Caractérisation moléculaire d'échantillons organiques complexes par spectrométrie de masse et chromatographie en phase liquide

Cédric Wolters

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Terre, Univers, Environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

« Cédric WOLTERS »

Thèse dirigée par **Véronique VUITTON** et
co-encadrée par **François-Régis ORTHOUS-DAUNAY**

préparée au sein du **Laboratoire de Planétologie et d'Astrophysique de Grenoble**
dans **l'École Doctorale Terre, Univers, Environnement**

Caractérisation moléculaire d'échantillons organiques complexes par spectrométrie de masse et chromatographie en phase liquide

Thèse soutenue publiquement le **26 Février 2021**,
devant le jury composé de :

M, Didier, Voisin
Professeur, Univ. Grenoble Alpes, Président du jury
M, Uwe, Meierhenrich
Professeur, Univ. De Nice, Rapporteur
Mme, Marie-Claire, Gazeau
Professeur, Univ. Paris-Est-Creteil, Rapporteur
M, Carlos, Afonso
Professeur, Univ. De Rouen, Membre
Mme, Claude, Geoffroy
Maître de Conférence, Univ. De Poitiers, Membre

Thèse réalisée au

Laboratoire de Planétologie et
d'Astrophysique de Grenoble
UMR5274 - CNRS/UGA
414 Rue de la Piscine
Bâtiment PhitemD
38400 Saint Martin d'Hères

Web : ipag.osug.fr

Sous la direction de

Véronique VUITTON, directrice de thèse
François-Régis ORTHOUS-DAUNAY,
co-encadrant de thèse
Michel VISO, référent CNES

Financement

Allocation de recherche du CNES
Allocation de recherche de l'ANR, projet RAHIIA-SSOM porté par Grégoire Danger, Univ. Aix-Marseille, PIIM

Caractérisation moléculaire d’échantillons organiques complexes par spectrométrie de masse et chromatographie en phase liquide

Résumé

Comment analyser un échantillon organique complexe ? Cette question générale semble simple de prime abord, mais requiert de s'intéresser de plus près à la notion de complexité afin de pouvoir comprendre et justifier les moyens utilisés pour la caractériser. En planétologie, et plus largement en astrophysique, l'ensemble des observations et observables indiquent que la matière qui compose les objets extraterrestres est composée d'un mélange de diverses molécules, et ce mélange est plus ou moins divers en fonction de l'objet. Observations et modélisations sont effectuées couramment pour tenter de comprendre ces objets et de contraindre leurs processus évolutifs. Caractériser la complexité moléculaire de tels objets nécessite des instruments de pointe, qui sont difficilement adaptables aux contraintes spatiales pour être placés sur une sonde, et cela requiert que l'objet étudié puisse être échantillonné. Or, la très grande majorité des objets d'intérêt ne peuvent pas être atteints dans un temps raisonnable. Dès lors, il faut un autre moyen d'étudier ces objets : c'est l'astrophysique de laboratoire. De nombreuses expériences tentent de simuler les objets et environnements dans lequel ils évoluent, et analysent l'évolution de la matière soumise à ces contraintes. Une partie des défis de ces expériences réside dans la caractérisation chimique des échantillons, et plus particulièrement dans leur caractérisation moléculaire. Dans le cadre de cette thèse, nous proposons d'utiliser la spectrométrie de masse haute résolution (HRMS) et la chromatographie en phase liquide haute performance (HPLC) pour caractériser des échantillons organiques complexes. Pour se faire, l'ensemble de la chaine analytique a été étudiée, depuis l'acquisition des données jusqu'à leur exploitation. Ainsi, nous proposons une optimisation de l'acquisition des données en Orbitrap, ainsi que des systématiques de traitement des données issues des analyses effectuées ESI-HRMS ainsi que pour des analyses effectuées en LDI-ICR. La chromatographie couplée à la spectrométrie de masse est un outil puissant pour accéder à la structure moléculaire des échantillons, et nécessite de développer des méthodes qui soient adaptés aux échantillons analysés. Nous proposons ainsi deux méthodes HPLC pour l'analyse des échantillons, qui ont été développées et validées pour l'analyse d'échantillons complexes. Cependant, aucun logiciel commercial ne permet l'analyse non supervisée de tels échantillons : un logiciel qui permette le traitement de ces données a ainsi été développé et permet de révéler la diversité moléculaire des échantillons sans supervision. Mais l'identification des molécules ainsi détectées n'est pas un processus aisé puisqu'il nécessite alors de posséder l'ensemble des isomères possibles pour chaque molécule détectée. Pour réduire cet espace des possibles, un outil de prédiction des temps de rétention est proposé qui se base sur la connaissance des propriétés physico-chimiques de composés connus afin de prédire, pour ces mêmes composés, leur temps de rétention théorique sur les méthodes utilisées. Ce travail présente dans une dernière partie l'application de l'ensemble des développements effectués au cours de ces trois années sur un jeu d'échantillons d'analogue d'aérosols atmosphériques de synthèse modélisant des exoplanètes de type super-Terres et mini-Neptunes. Depuis l'analyse de leur matière soluble, jusqu'à la comparaison entre phase soluble, insoluble et totale, l'analyse par spectrométrie de masse indique une grande diversité et des différences importantes entre échantillons, indiquant des processus de formation et d'évolution directement liés à la composition du mélange réactif. Enfin, l'analyse par chromatographie d'un de ces échantillons indique de multiples isomères, dont certains pouvant être annotés comme étant des molécules biologiques, potentiellement impliquées dans le processus de l'origine de la vie.

