds in different studies.
- a criterion for taking into account only peaks with an intensity above a threshold to be defined.
- a method for taking into account the response of internal standards, which has not yet been finalised.

In view of the analytical prerequisites described in this report, and although numerous quality controls have been put in place, the HRMS fingerprints acquired in the studies of the prospective surveillance

network [1], [2] would not allow reliable results to be obtained. This is because the analyses were not carried out in replicates in order to minimise the rate of false positives and false negatives.

Key words (thematic and geographical area):

Non-targeted analysis, HRMS fingerprinting, quality control



RÉPUBLIQUE
FRANÇAISE

Liberté
Égalité
Fraternité



Géosciences pour une Terre durable

brgm

Document à accès immédiat

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'applications et besoins méthodologiques [Titre]

Rapport final [provisoire / final]

BRGM/RP-73614-FR

Version 0 du 20 septembre 2024

Étude réalisée dans le cadre des projets d'appui aux politiques publiques

Soulier C., Huynh N., Bonnaud B. and Assoumani A. [Auteur]

Vérificateur :

Nom : Ghestem Jean Philippe

Fonction :

Date : 10/06/24.

Signature :

Approbateur :

Nom : Amalric Laurence

Fonction :

Date : 20/09/24.

Signature :

Le système de management de la qualité et de l'environnement du BRGM est certifié selon les normes ISO 9001 et ISO 14001.

Contact : qualite@brgm.fr



Avertissement

Ce rapport est adressé en communication exclusive au demandeur, au nombre d'exemplaires prévu.

Le demandeur assure lui-même la diffusion des exemplaires de ce tirage initial.

La communicabilité et la réutilisation de ce rapport sont régies selon la réglementation en vigueur et/ou les termes de la convention.

Le BRGM ne saurait être tenu comme responsable de la divulgation du contenu de ce rapport à un tiers qui ne soit pas de son fait et des éventuelles conséquences pouvant en résulter.

Votre avis nous intéresse

Dans le cadre de notre démarche qualité et de l'amélioration continue de nos pratiques, nous souhaitons mesurer l'efficacité de réalisation de nos travaux.

Aussi, nous vous remercions de bien vouloir nous donner votre avis sur le présent rapport en complétant le formulaire accessible par cette adresse <https://forms.office.com/r/yMgFcU6Ctg> ou par ce code :



Mots clés : Analyse non-ciblée, empreintes HRMS, contrôles qualité, sources de pollution

En bibliographie, ce rapport sera cité de la façon suivante :

Soulier C., Huynh N., Bonnaud B. and Assoumani A.[Auteur] 2024. Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'application et besoins méthodologiques.[Titre]Rapport provisoire[provisoire / final] V0. BRGM/RP-73614-FR, 47 p.

© BRGM, 2024, ce document ne peut être reproduit en totalité ou en partie sans l'autorisation expresse du BRGM.
IM003-MT008-P2-09/03/2023

Synthèse

L'analyse non-ciblée par chromatographie couplée à la spectrométrie de masse haute résolution (HRMS) permet l'acquisition d'une grande partie de l'information présente dans un extrait ou un échantillon : la masse d'une molécule associée au temps de rétention chromatographique. L'ensemble de ces couples (masse- temps de rétention) forme ce qu'on appelle une empreinte HRMS. Ces empreintes peuvent être utilisées pour faire de la recherche suspectée dans le but d'identifier des molécules par comparaison avec des bases de données et de la recherche non-ciblée afin de mettre en évidence certains couples par des traitements statistiques. Ce travail est un rapport bibliographique permettant de décrire quelques exemples d'application de traitement non-ciblé à partir des empreintes HRMS, de recenser les outils de traitement fournisseurs ou en libre accès existants et de synthétiser les prérequis analytiques nécessaires pour ce type de traitement.

Parmi les applications d'analyse non-ciblée, l'utilisation des empreintes globales, sans identification préalable des molécules, pour la caractérisation de sources de pollution se développe, parallèlement aux avancées des traitements statistiques de grands volumes de données « big data ». Cette étude a pour objectifs de faire un état des lieux des études par analyse non-ciblée des empreintes HRMS sur des échantillons environnementaux, un inventaire des logiciels fournisseurs et outils en libre accès, et de recenser les contrôles qualités à associer aux analyses pour obtenir des résultats fiables.

Les études impliquant l'analyse non-ciblée via les empreintes HRMS permettent d'obtenir des informations sur les évolutions spatiales, les évolutions temporelles et l'identification de sources de pollution.

Les outils de prétraitement, traitement et visualisation des empreintes HRMS, sont généralement des outils en libre accès. L'opérateur a un large éventail d'outils à sa disposition pour traiter les empreintes HRMS ; son choix doit être guidé par la compatibilité des formats de données entre les données acquises et le format d'entrée dans l'outil, du mode de langage utilisé (R, Python, etc.), du type de traitement des données souhaité (objectifs, prétraitement, visualisation, etc.) et aussi de l'accessibilité des outils en termes de prise en main pour les différents réglages et algorithmes à utiliser.

Les prérequis analytiques à appliquer sont impératifs pour obtenir des résultats fiables. Pour l'analyse non-ciblée, les différents guides relatifs aux critères d'assurance qualité recommandent pour la série d'analyse :

- Une randomisation des échantillons dans la séquence d'analyse pour répartir aléatoirement le biais analytique ;
- L'analyse de plusieurs réplicats d'un même échantillon pour minimiser les taux de faux positifs et faux négatifs ; le nombre de 2 à 3 réplicats est recommandé.
- La réalisation de blancs d'analyse à mesurer en début et fin de série, pour prendre en compte d'éventuelles contaminations au laboratoire ;
- La mise en œuvre de blancs terrain qui peuvent permettre de prendre en compte des contaminations pendant l'échantillonnage ;
- L'emploi d'étalons internes pour maîtriser et corriger la fluctuation des signaux (dus par exemple à l'encrassement de la source de l'équipement). Une réflexion sur l'emploi d'une liste minimale d'étalons internes communs est en cours à différents niveaux : européens mais aussi entre laboratoires qui harmonisent leurs travaux (Aquaref, GE SANA de la CIPR).

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

- L'injection d'échantillons de contrôle pour vérifier l'absence de biais, en vérifiant les intensités des pics obtenus.

Par rapport au traitement du signal, on peut souligner le besoin de disposer :

- d'un critère pour le traitement des blancs ; la suppression dans les échantillons, des signaux retrouvés dans les blancs, se fait avec un seuil différent selon les études.
- d'un critère pour ne prendre en compte que les pics d'une intensité supérieure à un seuil à définir.
- d'une méthode de prise en compte de la réponse des étalons internes en vue de normaliser la réponse des molécules. Celle-ci n'est pas encore aboutie à ce jour.

Au vu des prérequis analytiques décrits dans ce rapport et bien que de nombreux contrôles qualité aient été mis en place, les empreintes HRMS acquises dans les études du réseau de surveillance prospective [1], [2] ne permettraient pas d'obtenir des résultats fiables. En effet les analyses n'ont pas été réalisées en réplicats ce qui aurait permis de minimiser le taux de faux positifs et de faux négatifs.

Sommaire

1	Introduction	17
2	Contexte.....	18
2.1	Définitions	18
2.2	traitement des empreintes HRMS.....	19
2.2.1	<i>Evolutions spatiales</i>	19
2.2.2	<i>Evolutions temporelles</i>	22
2.2.3	<i>Discrimination de sources</i>	23
3	Quels sont les outils permettant le traitement des empreintes HRMS ?.....	27
3.1	Les outils commerciaux.....	28
3.1.1	<i>Waters</i>	28
3.1.2	<i>Agilent</i>	28
3.1.3	<i>Thermo Scientific</i>	29
3.1.4	<i>Sciex</i>	29
3.2	Les outils en libre accès	29
3.3	Avantages / inconvénients.....	34
4	Assurance qualité et contrôle qualité (QA/QC)	34
4.1	Les prérequis analytiques	34
4.1.1	<i>Randomisation de la séquence d'analyse</i>	34
4.1.2	<i>Réplicats d'analyse</i>	35
4.1.3	<i>Tri du signal</i>	35
4.1.4	<i>Emploi d'étalons internes</i>	36
4.1.5	<i>Analyses d'échantillons de contrôle</i>	37
4.2	Les formats de données.....	37
5	Conclusion	38

Liste des figures

Figure 1 : Illustration de l'empreinte HRMS d'un échantillon contenant des composés potentiels d'investigation basés sur le poids moléculaire (MW ; axe x) et la polarité (logKow ; axe y) de 699013 entrées de composés organiques (masse < 1200 Da) [11]. 18

Figure 2 : Modèles obtenus pour les 3 types de pollution identifiés le long de la Bode (eaux usées (WW), rivière Bode (BR) et pollution diffuse et aléatoire (DRI)) grâce à l'exploitation des analyses LC-HRMS par traitement non ciblé. Le fond gris représente l'évolution longitudinale de l'intensité des pics individuels détectés sur les sites d'échantillonnage (l'axe x représente la distance du site depuis le point amont de la rivière, en kilomètre). Les pourcentages correspondent au pourcentage de pics attribués à la source de pollution [17]. 20

Figure 3 : Comparaison des sites étudiés pour le nombre d'entités au-dessus des scores de rareté (RS) de 1000 et 5000 [18]. 21

Figure 4 : Distribution spatiale du DEET dans des échantillons d'eau de mer provenant de la mer Noire. Le profil révèle qu'il y a un apport terrestre de DEET dans la mer Noire. Le DEET a été détecté dans 39 (sur 54) échantillons d'eau analysés par LC-HRMS et en utilisant l'analyse non-ciblée. Les cercles indiquent les stations d'échantillonnage ; les petits cercles bleu clair

correspondent à des intensités faibles de DEET ; les grands cercles rouges indiquent des intensités élevées de DEET [19]. 22

Figure 5 : A : Sites d'échantillonnage le long du Rhin avec des résultats positifs (marques vertes) et négatifs (croix rouges) de la substance N-(chlorométhyl)-triéthylammonium cation. B : Sur la base de ces résultats, la source a été identifiée et l'émission stoppée..[12] 23

Figure 6 : Des analyses de données multivariées et des statistiques univariées ont été utilisées pour mettre en évidence les entités liées à l'origine géographique de grains de maïs. La classification utilisant seulement 20 entités sélectionnées avec l'algorithme Random Forest a permis de classer correctement 90,5 % des échantillons avec une validation croisée 10 fois répétée 100 fois. Les scores de l'ACP ont été créés en utilisant les aires des pics de ces 20 entités, après le prétraitement par la méthode de l'échelle de Pareto. 24

Figure 7 : Les résultats de l'ACP sont affichés à l'aide d'un biplot des scores et loadings. Un regroupement d'entités en fonction des types de bois mesurés et une séparation entre le bois tendre et le bois dur sont observés sur le premier axe . Une discrimination plus poussée des échantillons issus de la combustion de bois tendre est observée sur le deuxième axe. Les entités ont été attribuées au type de bois en fonction de leur emplacement dans le biplot [29]. Beech = hêtre, birch = bouleau et spruce = épicéa 25

Figure 8 : Application du NTS pour l'étude du comportement de composés chimiques en station d'épuration ; exemples des dix tendances définies pour décrire les profils d'intensité dans un cluster. Chaque tendance peut être associée au comportement spécifique d'un composé dans la station d'épuration et est décrite dans le titre de chaque tendance. Les profils d'intensité pour chaque post-traitement ont été regroupés séparément et sont donc généralement intitulés "Après post-traitement". CAS désigne le traitement conventionnel par boues activées [31]..... 26

Figure 9 : Analyse en composantes principales (a) scores plot et (b) loading plot de l'effluent de la station d'épuration avant et après ozonation pour une dose d'ozone spécifique de 0,5 gO₃/gDOC mesurée en ESI positive. Les première et deuxième composantes principales sont indiquées sur les axes x et y, respectivement. Une bonne séparation entre les échantillons avant ozonation (triangles rouges) et après ozonation (cercles bleus) a été obtenue le long de l'axe des x et a pu expliquer 47 % de la variance, ce qui a permis de classer les entités non ciblées comme des composés parents potentiels ayant subi une transformation ou des produits de transformation d'ozonation potentiels (ou ni l'un ni l'autre, cercles noirs). 27

Liste des tableaux

Tableau 1 : Liste non exhaustive des outils en libre accès disponibles pour traiter les empreintes HRMS 31

1 Introduction

Ce rapport fait suite aux travaux menés

- d'une part dans les programmes Aquaref 2016-2020, portant :
 - sur le positionnement d'Aquaref sur l'implémentation de l'analyse non-ciblée dans le champ de la surveillance environnementale de la qualité des eaux [1],
 - la description des outils et méthodologie [4],
 - les initiatives européennes existantes [5],
 - les recommandations en termes de contrôle qualité et d'informations à fournir lors d'une analyse non-ciblée avec retraitement en mode suspecté [6],
- et d'autre part dans le réseau national de surveillance prospective (RSP), portant sur l'applicabilité de l'analyse non-ciblée avec traitement en mode suspecté pour la surveillance prospective sur une campagne à large échelle [1].

Ces différentes études ont permis d'aborder l'analyse non-ciblée via les traitements en mode suspecté [6-9]. Ce type de traitement ne prend en compte qu'une petite partie du signal d'un échantillon. Pour prendre en compte l'intégralité du signal c'est à dire l'empreinte HRMS (*High Resolution Mass Spectrometry*), l'analyse non-ciblée est utilisée. Au sein des programmes Aquaref, ce sujet n'a été que partiellement abordé à cette date via un rapport sur des échantillons sédimentaires et biologiques [10] et un rapport sur des rejets de stations d'épuration [7].

L'exploration des empreintes HRMS via l'analyse non-ciblée nécessite un savoir-faire en traitement statistique des données afin d'adapter les outils disponibles pour le traitement de données en fonction des besoins scientifiques. L'évolution de ces outils est constante, notamment avec le développement de nombreux outils libres d'accès et leur partage sur des plateformes collaboratives. Cependant, la question de la comparabilité des données acquises se pose, d'autant plus que les formats de données brutes ne sont pas toujours compatibles avec tous les outils. De récents travaux montrent la volonté d'essayer d'harmoniser les pratiques en mettant en place des exigences minimales en termes d'acquisition des données pour garantir une interprétation fiable de ces données.

Ce rapport a pour objectifs de faire

- un état des lieux des études et résultats obtenus par analyse non-ciblée des empreintes HRMS sur des échantillons environnementaux,
- un inventaire des logiciels fournisseurs et outils en libre accès permettant de faire ce type d'analyse,
- et de recenser les contrôles qualités à associer aux analyses pour obtenir des résultats fiables.

Il est important de préciser que ce rapport n'entrera pas dans les détails des étapes précédant l'analyse et le traitement de données (e.g., plan d'échantillonnage, stratégie de prélèvement, conservation et préparation des échantillons...). Ces étapes ont cependant une importance capitale et doivent être réfléchies en amont en fonction des objectifs fixés par l'étude.

2 Contexte

2.1 DEFINITIONS

L'analyse non ciblée (NTA, *non-target analysis*) est souvent connue pour sa capacité à permettre l'identification de molécules, parfois nouvelles, sans l'absolue nécessité de disposer d'étalons analytiques. Cette particularité, par rapport aux analyses ciblées classiques, s'explique par l'utilisation de la spectrométrie de masse haute résolution (HRMS) qui permet l'acquisition de masses très précises, sans *a priori* sur le contenu des échantillons. Cette acquisition génère une empreinte HRMS composée des signaux souvent appelés entités, ou encore *features* ou marqueurs (i.e., à ce stade, souvent désignés par un couple masse sur charge (m/z) / temps de rétention), et qui correspondent potentiellement aux molécules présentes dans l'échantillon. Deux voies de traitement de l'empreinte HRMS sont possibles :

- 1) la voie « en mode suspecté » qui permet la comparaison de paramètres d'identification, tels que la masse exacte, le temps de rétention, le profil isotopique et les masses des fragments, à ceux déjà présents dans une source « référente » (bibliothèque ou base de données) ;
- 2) et la voie « non-ciblée » qui permet l'exploitation des entités afin de mettre en évidence les plus pertinentes au regard du questionnement scientifique visé.

Pour chaque échantillon, cela représente donc un nombre important de données (Figure 1), qui ne sont que partiellement exploitées lors de comparaison avec une base de données en mode suspecté. La voie de traitement non-ciblé des empreintes HRMS permet d'observer l'évolution globale de toutes les entités détectées dans une série d'échantillons afin de pouvoir dégager des tendances spatiales ou temporelles ou encore caractériser d'éventuelles sources de pollution.

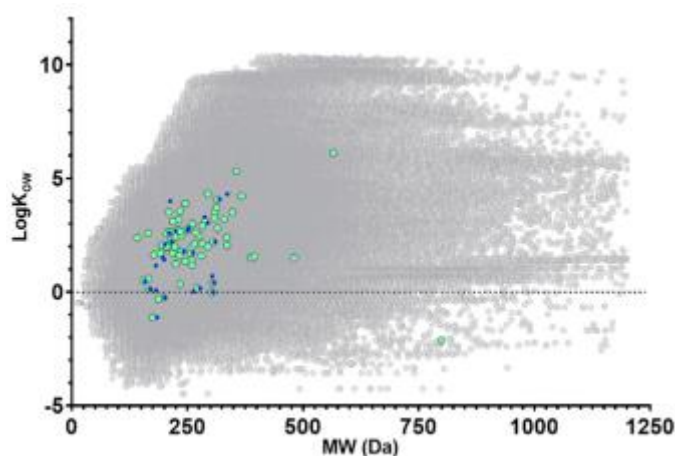


Figure 1 : Empreinte HRMS d'un échantillon illustrée via le poids moléculaire (MW ; axe x) et la polarité ($\log K_{ow}$; axe y). Chaque point représente un signal détecté ; les points gris sont les inconnus, les points verts sont identifiés via une comparaison avec une base de données et les points bleus sont les étalons internes. [11].

Les empreintes HRMS sont des données acquises et archivées par chaque laboratoire réalisant des analyses non-ciblées. Ces empreintes peuvent ensuite être traitées rétrospectivement par le laboratoire effectuant les analyses ou par tout autre organisme ayant accès aux données brutes.

2.2 TRAITEMENT DES EMPREINTES HRMS

Selon le questionnement scientifique visé, le traitement par la voie non ciblée consiste à comparer les empreintes HRMS entre elles, pour identifier les entités communes et les entités particulières, à l'aide de statistiques, logiciels fournisseurs ou en libres accès, scripts, etc... [12-15].

À l'heure actuelle la communauté scientifique utilise ce mode de traitement pour trois objectifs : les évolutions spatiales, les évolutions temporelles et l'identification de sources de pollution. Des exemples d'applications pour chacun de ces objectifs sont détaillés dans les paragraphes ci-dessous.

2.2.1 Evolutions spatiales

Un des grands enjeux environnementaux abordé par la comparaison d'empreintes HRMS est l'évolution spatiale de la pollution. Un exemple de l'utilisation d'empreintes HRMS pour évaluer l'évolution spatiale de la contamination le long du Rhin est présenté dans l'étude de Ruff et al. [16]. Après avoir modélisé la vitesse de déplacement de la masse d'eau le long du Rhin, la même masse d'eau a été analysée au cours de son déplacement, dans les différentes stations de surveillance le long du Rhin. Les empreintes HRMS acquises sur toutes les stations ont été comparées de l'amont vers l'aval. Pour chaque station, des entités ont été mises en évidence dès lors qu'elles n'étaient pas présentes en amont et répondaient à certains critères (incertitudes sur la masse, sur la fenêtre de temps de rétention, etc.). Cette étude a ainsi permis de mettre en évidence plusieurs zones à l'origine d'une pollution spécifique, qui a souvent été reliée à une activité industrielle.

De manière similaire, des travaux ont été menés sur la rivière Holtemme, un affluent de la Bode en Allemagne [17]. Cette rivière de 47 km de long prend sa source dans le parc national des montagnes du Harz jusqu'à sa confluence avec la Bode ; elle traverse une zone d'agriculture intensive et deux villes de taille moyenne dotées de stations d'épuration, qui se déversent dans la rivière. Une analyse de clustering a été réalisée via le package 'kml' sur R à partir de toutes les entités détectées ($n = 23485$) dans les empreintes HRMS des 16 échantillons prélevés le long de l'Holtemme. En amont de l'analyse de clustering la hauteur de pic de chaque entité a été normalisée par l'intensité du pic correspondant à l'étalon interne le plus proche en temps de rétention pour tenir compte des effets matriciels. Puis les hauteurs de pic normalisées ont été mises à l'échelle de la variance unitaire afin que toutes les entités aient une variance égale. Cette analyse a permis de distinguer trois groupes d'entités (Figure 2) parmi toutes les entités :

- 1) Le premier groupe rassemble des entités ($n = 9811$) ayant une intensité maximale en aval des deux stations d'épuration, et des intensités faibles ou inexistantes dans les eaux en amont. Ces entités sont donc considérées comme marqueur d'une pollution d'eaux usées. Ce groupe est appelé groupe eaux usées (WW) et est représenté par la courbe rouge (Figure 2). Celle-ci présente 2 pics qui se distinguent pour les sites 17 et 31 qui sont situés après une station d'épuration.
- 2) Un second groupe rassemble des entités ($n = 7776$) présentant une augmentation distincte et soudaine de leur intensité au dernier site d'échantillonnage dans la rivière, qui correspond à la zone de mélange avec la Bode. Ce groupe est appelé groupe Bode (BR) et est représenté par la courbe verte (Figure 2). L'évolution de cette courbe reste constante le long de tous les sites d'échantillonnage sur l'Holtemme, sauf pour le dernier site où des entités caractéristiques de la Bode sont détectées.
- 3) Un troisième groupe rassemble des entités ($n = 5910$) ayant des intensités plus élevées dans les stations en amont des stations d'épuration, avec une diminution en aval, potentiellement due à la dilution ; il est appelé groupe d'apport diffus (DRI) et est

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

représenté par la courbe bleue. Ce groupe contient principalement des entités non identifiées qui ont une intensité élevée tout au long du cours d'eau ; il pourrait potentiellement s'agir de composés présents naturellement dans l'Holtemme.

Le fond grisé représente toutes les entités détectées (n=23485) dans les tous les échantillons. Une entité est représentée par une courbe grise qui permet de visualiser son évolution le long du cours d'eau.

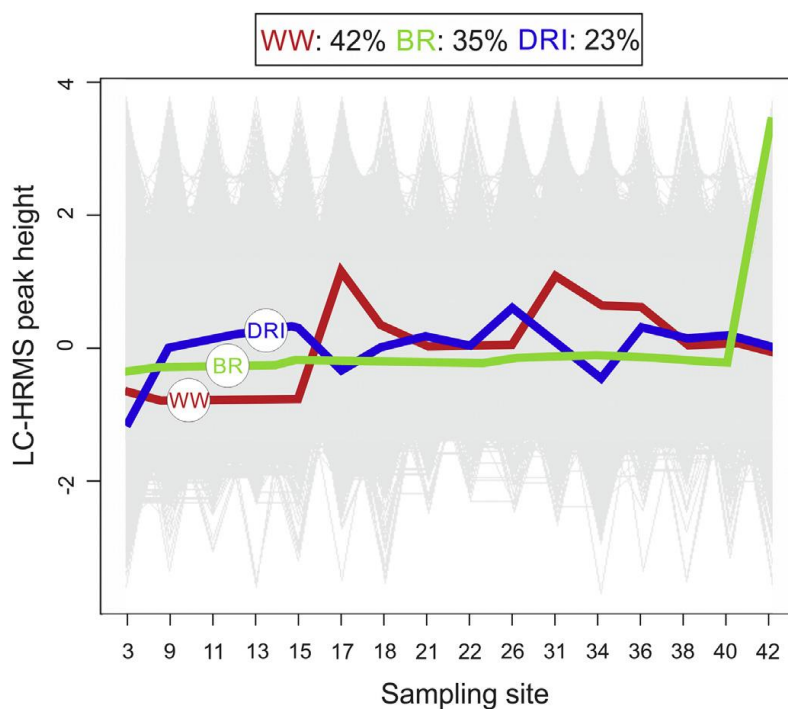


Figure 2 : Evolution des principaux groupes (eaux usées (WW), Bode (BR) et apport diffus (DRI)) le long de l'Holtemme. Le fond gris représente l'évolution de l'intensité des entités détectées sur tous les sites d'échantillonnage (l'axe x représente la distance du site depuis le point amont de la rivière, en kilomètre). Les pourcentages correspondent au pourcentage des entités attribuées à la source de pollution [17].

Dans une étude menée dans les bassins versants des rivières Saale et Mulde [18], qui sont d'importants affluents de l'Elbe en Allemagne, une nouvelle approche a été proposée pour identifier et classer par ordre de priorité les échantillons présentant une contamination importante spécifique à un site, sur la base des empreintes HRMS, sans connaissance préalable des composés présents. Pour cela des « scores de rareté » ont été calculés sur les échantillons. Ce score prend en compte toutes les entités apparaissant dans les 31 échantillons d'eau de surface et combine l'intensité maximale du signal d'un pic dans tous les échantillons avec sa fréquence de détection en un seul chiffre. Dans cette étude, environ 1 % des entités avaient des valeurs de scores de rareté supérieures à 1000 et ont donc été considérées en priorité comme représentant des composés spécifiques à un site. Pour donner la priorité aux sites présentant une contamination spécifique, le nombre d'entités présentant des scores de rareté supérieurs aux seuils de 5 000 et 1 000 a été comparé (Figure 3). Cette étude a permis de classer les entités à essayer d'identifier mais aussi de classer les sites présentant un nombre d'entités très élevé afin d'y mener des recherches plus approfondies.