Molecular characterization of complex organic samples by mass spectrometry and liquid chromatography

Abstract

How to analyse a complex organic sample? This general question seems simple at first glance but requires a closer look at the notion of complexity to be able to understand and justify the means used to characterise it. In planetology, and more widely in astrophysics, all the observations and observables indicate that the matter that makes up extraterrestrial objects is composed of a mixture of various molecules, and this mixture is more or less diverse and dense depending on the object. Observations and models are routinely done to try to understand these objects and to constrain their evolutionary processes, or to try to investigate their origin. Characterising the molecular complexity of such objects requires state-of-the-art instruments, which are difficult to adapt to space industry constraints in order to be placed on a probe, and this requires that the object under study can be sampled. However, most objects of interest cannot be reached in a reasonable time. Therefore, another way to study these objects is needed: laboratory astrophysics. Many experiments attempt to simulate the objects and environments in which they evolve and analyse the evolution of matter subjected to these constraints. Part of the challenges of these experiments lies in the chemical characterisation of the samples, and more particularly in their molecular characterisation. As part of this thesis, we proposed to use high-resolution mass spectrometry (HRMS) and high-performance liquid chromatography (HPLC) to characterise complex organic samples. To do so, the entire analytical chain was studied, from the data acquisition to its use. Thus, we proposed an optimisation of the data acquisition in Orbitrap, as well as the systematic processing of the data resulting from the analysis done by ESI-HRMS as well as for the analysis done by LDI-ICR. Chromatography coupled with mass spectrometry is a powerful tool for accessing the molecular structure of samples and requires developing methods that are suited to the samples analysed. Therefore we offered two HPLC methods for sample analysis, which have been developed and validated for the analysis of complex samples. However, no currently available commercial software allowed for the unsupervised analysis of such samples. Software to allow the processing of this data has now been developed and allows the molecular diversity of samples to be revealed without supervision. The identification of the detected molecules is not an easy process since it then requires having all the possible isomers for each molecule detected as standards for reference. To reduce the number of possibilities, a tool for predicting retention times was proposed. This was based on knowledge of the physico-chemical properties of known compounds to predict their theoretical retention time on the methods used. Lastly, this work presents the application of all the developments carried out during these three years on a set of samples of synthetic atmospheric aerosol analogues modelling exoplanets of the super-Earth and mini-Neptunes type. From the analysis of their soluble matter to the comparison between soluble, insoluble, and total phase, analysis by mass spectrometry indicates a great diversity and important differences between samples. This indicates processes of formtion and evolution related to the composition of the reactive mixture. Finally, chromatographic analysis of one of these samples indicates multiple isomers, some of which may be labelled as biological molecules, potentially involved in the process of the origin of life.

Table des matières

Index des figures

Index des tableaux

Index des équations

0. Introduction

Qu'est-ce qu'un échantillon organique complexe ? Si l'on cherche à transformer ces termes plutôt généraux en termes plus précis, on pourrait plutôt se demander ce que signifie un « mélange moléculaire complexe ». Cette formulation ne remplace que les deux premiers termes, l'idée de la complexité étant toujours présente. Ainsi, il convient de chercher à définir et délimiter individuellement les différents termes afin d'avoir une idée précise de ce que dont parle.