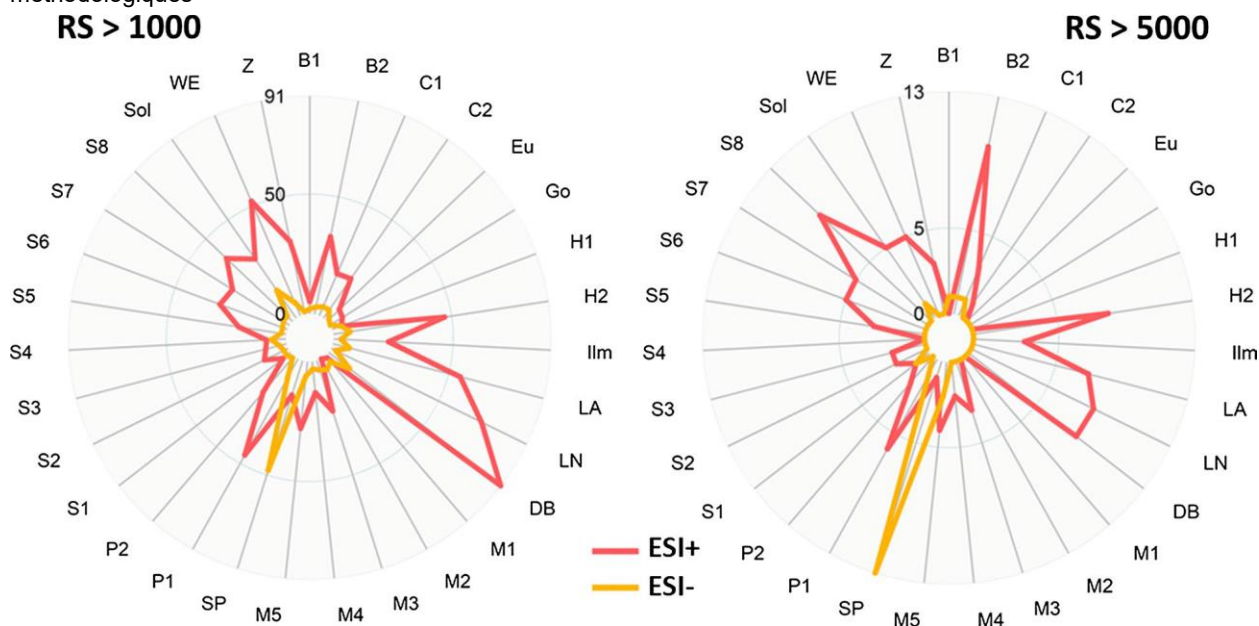


Figure 3 : Comparaison des sites étudiés pour le nombre d'entités au-dessus des scores de rareté (RS) de 1000 et 5000 [18].

D'autres outils ont également été développés afin de comparer des empreintes HRMS pour établir une potentielle évolution spatiale et repérer des zones d'attention particulière. Par exemple, certains auteurs ont proposé un classement basé sur la fréquence de détection des entités acquises [15] ; l'ensemble des données d'analyses de suivi de la mer noire a été utilisé pour générer des images représentant la distribution spatiale de chacun des signaux. Ces images ont ensuite servi d'entrée à un modèle de classification par réseau neuronal convolutif et l'étude a abouti à l'estimation de la charge de pollution par les produits chimiques apportés par les deux principaux fleuves entrants (Danube et Dniepr) et d'autres sources non identifiées jusqu'à présent. Le modèle présenté s'est avéré particulièrement utile pour hiérarchiser les signaux de composés inconnus permettant ainsi de focaliser les efforts d'élucidation structurale sur les signaux les plus pertinents selon le modèle.

Dans cette optique d'étude de pollution et dans le cadre des travaux réalisés au sein du réseau NORMAN¹, des travaux ont été menés afin d'archiver des empreintes HRMS et pouvoir confronter des données d'origines différentes pour repérer des contaminations spécifiques de certaines zones géographiques, ou, au contraire, des pollutions ubiquistes [19-21]. La Figure 4 est un exemple de résultat visualisé sous la forme d'une carte interactive par la *Digital Sample Freezing Platform* (DSFP) [19].

¹ <https://www.norman-network.net/>

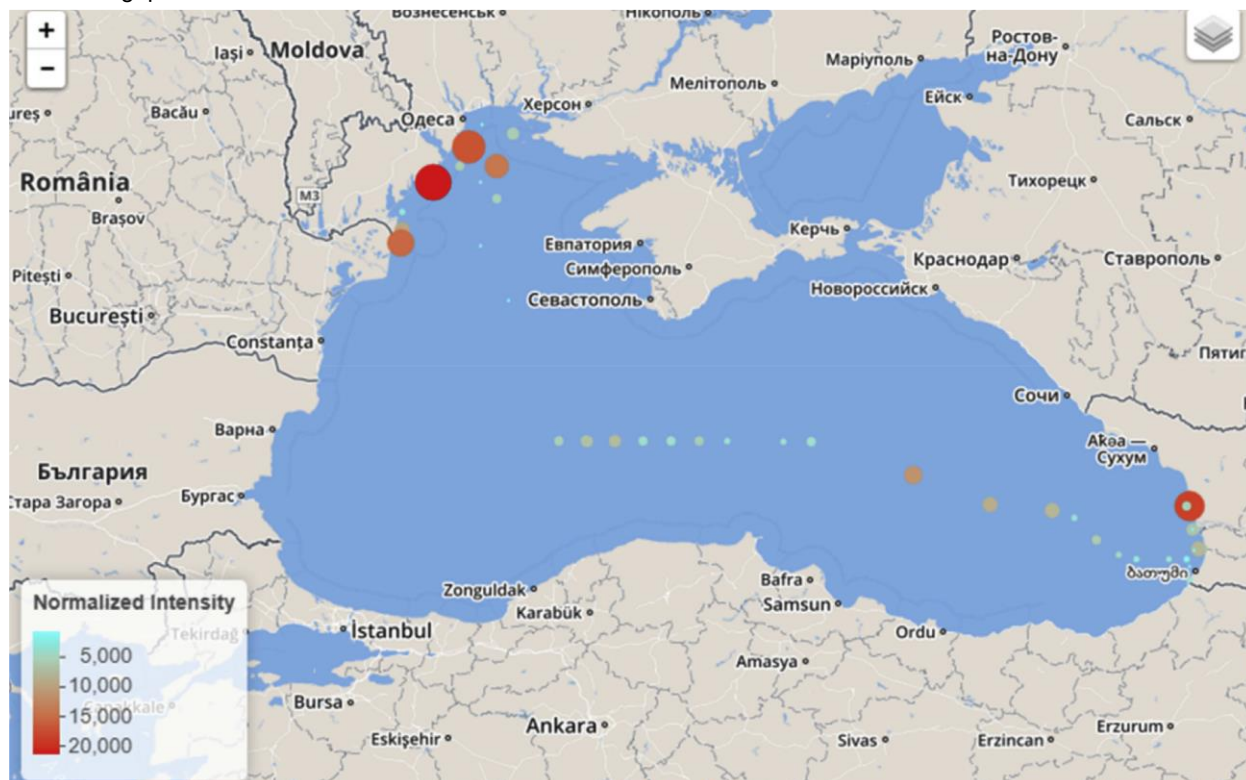


Figure 4 : Distribution spatiale du DEET dans des échantillons d'eau de mer provenant de la mer Noire. Le profil révèle qu'il y a un apport terrestre de DEET dans la mer Noire. Le DEET a été détecté dans 39 (sur 54) échantillons d'eau analysés par LC-HRMS et en utilisant l'analyse non-ciblée. Les cercles indiquent les stations d'échantillonnage ; les petits cercles bleu clair correspondent à des intensités faibles de DEET ; les grands cercles rouges indiquent des intensités élevées de DEET [19].

2.2.2 Evolutions temporelles

Les empreintes HRMS peuvent également être utilisées afin d'étudier des évolutions temporelles. Dans ce cas, ces études peuvent tirer parti des analyses rétrospectives, rendues possibles par l'analyse non-ciblée, afin d'étudier sur des périodes antérieures et sur des longues périodes les variations de signaux [20]. Cette stratégie nécessite cependant de pouvoir s'assurer de la comparabilité des analyses par l'utilisation de contrôles qualité, et de disposer d'un système d'archivage de données.

Les comparaisons d'empreintes HRMS par étude d'évolutions temporelles peuvent être réparties en deux catégories en fonction de leurs objectifs.

La première s'appuie sur la détection de valeurs « aberrantes » pour des signaux détectés régulièrement dans la série temporelle étudiée. Elle est souvent utilisée pour diagnostiquer des déversements accidentels en repérant des fortes variations soudaines dans l'intensité d'une entité. Pour cela, différentes méthodes peuvent être utilisées, certaines s'appuyant sur des tests statistiques [22] alors que d'autres s'appuient sur l'établissement de pseudos-cartes de contrôles pour définir des valeurs dites « normales » et repérer les signaux qui sortent de l'intervalle défini comme normal [12], [16], [23].

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

Pour illustrer cette catégorie, on peut citer la Commission Internationale pour la Protection du Rhin² (CIPR) au sein de laquelle des stations de mesure le long du Rhin déterminent depuis plusieurs années la pollution des masses d'eau, sur la base d'une sélection de substances ciblées définie *a priori*. Ces dernières années, l'analyse non ciblée par spectrométrie de masse à haute résolution couplée à la chromatographie liquide (LC-HRMS), a été mise en place avec succès dans les stations de mesure suisses et allemandes sur le Rhin [12], [16]. Ces stations de mesure sont pionnières dans l'application de l'analyse non-ciblée. Des échantillons sont prélevés et analysés quotidiennement. Pour chaque station de mesure, les empreintes HRMS obtenues chaque jour sont comparées afin de mettre en évidence l'apparition d'une entité ou l'augmentation d'une entité potentielle. Lorsque c'est le cas, une tentative d'identification est menée afin de rechercher la source potentielle de cette pollution (voir exemple en Figure 5).

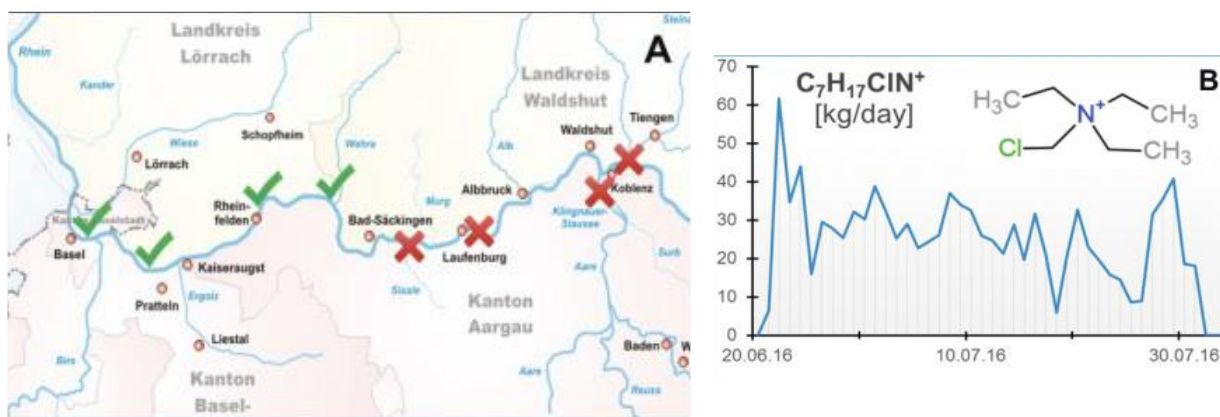


Figure 5 : A : Sites d'échantillonnage le long du Rhin avec des résultats positifs (marques vertes) et négatifs (croix rouges) de la substance N-(chlorométhyl)-triéthylammonium cation. B : Sur la base de ces résultats, la source a été identifiée et l'émission stoppée..[12]

Une seconde catégorie d'études temporelles en NTS consiste à étudier une variation constante d'un signal. Ce type d'étude est généralement retrouvé pour l'analyse de matrices biologiques et permet notamment de détecter et éventuellement identifier des substances bioaccumulables lorsqu'un signal augmente régulièrement au cours de la série temporelle [24]. Au contraire, certains auteurs choisissent de se concentrer sur les signaux dont la diminution est régulière au cours de la série temporelle. Ces signaux peuvent notamment donner une indication sur de potentielles mesures de restrictions ou des changement de consommation qui induisent une diminution d'un certain type de composés dans l'environnement [25].

2.2.3 Discrimination de sources

Outre ces études d'évolution spatiales ou temporelles, l'étude d'empreintes HRMS peut aussi permettre de différencier des sources de pollution. Les empreintes sont ainsi comparées afin de mettre en évidence des entités qui permettraient de différencier des groupes d'échantillons.

Notamment, de nombreuses études dans le domaine agroalimentaire sont effectuées en ce sens afin de déterminer l'origine de matières premières ou encore de type de matières premières utilisées dans les produits. Là où certaines ont permis de remonter à l'origine géographique de plus d'une centaine de grains de maïs (Figure 6) [26], d'autres ont permis la distinction entre des

² Commission Internationale pour la Protection du Rhin (CIPR) : <https://www.iksr.org/fr/iksr/rhein-2040/rhine-project-non-target-screening>

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

raisins d'origine biologique ou non [27], ou encore la distinction entre différentes variétés d'huile d'olive [28].

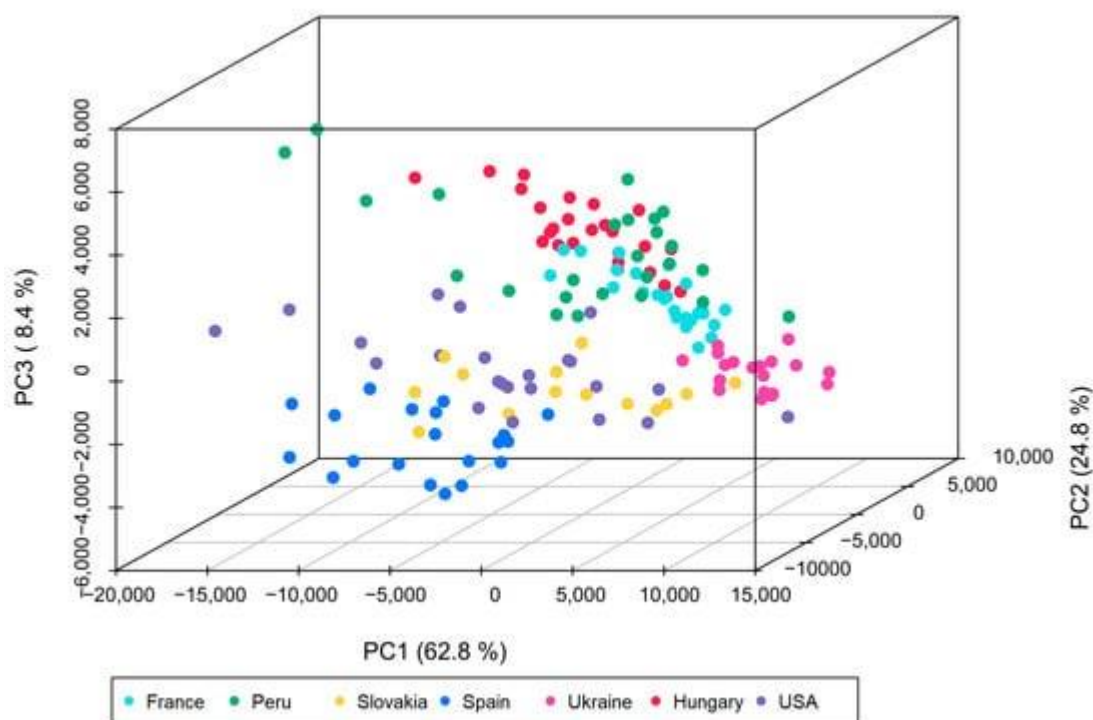


Figure 6 : Des analyses de données multivariées et des statistiques univariées ont été utilisées pour mettre en évidence les entités liées à l'origine géographique de grains de maïs. La classification utilisant seulement 20 entités sélectionnées avec l'algorithme Random Forest a permis de classer correctement 90,5 % des échantillons avec une validation croisée 10 fois répétée 100 fois. Les scores de l'ACP ont été créés en utilisant les aires des pics de ces 20 entités, après le prétraitement par la méthode de l'échelle de Pareto.

Dans un autre type d'étude, la comparaison d'empreintes HRMS a permis de déterminer la nature de matières premières brûlées (Figure 7) [29] et potentiellement d'en déduire le procédé de combustion [30].

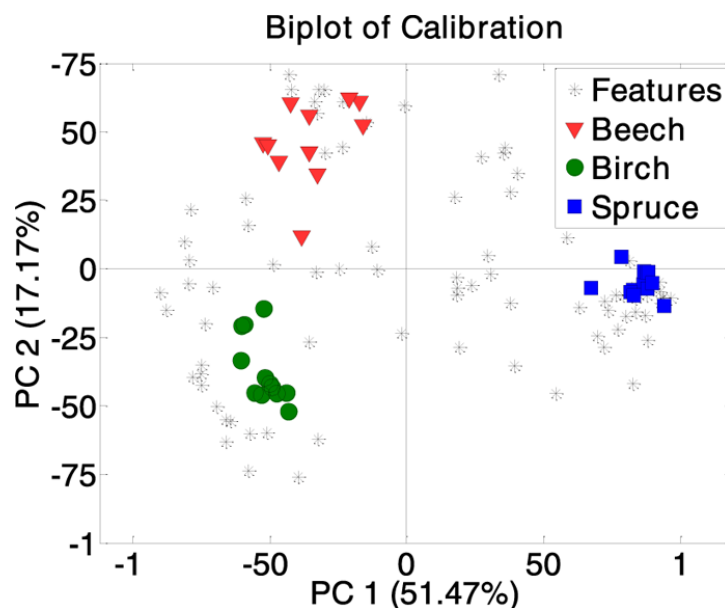


Figure 7 : Les résultats de l'ACP sont affichés à l'aide d'un biplot des variables et observations. Un regroupement d'entités en fonction des types de bois mesurés et une séparation entre le bois tendre et le bois dur sont observés sur le premier axe. Une discrimination plus poussée des échantillons issus de la combustion de bois tendre est observée sur le deuxième axe. Les entités ont été attribuées au type de bois en fonction de leur emplacement dans le biplot [29]. Beech = hêtre, birch = bouleau et spruce = épicéa

Pour suivre les entités dans des échantillons d'eau de stations d'épuration avec un post-traitement d'ozonation, une étude a permis de regrouper les entités par regroupement hiérarchique de type *clustering* et ainsi de mettre en évidence des tendances de comportement (dégradation, formation, ...) au long des traitements de la station [31]. La Figure 8 présente les 10 tendances mises en évidence dans cette étude.

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

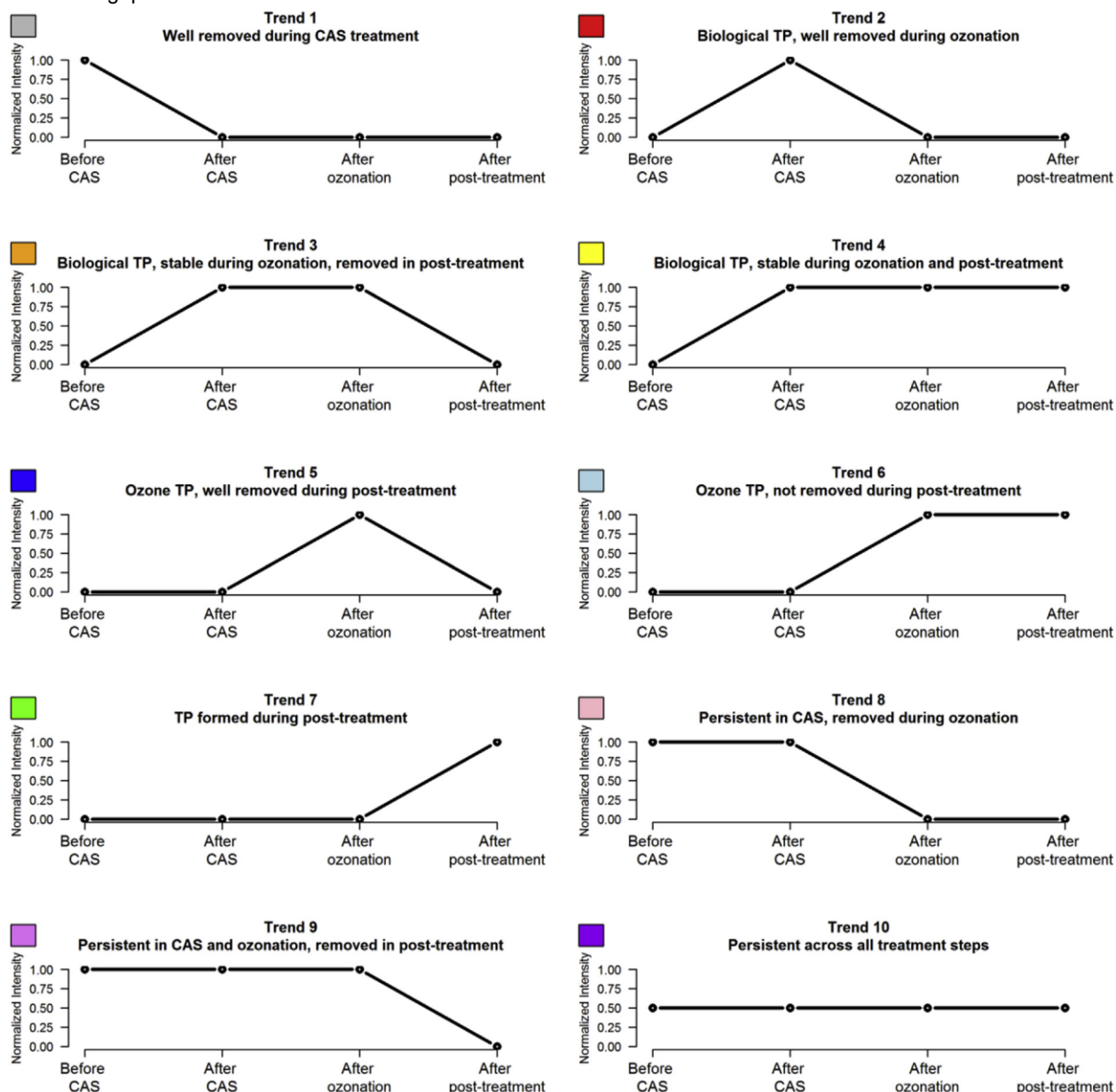


Figure 8 : Application du NTS pour l'étude du comportement de composés chimiques en station d'épuration ; exemples des dix tendances définies pour décrire les profils d'intensité dans un cluster. Chaque tendance peut être associée au comportement spécifique d'un composé dans la station d'épuration et est décrite dans le titre de chaque tendance. Les profils d'intensité pour chaque post-traitement ont été regroupés séparément et sont donc généralement intitulés "Après post-traitement". CAS désigne le traitement conventionnel par boues activées [31].

Au cours de cette même étude, une comparaison des entités avant et après traitement d'ozonation a permis de mettre en évidence 1552 entités comme potentielles molécules parents ayant subi une transformation ou potentiels produits de transformation d'ozonation (Figure 9), sur 6207 entités [31].

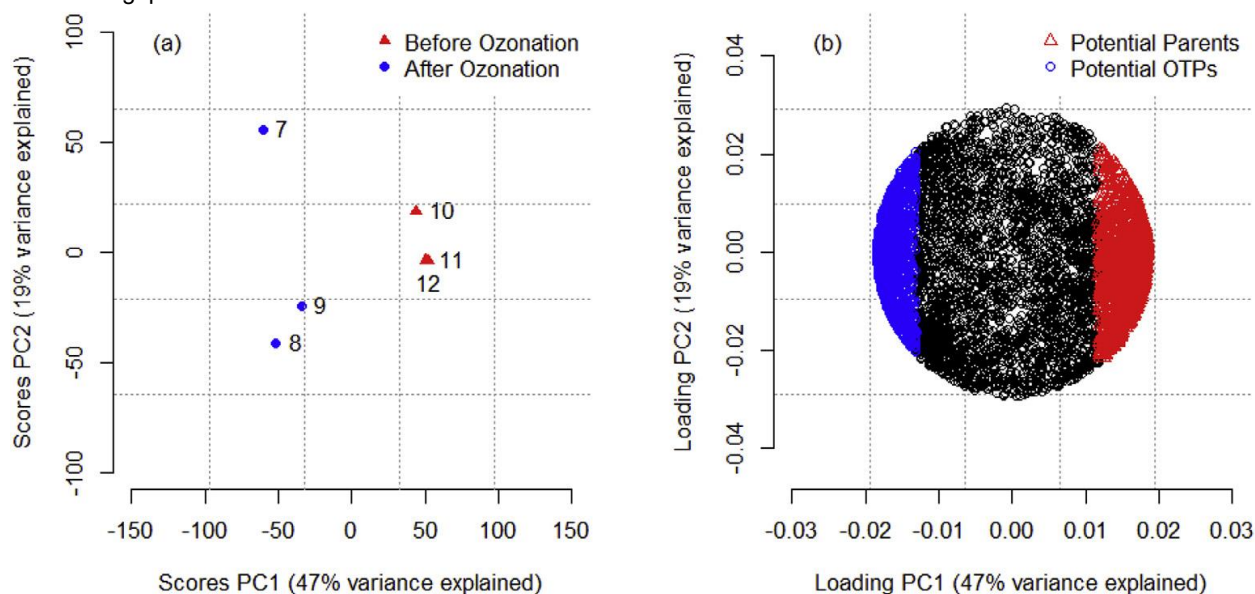


Figure 9 : Analyse en composantes principales (a) scores plot et (b) loading plot de l'effluent de la station d'épuration avant et après ozonation pour une dose d'ozone spécifique de 0,5 gO₃/gDOC mesurée en ESI positive. Les première et deuxième composantes principales sont indiquées sur les axes x et y, respectivement. Une bonne séparation entre les échantillons avant ozonation (triangles rouges) et après ozonation (cercles bleus) a été obtenue le long de l'axe des x et a pu expliquer 47 % de la variance, ce qui a permis de classer les entités non ciblées comme des composés parents potentiels ayant subi une transformation ou des produits de transformation d'ozonation potentiels (ou ni l'un ni l'autre, cercles noirs).

Enfin, d'autres études comparent les empreintes HRMS de sites d'usages différents afin de trouver des empreintes spécifiques à un type d'activité. Cette méthode a par exemple permis de distinguer des sites urbains avec des différences en termes de trafic routier [32]. Ces études se différencient des évolutions spatiales dans le sens où seule la connaissance de l'utilisation du site est prise en compte et les échantillons correspondants sont comparés strictement sur la base de cette connaissance.

3 Quels sont les outils permettant le traitement des empreintes HRMS ?

Après l'acquisition des données, un prétraitement est extrêmement important pour réduire la quantité et la complexité des données [33]. Plusieurs étapes sont cruciales :

- Une conversion des données brutes est nécessaire si le traitement des empreintes HRMS est effectué à l'aide d'outils en libre accès ou n'acceptant pas le format de données d'acquisition [10].
- La détection des pics (« *Peak detection* »), étape au cours de laquelle les informations relatives à la masse et aux pics chromatographiques sont combinées pour produire une entité tridimensionnelle, c'est-à-dire une entité composée de la masse, du temps de rétention et de l'intensité/de l'aire [11].
- Le regroupement des entités, aussi appelé « *componentisation* », étape qui permet de regrouper les isotopes, adduits, ions multichargés et les fragments en source pour définir les composants, c'est-à-dire le regroupement de tous les signaux susceptibles d'appartenir à une structure moléculaire unique [33]. Une « *componentisation* » efficace

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

peut potentiellement réduire le nombre total d'entités et s'appuie sur des tolérances définies par l'utilisateur pour le temps de rétention, la différence de masse (c'est-à-dire les isotopes et les adduits) et la similarité du profil chromatographique [11].

- La soustraction du blanc, étape importante pour diminuer le nombre d'entités. Si certaines entités sont détectées dans les blancs, il est possible de filtrer les données brutes sur une valeur d'aire par exemple pour ne pas voir cette entité dans les échantillons.
- L'alignement du temps de rétention et de la masse des entités sur plusieurs échantillons permet de donner la priorité à des entités spécifiques en vue d'une évaluation plus approfondie par la suite [33].

Il existe plusieurs possibilités, tant commerciales (par exemple Waters, ThermoFisher, Agilent, SCIEX) qu'en libre accès pour réaliser ces différentes étapes. Celles-ci sont brièvement détaillées dans les paragraphes ci-dessous.

3.1 LES OUTILS COMMERCIAUX

3.1.1 Waters

Les instruments analytiques provenant de chez Waters peuvent être pilotés par MassLynx pour les anciens modèles ou UNIFI pour les plus récents. Le traitement des empreintes HRMS se fait désormais via UNIFI. Si l'appareil analytique est piloté par MassLynx, un import des données brutes dans UNIFI est nécessaire.

UNIFI permet de faire un prétraitement des empreintes HRMS : détection des entités, componentisation et alignement. Il est ensuite possible de faire un traitement suspect par comparaison avec des bases de données pour identifier des molécules. L'alignement des entités sur l'ensemble des échantillons est réalisé uniquement pour les molécules identifiées par traitement suspect. Le traitement non-ciblé des empreintes HRMS est quant à lui possible grâce à l'outil MultiVariate Analysis (MVA) qui permet de générer une matrice contenant toutes les entités de l'ensemble des échantillons, en essayant de faire l'alignement du mieux possible à partir des filtres choisis. Cette matrice peut ensuite être exportée dans EZInfo (Umetrics), outil complémentaire au logiciel UNIFI, qui permet de réaliser de nombreuses analyses multivariées. Parmi celles-ci, on retrouve les Analyses par Composante Principales (ACP), les régressions par moindres carrés discriminantes (Partial least squares-discriminant analysis ou PLS-DA) et régressions par moindres carrés orthogonales discriminantes (Orthogonal partial least squares discriminant analysis ou OPLS-DA). Différentes visualisations graphiques des scores sont disponibles (loadings plot, bi-plot, s-plot) ainsi que les scores des prédictions générés par le modèle [34]. Les données et sélections d'entités réalisées dans EZInfo ne sont pas exportables, il est nécessaire soit de retransférer les données dans UNIFI, copier et coller manuellement dans un tableur, et récupérer via des captures d'écran soit d'imprimer au format *.pdf [34].

3.1.2 Agilent

Les instruments analytiques provenant de chez Agilent sont pilotés par *MassHunter*. Le traitement des données HRMS peut se faire directement sous *MassHunter Qualitative* pour l'extraction de pics d'intérêts à des fins d'identifications en traitement non-ciblé ou par traitement suspect via l'algorithme Molecular Feature Extraction (MFE).

Pour ce qui est du traitement des empreintes HRMS, l'extraction des données est effectuée par *Profiler*: détection, componentisation et alignement. Cette extraction se fait en réalité via un

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

algorithme en 8 étapes qui inclut l'application ou non de différents filtres et une vérification par extraction récursive. Une fois les données extraites, elles sont agrégées sous forme d'une matrice contenant toutes les entités de l'ensemble des échantillons. Cette dernière peut être exportée en format *.csv pour une analyse dans des logiciels tiers (e.g., excel, R...) ou sous format *.cef pour utilisation sous la suite *Mass Profiler* (Agilent).

Mass Profiler est un logiciel d'analyse statistique permettant de comparer différentes populations (échantillons ou groupes d'échantillons). Différentes visualisations (e.g., nuages de points, diagrammes de Venn) et tests statistiques uni- ou multivariés (e.g., t-test, ANOVA, hierarchical clustering, PLS-DA...) sont proposées dans ce logiciel et permettent de sélectionner des entités par ID (i.e., couple masse/temps de rétention). Ces ID sont liées, grâce au logiciel ID Browser, aux bases de données utilisées pour le traitement suspect et permettent ainsi la tentative d'identification de ces entités sélectionnées.

3.1.3 Thermo Scientific

Les instruments analytiques de type Orbitrap provenant de chez ThermoFisher Scientific sont généralement pilotés par Xcalibur.

Le traitement des empreintes HRMS est fait par le logiciel d'identification de petites molécules Compound Discoverer. Ce logiciel permet d'accéder à de nombreuses ressources en ligne et hors ligne, et utiliser des algorithmes intelligents lorsqu'il n'y a pas de correspondance spectrale directe pour aider à identifier un composé inconnu. Le *workflow* non ciblé de ce logiciel permet de simplifier le processus de sélection de détection des pics, de soustraction des blancs, de regroupement des entités (componentisation), de proposition de formules moléculaires, de comparaison des profils isotopiques, d'évaluation des adduits, d'attribution et de comparaison des profils de fragmentation et de recherche de bases de données, le tout dans un seul logiciel au lieu de plusieurs. Le logiciel Mass Frontier complète Compound Discoverer par l'accès et la création de bibliothèques spectrales.

3.1.4 Sciex

Les instruments analytiques provenant de chez Sciex sont pilotés par Sciex OS qui permet également un traitement suspect des données via le module Analytics qui permet la confrontation des spectres acquis avec différentes bibliothèques. Pour ce qui est de l'exploitation des empreintes HRMS, le logiciel PeakView (Sciex) est utilisé afin de générer une matrice contenant les entités extraites dans tous les échantillons. Cette matrice peut ensuite être exportée dans le logiciel MarkerView (Sciex) pour des exploitations statistiques. Le logiciel propose notamment les ACP avec une représentation graphique de scores associée (loadings plots). A partir de cette représentation des scores, la présence des entités sélectionnées peut être visualisée par le biais des profils plots (i.e., représentation des aires relatives des entités dans chaque échantillon) ou par le biais de boîtes à moustache pour comparer plusieurs groupes d'échantillons. MarkerView propose également un groupement des variables par composantes principale (PCVG pour principal component variable grouping) afin de déterminer les entités qui évoluent de manière similaire statistiquement. Une sélection statistique des entités d'intérêt peut ainsi être effectuée et ces dernières peuvent être ré-exportées vers le module Analytics pour une tentative d'identification.

3.2 LES OUTILS EN LIBRE ACCES

De nombreux outils en libre accès sont disponibles ; une liste non-exhaustive est présentée dans le Tableau 1. Chaque outil a des spécificités par exemple en termes de langage (R, Python,

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

Matlab, etc...), de format de données compatibles, de logarithme utilisé pour le prétraitement des données, la visualisation des résultats. Certains outils offrent la possibilité de voir chaque étape et parfois sélectionner un paramètre plutôt qu'un autre (patRoos par exemple).

Tableau 1 : Liste non exhaustive des outils en libre accès disponibles pour traiter les empreintes HRMS

Nom	Langage	Compatibilité / Accessibilité	Prétraitement de données	Traitements statistiques / visualisation	Références	Téléchargement / accès
xcms	R	*.mzXML, *.mzData, *.cdf, *.d (Agilent et Bruker), *.wiff (Sciex), *.RAW (Waters et Thermo)	<p><u>Peak détection</u> : algorithme <i>centwave</i></p> <p><u>Componentisation</u> : adduits et isotopes sont annotés mais pas soustraits</p> <p><u>Alignement</u> : méthode <i>adjustRtime</i> avec des algorithmes différents dont <i>Orbiwarp</i> permettant un alignement à partir de groupes d'entités présents dans tous les échantillons et répartis uniformément sur le profil chromatographique.</p> <p><u>Filtration</u> : Possibilité de filtrer la matrice finale en temps de rétention, de soustraire des entités particulières</p>		[35], [36], [37], [38], [39], [40]	https://bioconductor.org/packages/release/bioc/html/xcms.html
CAMERA	R	Objet R	<p>Componentisation uniquement</p> <p>Ce package R peut être utilisé après avoir réalisé le peak picking via un autre outil.</p> <p>Permet d'annoter les spectres pour les isotopes et adduits.</p>		[40], [41], [42]	https://www.biocconductor.org/packages/release/bioc/html/CAMERA.html
Non Target	R	Objet R	<p><u>Componentisation</u> : inclus aussi les séries d'homologues</p> <p>Permet également de tracer et de filtrer les données HRMS pour les défauts de masse, les distances fréquentes ou caractéristiques de m/z</p>		[43], [44], [45]	https://github.com/blosloos/nontargetenvibee
TimeCourse	R	Matrice contenant les features (généralement *.csv) préalablement transformée	R-package utilisé pour prioriser les entités à identifier à l'aide de fonctions d'analyse des données et d'affichage graphique pour les données temporelles	<p>Profiling par étude des tendances en intensité</p> <p>Étude des tendances de Bayes</p>	[42], [46]	https://www.biocconductor.org/packages/release/bioc/html/timecourse.html
MZMine	Java	*.mzML, *.mzXML,	<u>Peak detection</u> : plusieurs algorithmes disponibles en fonction du format des données	Scatter plot, distribution en m/z et	[36], [45],	http://mzmine.github.io/

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'application et besoins méthodologiques

		*.mzDATA, *.mzTab, *.csv	<u>Componentisation</u> : les isotopes sont assignés à la masse exacte et les adduits supprimés <u>Alignement</u> : par l'algorithme <i>join aligner</i> <u>Normalisation</u> : par les hauteurs ou aires des entités pour une injection <u>Filtration</u> : sur réplacats, par perte de neutres, par distribution isotopique et soustraction de blancs	temps de rétention, défaut de masse de Kendrick, van Krevelen plot Profiling des intensités Annotation de pics	[47], [48]	
Open Source Peak Picking	R	*.mzML *,.mzXML,	Détermination des faux positifs <u>Peak detection</u> <u>Componentisation</u> <u>Alignement</u>		[14]	
patRoön	R	*.mzXML, *.mzML, *.d (Bruker)	<u>Peak detection</u> : s'appuie sur xcms ou OpenMS <u>Componentisation</u> : est réalisée avec CAMERA, nontarget, RAMClustR ou OpenMS <u>Alignement</u> s'appuient sur xcms ou OpenMS <u>Filtration</u> : soustraction de blancs, filtres sur réplacats, intensité, scores de componentisation...	Profiling par étude des tendances en intensité Diagramme de Venn, Upset et Chord Dendrogrammes heatmaps	[49]	https://github.com/rickhelmus/patRoön
MetaboAnalyst	R	*.mzXML, *.mzML, *.CDF, *.mzDATA	MetaboAnalyst est une plateforme web dédiée à l'analyse complète des données métabolomiques, à l'interprétation et à l'intégration avec d'autres données omiques. <u>Peak detection</u> : s'appuie sur xcms ou OpenMS <u>Componentisation</u> est réalisée avec CAMERA, <u>Alignement</u> s'appuient sur xcms ou OpenMS <u>Normalisation</u> : par la médiane des intensité pour chaque injection et log transformation	Student, ANOVA, étude des tendances de Bayes, fold-change, volcano plot, PCA, PLS-DA, OPLS-DA, dendrogrammes, heatmap, SVM, Random forest	[50]	https://www.metaboanalyst.ca/
W4M	Matlab	*.mzXML ou données prétraitées par ailleurs et reformatées sous 3 fichiers *.csv spécifiques (décrits sur W4M)	Portail collaboratif dédié au traitement des données métabolomiques, à l'analyse et à l'annotation pour la communauté de la métabolomique. <u>Peak détection</u> via xcms <u>Componentisation</u> via CAMERA <u>Alignement</u> via xcms	Student, Wilcoxon, ANOVA, Kruskal-Wallis, Pearson, Spearman PCA, PLS(-DA), OPLS(-DA)	[38], [51]	https://workflow4metabolomics.org/

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'application et besoins méthodologiques

OpenMS	Python , C++	*.mzML	Permet la gestion, l'analyse et la visualisation des données LC/MS.		[52]	https://openms.de/
Envimass	R	*.mzXML, *.raw (Thermo)	<u>Peak detection</u> : paramétrage automatique avec soustraction du bruit de fond <u>Componentisation</u> : les isotopes et les adduits sont assignés à la masse exacte <u>Alignement</u> : en temps de rétention et avec une recalibration en masse possible <u>Normalisation</u> : basée sur la médiane des intensités ou par utilisation d'étalons internes <u>Filtration</u> : soustraction de blancs, filtration sur réplicats	Profiling par étude des tendances en intensité ou par tendance de détection Fold change, volcano plots, Venn PCA, dendrogrammes, heatmaps	[36], [53], [54], [55]	https://www.envibee.ch/eng/enviMass/overview.htm
RAMClustR	Matlab , R	xcms object ou *.csv respectant certaines conditions détaillées dans l'article correspondant	<u>Componentisation uniquement</u> : par création d'une matrice de similarité entre les entités d'une même injection d'une part, et la corrélation entre les entités d'injections différentes en utilisant le coefficient de corrélation de Person		[56]	https://github.com/sneumann/RAMClustR

3.3 AVANTAGES / INCONVENIENTS

Les outils commerciaux sont très pratiques et ne nécessitent pas de convertir les formats de données. Cependant, certains de ces outils restent incomplets lorsqu'il s'agit de faire des traitements non-ciblés. De manière générale, il est difficile de contrôler leur algorithme en modifiant certains paramètres clés comme cela peut être fait sur des outils en libre accès.

Les outils en libre accès sont nombreux et, pour sélectionner le ou les plus adaptés, l'opérateur doit tenir compte : i) de l'objectif du travail (nécessité de faire toutes les étapes du prétraitement des données ou qu'une partie, besoin de traitements statistiques ou non, sur quel type de données porte le travail, etc...) et ii) des outils les plus abordables en lien avec ses connaissances (format de données, langage utilisé, aide possible dans la bibliographie ou par l'outil en lui-même).

Certains outils peuvent présenter un avantage :

- xcms permet de traiter les données sans utiliser les ressources locales et de partager des données avec la communauté ; cela est simplifié du fait de sa disponibilité en ligne.
- patRoön permet de regrouper plusieurs algorithmes existants sous une même plateforme avec une harmonisation des formats de données. Cela facilite le passage d'un algorithme à l'autre, notamment lorsqu'il y a une volonté de faire des comparaisons.

4 Assurance qualité et contrôle qualité (QA/QC)

Pour toute analyse, les conditions d'échantillonnage, de préparation et d'analyse d'un échantillon peuvent impacter le résultat. Afin d'évaluer et/ou de corriger ces impacts, il est nécessaire d'utiliser des pratiques d'assurance et de contrôle qualité. Dans le cas de la comparaison d'empreintes, inter-laboratoires ou intra-laboratoire, il faut pouvoir valider les données utilisées. Pour cela, des exigences minimales de qualité doivent être établies [57].

Il existe des guides définissant des recommandations en terme de QA/QC pour l'utilisation des données d'analyse non-ciblée, que ce soit en métabolomique [58-63], dans le domaine de l'environnement [6], [11] ou d'un point de vue général, c'est-à-dire sans cibler sur un domaine en particulier [64].

4.1 LES PREREQUIS ANALYTIQUES

4.1.1 Randomisation de la séquence d'analyse

La randomisation des échantillons au sein d'une séquence d'analyse permet de répartir aléatoirement sur les différents échantillons le biais analytique dû à l'encrassement de la source par exemple. Si le biais est trop important, il peut masquer ou augmenter de manière fictive la variabilité que l'on cherche à étudier [65-66].

4.1.2 Réplicats d'analyse

L'analyse de plusieurs réplicats d'un même échantillon permet de définir des critères stricts pour effectuer un premier tri des données et permet de minimiser le taux de faux positifs et de faux négatifs lors de l'étape de détection des pics [67]. Alors que certaines études recommandent de faire au moins deux, voire trois réplicats [67], d'autres indiquent qu'il n'est pas obligatoire d'effectuer des réplicats dans le cas où la méthode est bien validée et que la séquence d'analyse comporte des contrôles qualités qui permettent de prouver les performances de l'analyse [61], [68].

Les critères pour trier les données à partir des résultats obtenus sur les différents réplicats d'analyse varient selon les publications. Par exemple, le tri peut être effectué sur le critère de la présence d'un analyte dans tous les réplicats [69], [31], [70].

4.1.3 Tri du signal

4.1.3.1 Blancs

Afin de s'assurer de la qualité des résultats, il est nécessaire d'avoir recours à l'utilisation de blancs pour vérifier l'absence de contamination lors des différentes étapes du processus analytique. Pour cela, il existe plusieurs types de blancs dont les blancs analytiques [71].

Dans le cas d'une analyse HRMS, ces derniers doivent être analysés au début et à la fin d'une séquence d'analyse [61]. L'analyse de blancs analytiques pendant la séquence n'est pas recommandée : les modifications de composition de matrice, entre celle des échantillons et des blancs, peuvent entraîner un déconditionnement partiel du système (décalage du temps rétention, modifications de la largeur des pics et de la résolution chromatographique...) [61]. Il existe également des blancs terrain qui permettent de prendre en compte toute contamination de l'échantillonnage à l'analyse et des blancs d'extraction permettant de prendre en compte les éventuelles contaminations au cours de l'extraction de l'échantillon et pendant le stockage avant analyse [6]. Ces derniers, ayant la même composition que les échantillons, peuvent être analysés tout au long de la séquence d'analyse.

L'utilisation des données associées aux blancs pour trier le signal est effectuée différemment selon les études. Une première possibilité est la suppression des pics par comparaison avec les blancs avec un seuil fixé par les auteurs. Ce dernier peut être basé sur le rapport d'intensité échantillon/blanc qui doit être, selon les études, supérieur à 10 [18], supérieur à 5 [69], [72] ou encore supérieur à 4 [16]. Le seuil peut également être calculé à partir de la moyenne et de l'écart-type de l'intensité des blancs [17]. Dans l'étude de Schollé et al. [31], tous les pics détectés dans les échantillons et également présents dans les blancs sont supprimés, peu importe leur intensité.

4.1.3.2 Seuils d'intensité

Il est possible d'effectuer un tri supplémentaire, basé sur l'intensité des pics, en plus du tri effectué sur le critère des blancs. Ce tri permet de retirer le bruit de fond. Le seuil choisi dépend des études. Il peut être fixé à partir d'un rapport d'aire/hauteur minimal [18] ou d'une intensité minimale [17]. Ces seuils peuvent être dépendants du mode d'ionisation, de l'équipement. Par exemple, dans l'étude de Beckers et al. [17], le seuil d'intensité minimale est fixé à 5 000 en mode d'ionisation négatif ou de 50 000 en mode d'ionisation positif.

4.1.3.3 Autres

Les pics des différents isotopes d'un même composé peuvent également être identifiés et triés si les méthodes utilisées pour le retraitement le permettent [17], [31].

4.1.4 Emploi d'étalons internes

Lors de l'analyse, il peut exister des dérives, notamment des effets de suppression d'ionisation dus à l'encrassement de la source et qui modifient le rendement d'ionisation au sein d'une même analyse ou d'expériences différentes [65]. Il faut donc maîtriser et corriger ces dérives afin qu'elles n'entravent pas l'interprétation des données. Les fluctuations des signaux observés peuvent être évaluées par des étalons internes ajoutés dans tous les échantillons.

4.1.4.1 Choix des étalons internes

L'ajout d'étalons marqués dans les échantillons permet d'évaluer l'impact des conditions de préparation et/ou d'analyse (extraction, injection, stabilité, effets matrices...) [6]. Les étalons internes doivent présenter une similitude structurale avec les composés d'intérêt et sont choisis pour représenter l'ensemble des composés à analyser. Dans le cas de l'analyse non ciblée, il est impossible de représenter l'ensemble des composés pouvant être présents dans les échantillons. Dans de précédents travaux, les étalons internes ont été sélectionnés afin de représenter le temps d'analyse chromatographique, sur les deux modes d'ionisation (ESI+ et ESI-) et si possible avec les deux types de chromatographies (GC et LC) [6], [69]. Des étalons internes ont déjà été utilisés lors d'études de comparaison d'empreintes (source/temporel) [15]. Dans d'autres études, seul un étalon interne est utilisé pour tous les composés identifiés en mode positif et un autre pour tous les composés identifiés en mode négatif [64].

La question de la mise en place d'une liste de base d'étalons internes commune à tous les laboratoires (avec ajout d'étalons propres au laboratoire si besoin) se pose aujourd'hui. En effet, une liste minimale commune d'étalons internes permettrait une meilleure inter-comparaison des différents jeux de données [6]. Dans le cas d'analyse HRMS, une proposition d'une liste d'étalons internes testés dans 3 matrices alimentaires a été proposée [73]. Elle regroupe 89 composés très divers en termes de composition élémentaire, de poids moléculaire, de temps de rétention et de modes d'ionisation. Cependant, cette liste contient des composés non marqués, qui peuvent être déjà présents dans les matrices analysées ; cela remet en question leur pertinence.

4.1.4.2 Utilisation des données des étalons internes

L'utilisation des données des étalons marqués utilisés pose encore question. Une normalisation des pics des composés peut être effectuée avec les aires des pics des étalons marqués. Cependant, l'attribution des étalons marqués aux composés pose problème. Il existe plusieurs stratégies d'attribution des étalons marqués utilisées principalement dans le domaine de la métabolomique [74] :

- L'attribution des étalons internes peut se faire en fonction des propriétés physico-chimiques comme le temps de rétention [17], [75], la classe des composés [76] ou le ratio m/z . Cependant, cette normalisation peut ne pas permettre de contrôler entièrement la variation observée [74].
- La stratégie B-MIS (best-matched internal standard) permet de faire correspondre les étalons marqués aux composés qui se comportent de manière similaire pendant l'analyse [77]. Cette méthode est effectuée à partir de mesures répétées d'un échantillon pool, c'est-à-dire un échantillon composé de tous les échantillons à analyser.
- CCMN (Cross-contribution Compensating Multiple standard Normalization) est un modèle se basant sur les effets systématiques de "cross-contribution" [78].

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'application et besoins méthodologiques

- NOMIS (Normalization using Optimal Selection of Multiple Internal Standards) est un modèle se basant sur les similarités dans les variations en utilisant une régression linéaire multiple [79].
- RUV-random (Remove Unwanted Variance-random) [80].

Drotleff et Lämmerhofer [74] effectuent une comparaison de ces stratégies de normalisation et proposent un arbre de décision pour la sélection du type de normalisation. En général, l'inclusion d'échantillons de contrôle de qualité dans les analyses non-ciblées, en respectant un rapport échantillon/contrôle de qualité raisonnable, s'est avérée bénéfique. Ainsi, non seulement les performances de l'instrument sont facilement contrôlées et la normalisation basée sur le contrôle de qualité appliqué, mais l'attribution des étalons internes est également exécutée avec une précision accrue. En résumé, la mise en œuvre générale des contrôles de qualité dans les essais non-ciblés permet un mécanisme de contrôle efficace de l'intégrité des données après normalisation.

Dans certaines études, des étalons internes sont ajoutés aux échantillons sans normalisation des données. Seule une vérification du signal est effectuée avec un suivi de la stabilité du m/z , du temps de rétention et de l'intensité des pics [81].

4.1.5 Analyses d'échantillons de contrôle

D'autres méthodes peuvent être utilisées pour vérifier la stabilité de l'instrument ou pour corriger les signaux.

Une de ces solutions est le recours à un mélange de l'ensemble des échantillons analysés. Ils sont injectés de façon régulière au cours de la séquence d'acquisition, afin d'observer le biais analytique à partir des intensités des différents analytes selon l'ordre de passage [65] et de vérifier la reproductibilité du système par exemple avec des ACP [82]. Cette méthode offre également la possibilité d'intégrer des données provenant de différentes séries analytiques. Cette méthode est largement utilisée dans le domaine de la métabolomique. Ces pools peuvent être enrichis en étalons internes [77].

Un contrôle de la dérive des temps de rétention au cours du temps peut être effectué à partir d'un mélange de composés connus [19].

4.2 LES FORMATS DE DONNEES

Il est nécessaire de garder à l'esprit que les échantillons sont souvent prélevés et analysés pour répondre à une question scientifique précise.

Pour pouvoir échanger les jeux de données ou effectuer de la comparaison d'empreintes chimiques, il est nécessaire d'harmoniser les bases de données. Pour cela, il faut convertir les données brutes obtenues par l'appareil dans un format commun. Les deux formats utilisés sont les formats .mzML et .mzXML [83]. Ce sont des formats open-source largement testés qui peuvent être facilement adaptés aux différentes versions et types des instruments d'analyse. Les données converties en ce format comprennent toutes les informations provenant d'une analyse, y compris les spectres et les métadonnées associées.

Pour les études à long terme ou les études inter-laboratoires, il est essentiel de bien définir les protocoles opératoires et/ou d'utiliser des protocoles normalisés pour assurer la cohérence des données et permettre l'échange de données [59]. Les informations obligatoires pour l'échange de base de données seraient :

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution): exemples d'application et besoins méthodologiques

- La description des échantillons et de la campagne associée : il est nécessaire de donner des informations précises sur l'échantillon (date, lieu, météo, opérateur, ...), la méthode d'échantillonnage et s'il existe des blancs terrains associés.
- La méthode de préparation des échantillons. En effet, l'empreinte chimique d'un échantillon est dépendante de la méthode de préparation des échantillons [4], [82]. Afin de comparer deux empreintes entre elles, il serait nécessaire d'utiliser la même préparation d'échantillon avec par exemple une méthode par matrice [57].
- La méthode d'analyse (appareil utilisé, phase mobile, gradient, mode d'ionisation, etc...).
- Le traitement des données : le traitement du signal est souvent dépendant du logiciel utilisé et des choix effectués par l'opérateur (seuils d'intensité, utilisation des étalons internes...). Ainsi, pour un même échantillon analysé, l'empreinte chimique pourra être différente en termes de pics sélectionnés d'un traitement à l'autre. Pour une comparaison d'empreinte chimique, il serait alors nécessaire de pouvoir avoir accès aux données brutes sans traitement.
- Les informations sur les pratiques d'assurance et de contrôle qualité utilisés [57] avec par exemple les informations sur la dernière calibration de l'appareil et les contrôles qualité utilisés. Ces derniers doivent être communs si possible et clairement définis.

Une approche d'harmonisation des temps de rétention obtenus pour les mêmes composés dans diverses conditions de chromatographie liquide a été proposée [84]. Cela permettrait de comparer les temps de rétention, quelle que soit la méthode d'analyse utilisée. L'harmonisation des temps de rétention est basée sur l'étalonnage des méthodes LC via 18 calibrants dans des indices de temps de rétention prédéfinis (RTI compris entre 1 et 1000). Cette normalisation des temps de rétention est utilisée sur la DSFP NORMAN [19].

5 Conclusion

Un état des lieux non exhaustif d'études et résultats obtenus par traitement non-ciblé des empreintes HRMS a été réalisé dans ce rapport. Cela permet de mettre en évidence les différents moyens entrepris pour chaque étude afin de répondre aux objectifs et les mener à bien.

Les outils de prétraitement, traitement et visualisation des données utilisés dans ces études pour le traitement non-ciblé des empreintes HRMS sont généralement des outils en libre accès. Certains outils sont listés dans ce rapport avec les études associées, afin d'avoir accès aux différents paramètres de réglage. L'opérateur a un large éventail d'outils à sa disposition pour traiter les empreintes HRMS ; son choix doit être guidé par la compatibilité des formats de données, du mode de langage utilisé (R, Python, etc.), du type de traitement des données souhaité (objectifs, prétraitement, visualisation, etc.) et aussi de l'accessibilité des outils en termes de prise en main pour les différents réglages et algorithmes à utiliser.

Bien avant le choix des outils à manipuler pour le traitement non-ciblé des empreintes HRMS, les prérequis analytiques à appliquer sont impératifs pour obtenir des résultats fiables.

Au vu des prérequis analytiques décrits dans ce rapport et bien que de nombreux contrôles qualité aient été mis en place, les empreintes HRMS acquises dans les précédentes études du RSP [1], [2] ne permettraient pas d'obtenir des résultats fiables après des traitements non-ciblés ; les analyses n'étaient pas réalisées en réplicats qui permettent de minimiser le taux de faux positifs et de faux négatifs.

Utilisation des empreintes HRMS (spectrométrie de masse haute résolution) : exemples d'application et besoins méthodologiques

- [1] A. Togola, C. Guillemain, F. Lestremau, C. Coureau, et C. Soulier, « Applicabilité de la technique de « screening non ciblé » pour la surveillance prospective », Réseau de Surveillance prospective-AQUAREF BRGM/RP-70108-FR, 2020.
- [2] A. Togola, F. Lestremau, C. Soulier, et C. Coureau, « Acquisition des données de screening environnemental dans le cadre du réseau de surveillance prospective », Réseau de Surveillance prospective-AQUAREF Rapport BRGM/RP-69967-FR, 2020.
- [3] A. Togola, S. Lardy-Fontan, F. Lestremau, et C. Soulier, « Rapport de positionnement sur l'utilisation de la spectrométrie de masse haute résolution pour le criblage environnemental », Rapport AQUAREF BRGM/RP-65420-FR 75, 2015.
- [4] C. Guillemain *et al.*, « Comparaison et impact de différentes méthodes d'extraction d'eau sur l'analyse non ciblée par chromatographie liquide couplée à la spectrométrie de masse haute résolution », Aquaref - INRAE, Research Report, juill. 2019. Consulté le: 29 novembre 2022. [En ligne]. Disponible sur: <https://hal.archives-ouvertes.fr/hal-03126711>
- [5] A. Togola, F. Lestremau, C. Margoum, V. Dulio, C. Miège, et C. Soulier, « Synthèse des actions européennes et nationales sur les techniques d'analyse chimique non ciblée | AQUAREF - Laboratoire national de référence pour la surveillance des milieux aquatiques », Rapport AQUAREF 2018, 2018.
- [6] A. Togola, S. Lardy-Fontan, C. Soulier, C. Margoum, C. Guillemain, et F. Lestremau, « Critères QA/QC pour le traitement suspect d'analyses non ciblées: note de recommandations », Rapport AQUAREF 2021 – BRGM RP-71775, 2021.
- [7] F. Lestremau et A. Assoumani, « Analyse non-ciblée d'eaux de rejets provenant de la campagne Emergents Nationaux », Réseau de Surveillance prospective-AQUAREF Ineris-203218-2712999-v2.0, 54 p21/12/2021, 2021.
- [8] F. Lestremau, « Evaluation de la faisabilité technique du transfert des données acquises par LC/HRMS vers le Digital Sample Freezing Platform (DSFP) », Réseau de Surveillance prospective-AQUAREF Ineris-203218-v1.0, 28/09/2021, 2021.
- [9] A. Togola, F. Lestremau, C. Coureau, et C. Soulier, « Propositions et mise en œuvre de traitements des données de screening environnemental: action EMNAT-NTS », Réseau de Surveillance prospective-AQUAREF BRGM/RP69968-FR, 2020.
- [10] S. Merel, C. Rigal, T. Ducrocq, C. Miège, R. Singh, et Y. Aminot, « Synthèse bibliographique sur l'analyse non ciblée d'échantillons environnementaux sédimentaires et biologiques », p. 42, 2021.
- [11] B. Schulze *et al.*, « An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples », *TrAC Trends Anal. Chem.*, vol. 133, p. 116063, déc. 2020, doi: 10.1016/j.trac.2020.116063.
- [12] S. Ruppe, D. S. Griesshaber, I. Langlois, H. P. Singer, et J. Mazacek, « Detective Work on the Rhine River in Basel – Finding Pollutants and Polluters: Highlights of Analytical Sciences in Switzerland », *CHIMIA*, vol. 72, n° 7-8, p. 547, août 2018, doi: 10.2533/chimia.2018.547.
- [13] C. Soulier *et al.*, « La spectrométrie de masse haute résolution pour la recherche de micropolluants organiques dans l'environnement: Screening of organic micropollutants in

- environment by high resolution mass spectrometry », *Tech. Sci. Méthodes*, vol. 6, n° 6, p. 43-54, juin 2021, doi: 10.36904/tsm/202106043.
- [14] C. Dietrich, A. Wick, et T. A. Ternes, « Open-source feature detection for non-target LC–MS analytics », *Rapid Commun. Mass Spectrom.*, vol. 36, n° 2, p. e9206, 2022, doi: 10.1002/rcm.9206.
- [15] N. Alygizakis, T. Giannakopoulos, N. S. Thomaidis, et J. Slobodnik, « Detecting the sources of chemicals in the Black Sea using non-target screening and deep learning convolutional neural networks », *Sci. Total Environ.*, vol. 847, p. 157554, nov. 2022, doi: 10.1016/j.scitotenv.2022.157554.
- [16] M. Ruff, M. S. Mueller, M. Loos, et H. P. Singer, « Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry – Identification of unknown sources and compounds », *Water Res.*, vol. 87, p. 145-154, déc. 2015, doi: 10.1016/j.watres.2015.09.017.
- [17] L.-M. Beckers, W. Brack, J. P. Dann, M. Krauss, E. Müller, et T. Schulze, « Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis », *Sci. Total Environ.*, vol. 727, p. 138388, juill. 2020, doi: 10.1016/j.scitotenv.2020.138388.
- [18] M. Krauss, C. Hug, R. Bloch, T. Schulze, et W. Brack, « Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score », *Environ. Sci. Eur.*, vol. 31, n° 1, p. 45, déc. 2019, doi: 10.1186/s12302-019-0231-z.
- [19] N. A. Alygizakis *et al.*, « NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in “digitally frozen” environmental samples », *TrAC Trends Anal. Chem.*, vol. 115, p. 129-137, juin 2019, doi: 10.1016/j.trac.2019.04.008.
- [20] N. A. Alygizakis *et al.*, « Exploring the Potential of a Global Emerging Contaminant Early Warning Network through the Use of Retrospective Suspect Screening with High-Resolution Mass Spectrometry », *Environ. Sci. Technol.*, vol. 52, n° 9, p. 5135-5144, mai 2018, doi: 10.1021/acs.est.8b00365.
- [21] W. Brack *et al.*, « High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources », *Environ. Sci. Eur.*, vol. 31, n° 1, p. 62, sept. 2019, doi: 10.1186/s12302-019-0230-0.
- [22] V. Nikolopoulou, R. Aalizadeh, M.-C. Nika, et N. S. Thomaidis, « TrendProbe: Time profile analysis of emerging contaminants by LC-HRMS non-target screening and deep learning convolutional neural network », *J. Hazard. Mater.*, vol. 428, p. 128194, avr. 2022, doi: 10.1016/j.jhazmat.2021.128194.
- [23] T. Bader, W. Schulz, T. Lucke, W. Seitz, et R. Winzenbacher, « Application of Non-Target Analysis with LC-HRMS for the Monitoring of Raw and Potable Water: Strategy and Results », in *Assessing Transformation Products of Chemicals by Non-Target and Suspect Screening – Strategies and Workflows Volume 2*, vol. 1242, 0 vol., in ACS Symposium Series, no. 1242, vol. 1242., American Chemical Society, 2016, p. 49-70. doi: 10.1021/bk-2016-1242.ch003.

- [24] M. M. Plassmann, E. Tengstrand, K. M. Åberg, et J. P. Benskin, « Non-target time trend screening: a data reduction strategy for detecting emerging contaminants in biological samples », *Anal. Bioanal. Chem.*, vol. 408, n° 16, p. 4203-4208, juin 2016, doi: 10.1007/s00216-016-9563-3.
- [25] W. Dürig, N. A. Alygizakis, F. Menger, O. Golovko, K. Wiberg, et L. Ahrens, « Novel prioritisation strategies for evaluation of temporal trends in archived white-tailed sea eagle muscle tissue in non-target screening », *J. Hazard. Mater.*, vol. 424, p. 127331, févr. 2022, doi: 10.1016/j.jhazmat.2021.127331.
- [26] D. Schütz, E. Achten, M. Creydt, J. Riedl, et M. Fischer, « Non-Targeted LC-MS Metabolomics Approach towards an Authentication of the Geographical Origin of Grain Maize (*Zea mays* L.) Samples », *Foods*, vol. 10, n° 9, Art. n° 9, sept. 2021, doi: 10.3390/foods10092160.
- [27] L. Valderrama, B. Demczuk Jr., P. Valderrama, et E. Carasek, « An Eco-Friendly Proposal by Integrating Chromatographic Fingerprinting and Multivariate Control Chart in a Non-Target Analysis to Evaluate Grape Juices from Different Farming Practices », *J. Braz. Chem. Soc.*, vol. 33, p. 401-405, mars 2022, doi: 10.21577/0103-5053.20210153.
- [28] N. P. Kalogiouri, R. Aalizadeh, et N. S. Thomaidis, « Application of an advanced and wide scope non-target screening workflow with LC-ESI-QTOF-MS and chemometrics for the classification of the Greek olive oil varieties », *Food Chem.*, vol. 256, p. 53-61, août 2018, doi: 10.1016/j.foodchem.2018.02.101.
- [29] B. A. Weggler *et al.*, « Untargeted Identification of Wood Type-Specific Markers in Particulate Matter from Wood Combustion », *Environ. Sci. Technol.*, vol. 50, n° 18, p. 10073-10081, sept. 2016, doi: 10.1021/acs.est.6b01571.
- [30] C. Noblet *et al.*, « Highlight of characteristic residential wood heating and garden green waste burning aerosol molecular markers using non-target screening strategies based on high resolution mass spectrometry combined with multivariate analyses », *Soumis à ES&T*.
- [31] J. E. Schollée, M. Bourgin, U. von Gunten, C. S. McArdell, et J. Hollender, « Non-target screening to trace ozonation transformation products in a wastewater treatment train including different post-treatments », *Water Res.*, vol. 142, p. 267-278, oct. 2018, doi: 10.1016/j.watres.2018.05.045.
- [32] J. Gasperi *et al.*, « Micropollutants in Urban Runoff from Traffic Areas: Target and Non-Target Screening on Four Contrasted Sites », *Water*, vol. 14, n° 3, p. 394, janv. 2022, doi: 10.3390/w14030394.
- [33] J. Hollender, E. L. Schymanski, H. P. Singer, et P. L. Ferguson, « Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? », *Environ. Sci. Technol.*, vol. 51, n° 20, p. 11505-11512, oct. 2017, doi: 10.1021/acs.est.7b02184.
- [34] N. Huynh, « Caractérisation des eaux résiduaires urbaines par spectrométrie de masse haute résolution: influence de la stratégie analytique, limitations et perspectives », 2021.
- [35] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, et G. Siuzdak, « XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification », *Anal. Chem.*, vol. 78, n° 3, p. 779-787, févr. 2006, doi: 10.1021/ac051437y.

- [36] L. L. Hohrenk, F. Itzel, N. Baetz, J. Tuerk, M. Vosough, et T. C. Schmidt, « Comparison of Software Tools for Liquid Chromatography–High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples », *Anal. Chem.*, vol. 92, n° 2, p. 1898-1907, janv. 2020, doi: 10.1021/acs.analchem.9b04095.
- [37] J. Chaker, E. Gilles, T. Léger, B. Jégou, et A. David, « From Metabolomics to HRMS-Based Exposomics: Adapting Peak Picking and Developing Scoring for MS1 Suspect Screening », *Anal. Chem.*, vol. 93, n° 3, p. 1792-1800, janv. 2021, doi: 10.1021/acs.analchem.0c04660.
- [38] M. Gaugain *et al.*, « A non-targeted LC-HRMS approach for detecting exposure to illegal veterinary treatments: The case of cephalosporins in commercial laying Hens », *J. Chromatogr. A*, vol. 1599, p. 161-171, août 2019, doi: 10.1016/j.chroma.2019.04.022.
- [39] G. Libiseller *et al.*, « IPO: a tool for automated optimization of XCMS parameters », *BMC Bioinformatics*, vol. 16, n° 1, p. 118, déc. 2015, doi: 10.1186/s12859-015-0562-8.
- [40] K. Mottershead et T. H. Miller, « Application of deep learning to support peak picking during non-target high resolution mass spectrometry workflows in environmental research », *Environ. Sci. Adv.*, vol. 2, n° 6, p. 877-885, 2023, doi: 10.1039/D3VA00005B.
- [41] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, et S. Neumann, « CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets », *Anal. Chem.*, vol. 84, n° 1, p. 283-289, janv. 2012, doi: 10.1021/ac202450g.
- [42] N. A. Alygizakis, P. Gago-Ferrero, J. Hollender, et N. S. Thomaidis, « Untargeted time-pattern analysis of LC-HRMS data to detect spills and compounds with high fluctuation in influent wastewater », *J. Hazard. Mater.*, vol. 361, p. 19-29, janv. 2019, doi: 10.1016/j.jhazmat.2018.08.073.
- [43] M. Loos, J. Hollender, E. L. Schymanski, M. Ruff, et H. P. Singer, « Bottom-up peak grouping for unknown identification from high-resolution mass spectrometry data », présenté à ASMS 2012 Annual Conference Vancouver, Oral Session Informatics: Identification, 2012.
- [44] M. Loos et H. Singer, « Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data », *J. Cheminformatics*, vol. 9, n° 1, p. 12, déc. 2017, doi: 10.1186/s13321-017-0197-z.
- [45] M. Krauss *et al.*, « Assessing background contamination of sample tubes used in human biomonitoring by non-targeted liquid chromatography–high resolution mass spectrometry », *Environ. Int.*, vol. 183, p. 108426, janv. 2024, doi: 10.1016/j.envint.2024.108426.
- [46] Y. C. Tai, *timecourse*. (2017). Bioconductor. doi: 10.18129/B9.BIOC.TIMECOURSE.
- [47] T. Pluskal, S. Castillo, A. Villar-Briones, et M. Orešič, « MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data », *BMC Bioinformatics*, vol. 11, n° 1, p. 395, déc. 2010, doi: 10.1186/1471-2105-11-395.
- [48] M. Hu, M. Krauss, W. Brack, et T. Schulze, « Optimization of LC-Orbitrap-HRMS acquisition and MZmine 2 data processing for nontarget screening of environmental samples using design of experiments », *Anal. Bioanal. Chem.*, vol. 408, n° 28, p. 7905-7915, nov. 2016, doi: 10.1007/s00216-016-9919-8.

- [49] R. Helmus, T. L. Ter Laak, A. P. Van Wezel, P. De Voogt, et E. L. Schymanski, « patRoom: open source software platform for environmental mass spectrometry based non-target screening », *J. Cheminformatics*, vol. 13, n° 1, p. 1, déc. 2021, doi: 10.1186/s13321-020-00477-w.
- [50] J. Xia, N. Psychogios, N. Young, et D. S. Wishart, « MetaboAnalyst: a web server for metabolomic data analysis and interpretation », *Nucleic Acids Res.*, vol. 37, n° Web Server, p. W652-W660, juill. 2009, doi: 10.1093/nar/gkp356.
- [51] F. Giacomoni *et al.*, « Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics », *Bioinformatics*, vol. 31, n° 9, p. 1493-1495, mai 2015, doi: 10.1093/bioinformatics/btu813.
- [52] H. L. Röst *et al.*, « OpenMS: a flexible open-source software platform for mass spectrometry data analysis », *Nat. Methods*, vol. 13, n° 9, p. 741-748, sept. 2016, doi: 10.1038/nmeth.3959.
- [53] Bloosloos, U. Schmitt, et Dutchjes, *blosloos/enviMass: enviMass version 3.5*. (5 avril 2018). Zenodo. doi: 10.5281/ZENODO.1213098.
- [54] K. Kiefer, L. Du, H. Singer, et J. Hollender, « Identification of LC-HRMS nontarget signals in groundwater after source related prioritization », *Water Res.*, vol. 196, p. 116994, mai 2021, doi: 10.1016/j.watres.2021.116994.
- [55] J. E. Schollée, M. Bourgin, U. Von Gunten, C. S. McArdell, et J. Hollender, « Non-target screening to trace ozonation transformation products in a wastewater treatment train including different post-treatments », *Water Res.*, vol. 142, p. 267-278, oct. 2018, doi: 10.1016/j.watres.2018.05.045.
- [56] C. D. Broeckling, F. A. Afsar, S. Neumann, A. Ben-Hur, et J. E. Prenni, « RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data », *Anal. Chem.*, vol. 86, n° 14, p. 6812-6817, juill. 2014, doi: 10.1021/ac501530d.
- [57] J. Hollender *et al.*, « High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management », *Environ. Sci. Eur.*, vol. 31, n° 1, p. 42, déc. 2019, doi: 10.1186/s12302-019-0225-x.
- [58] M. Bouhifd *et al.*, « Quality Assurance of Metabolomics », p. 9, 2015.
- [59] W. B. Dunn *et al.*, « Quality assurance and quality control processes: summary of a metabolomics community questionnaire », *Metabolomics*, vol. 13, n° 5, p. 50, mai 2017, doi: 10.1007/s11306-017-1188-9.
- [60] D. Broadhurst *et al.*, « Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies », *Metabolomics*, vol. 14, n° 6, p. 72, 2018, doi: 10.1007/s11306-018-1367-3.
- [61] D. Dudzik, C. Barbas-Bernardos, A. García, et C. Barbas, « Quality assurance procedures for mass spectrometry untargeted metabolomics. a review », *J. Pharm. Biomed. Anal.*, vol. 147, p. 149-173, janv. 2018, doi: 10.1016/j.jpba.2017.07.044.

- [62] R. D. Beger *et al.*, « Towards Quality Assurance and Quality Control in Untargeted Metabolomics Studies », p. 7, 2020.
- [63] J. A. Kirwan *et al.*, « Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management », *Metabolomics*, vol. 18, n° 9, p. 70, août 2022, doi: 10.1007/s11306-022-01926-3.
- [64] B. Ng, N. Quinete, et P. R. Gardinali, « Assessing accuracy, precision and selectivity using quality controls for non-targeted analysis », *Sci. Total Environ.*, vol. 713, p. 136568, avr. 2020, doi: 10.1016/j.scitotenv.2020.136568.
- [65] J. Cotton, « Application de l'analyse métabolomique à la détection ciblée et globale de contaminants organiques dans des matrices agroalimentaires et environnementales par spectrométrie de masse à ultra-haute résolution », Université Pierre et Marie Curie - Paris VI, 2015.
- [66] S. Ríos Peces *et al.*, « Untargeted LC-HRMS-Based Metabolomics for Searching New Biomarkers of Pancreatic Ductal Adenocarcinoma: A Pilot Study », *SLAS Discov.*, vol. 22, n° 4, p. 348-359, avr. 2017, doi: 10.1177/1087057116671490.
- [67] T. Bader, W. Schulz, K. Kümmerer, et R. Winzenbacher, « General strategies to increase the repeatability in non-target screening by liquid chromatography-high resolution mass spectrometry », *Anal. Chim. Acta*, vol. 935, p. 173-186, sept. 2016, doi: 10.1016/j.aca.2016.06.030.
- [68] E. J. Want *et al.*, « Global metabolic profiling of animal and human tissues via UPLC-MS », *Nat. Protoc.*, vol. 8, n° 1, Art. n° 1, janv. 2013, doi: 10.1038/nprot.2012.135.
- [69] T. Bader, W. Schulz, K. Kümmerer, et R. Winzenbacher, « LC-HRMS Data Processing Strategy for Reliable Sample Comparison Exemplified by the Assessment of Water Treatment Processes », *Anal. Chem.*, vol. 89, n° 24, p. 13219-13226, déc. 2017, doi: 10.1021/acs.analchem.7b03037.
- [70] V. Hinnenkamp, P. Balsaa, et T. C. Schmidt, « Target, suspect and non-target screening analysis from wastewater treatment plant effluents to drinking water using collision cross section values as additional identification criterion », *Anal. Bioanal. Chem.*, vol. 414, n° 1, p. 425-438, janv. 2022, doi: 10.1007/s00216-021-03263-1.
- [71] B. Lalere, V. Le Diuron, E. Alasonati, et S. Lardy-Fontan, « Les blancs analytiques : définition et recommandations quant à leur utilisation. », 2018.
- [72] E. Eysseric, F. Beaudry, C. Gagnon, et P. A. Segura, « Non-targeted screening of trace organic contaminants in surface waters by a multi-tool approach based on combinatorial analysis of tandem mass spectra and open access databases », *Talanta*, vol. 230, p. 122293, août 2021, doi: 10.1016/j.talanta.2021.122293.
- [73] A. M. Knolhoff, J. H. Premo, et C. M. Fisher, « A Proposed Quality Control Standard Mixture and Its Uses for Evaluating Nontargeted and Suspect Screening LC/HR-MS Method Performance », *Anal. Chem.*, vol. 93, n° 3, p. 1596-1603, janv. 2021, doi: 10.1021/acs.analchem.0c04036.

- [74] B. Drotleff et M. Lämmerhofer, « Guidelines for Selection of Internal Standard-Based Normalization Strategies in Untargeted Lipidomic Profiling by LC-HR-MS/MS », *Anal. Chem.*, vol. 91, n° 15, p. 9836-9843, août 2019, doi: 10.1021/acs.analchem.9b01505.
- [75] S. Bijlsma *et al.*, « Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation », *Anal. Chem.*, vol. 78, n° 2, p. 567-574, janv. 2006, doi: 10.1021/ac051495j.
- [76] T. Cajka, J. T. Smilowitz, et O. Fiehn, « Validating Quantitative Untargeted Lipidomics Across Nine Liquid Chromatography-High-Resolution Mass Spectrometry Platforms », *Anal. Chem.*, vol. 89, n° 22, p. 12360-12368, nov. 2017, doi: 10.1021/acs.analchem.7b03404.
- [77] A. K. Boysen, K. R. Heal, L. T. Carlson, et A. E. Ingalls, « Best-Matched Internal Standard Normalization in Liquid Chromatography–Mass Spectrometry Metabolomics Applied to Environmental Samples », *Anal. Chem.*, vol. 90, n° 2, p. 1363-1369, janv. 2018, doi: 10.1021/acs.analchem.7b04400.
- [78] H. Redestig *et al.*, « Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data », *Anal. Chem.*, vol. 81, n° 19, p. 7974-7980, oct. 2009, doi: 10.1021/ac901143w.
- [79] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, et M. Orešič, « Normalization method for metabolomics data using optimal selection of multiple internal standards », *BMC Bioinformatics*, vol. 8, n° 1, p. 93, mars 2007, doi: 10.1186/1471-2105-8-93.
- [80] A. M. De Livera *et al.*, « Statistical methods for handling unwanted variation in metabolomics data », *Anal. Chem.*, vol. 87, n° 7, p. 3606-3615, avr. 2015, doi: 10.1021/ac502439y.
- [81] M. Liu *et al.*, « Chemical fingerprinting of organic micropollutants in different industrial treated wastewater effluents and their effluent-receiving river », *Sci. Total Environ.*, vol. 838, p. 156399, sept. 2022, doi: 10.1016/j.scitotenv.2022.156399.
- [82] N. Huynh, E. Caupos, C. Soares Peirera, J. Le Roux, A. Bressy, et R. Moilleron, « Evaluation of Sample Preparation Methods for Non-Target Screening of Organic Micropollutants in Urban Waters Using High-Resolution Mass Spectrometry », *Molecules*, vol. 26, n° 23, p. 7064, nov. 2021, doi: 10.3390/molecules26237064.
- [83] L. Martens *et al.*, « mzML—a Community Standard for Mass Spectrometry Data », *Mol. Cell. Proteomics MCP*, vol. 10, n° 1, p. R110.000133, janv. 2011, doi: 10.1074/mcp.R110.000133.
- [84] R. Aalizadeh *et al.*, « Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening », *Anal. Chem.*, vol. 93, n° 33, p. 11601-11611, août 2021, doi: 10.1021/acs.analchem.1c02348.



**RÉPUBLIQUE
FRANÇAISE**

*Liberté
Égalité
Fraternité*

Centre scientifique et technique

3, avenue Claude-Guillemin

BP 36009

45060 – Orléans Cedex 2 – France

Tél. : 02 38 64 34 34

Direction régionale ou UTAM

Adresse

Tél. :

www.brgm.fr



Géosciences pour une Terre durable

brgm