Qu'est-ce que la complexité ? Ce terme n'est pas trivial et fait appel à de nombreux aspects qui ne sont pas forcément évidents. Nous nous sommes tous, à un moment ou à un autre, heurté à ce mur de verre : nous sommes capables de donner un sens à ce mot, mais la capacité à en faire le tour demande bien plus d'efforts que de définir n'importe quel autre concept scientifique. Premièrement, la complexité n'a pas de barrière, c'est-à-dire qu'elle est présente à tous les niveaux et non restreinte uniquement en ce que certains appellent les « sciences dures » : la description des interactions humaine peut très certainement être décrite comme complexe par exemple. Ainsi, la séparation artificielle qui est créée quelques fois serait plus un défaut de compréhension plutôt qu'une séparation profonde des différents univers de connaissance, comme présenté par Richard Feynman, prix Nobel de Physique, extrait de [1] :

I have a friend who's an artist and has sometimes taken a view which I don't agree with very well. He'll hold up a flower and say "look how beautiful it is," and I'll agree. Then he says "I as an artist can see how beautiful this is but you as a scientist take this all apart and it becomes a dull thing," and I think that he's kind of nutty. First of all, the beauty that he sees is available to other people and to me too, I believe. I can appreciate the beauty of a flower. At the same time, I see much more about the flower than he sees. I could imagine the cells in there, the complicated actions inside, which also have a beauty. I mean it's not just beauty at this dimension, at one centimeter; there's also beauty at smaller dimensions, the inner structure, also the processes. The fact that the colors in the flower evolved in order to attract insects to pollinate it is interesting; it means that insects can see the color. It adds a question: does this aesthetic sense also exist in the lower forms? Why is it aesthetic? All kinds of interesting questions which the science knowledge only adds to the excitement, the mystery and the awe of a flower. It only adds. I don't understand how it subtracts.

Plusieurs points ressortent de cette description : la complexité a plusieurs niveaux qui forment un tout interconnecté. On commence également à apercevoir ce qui relève des composants et ce qui relève des processus, comme énoncé par Nancy Hinman [1] :

Complexity is the nexus between parts and processes. If the parts are divided, then the processes won't occur. And conversely if the processes are occurring then the parts cannot be dismantled

---

La complexité serait donc la description d'un système dans son ensemble, incluant toutes ces parties et processus de fonctionnement. Chaque composant peut interagir localement avec d'autres composants de façon très simple, définissant une étape du processus. Mais décrire cette étape simple ne permet pas de décrire le système dans son ensemble, puisque c'est l'ensemble de ces étapes qui permet de définir le système. Si l'on considère un être vivant, système complexe que chacun d'entre nous peut appréhender, est-ce que décrire le cycle de Krebs, processus indispensable pour le métabolisme, est suffisant pour décrire un être vivant ? Probablement non. Un être vivant, qu'il soit animal, végétal ou autre, est bien plus qu'un cycle de Krebs, et la complexité à décrire précisément ce qu'est un être vivant est bien là : quelques processus et composants, alors briques élémentaires du système, contribuent à créer l'architecture globale, mais ne sont pas suffisantes en elles-mêmes pour reproduire l'architecture complète. A l'inverse, sans ces quelques briques élémentaires, cette même architecture n'est pas complète. L'interdépendance des processus et éléments est la clef de voûte de la description de la complexité.

Si l'on revient à notre question initiale, on se doit désormais de définir le terme « moléculaire ». Ce terme fait appel directement à des notions de chimie, où la molécule est déjà un système complexe d'atomes qui forment une structure stable. La description des composants et processus qui forment une molécule est complexe et fait appel à des notions de physique quantique qui ne sont pas l'objet de ce travail. Cependant, décrire le terme « moléculaire » en se réduisant à décrire physiquement ce qu'est une molécule est probablement incomplet. On se doit également de considérer l'ensemble des réactions et processus chimiques qui permettent d'agir sur les molécules, les changeant en d'autres molécules. Une molécule, constituée d'atomes, est ainsi la résultante de multiples étapes qui ont permis d'assembler ces atomes en la molécule étudiée. Et pour ajouter à la complexité, les différentes étapes qui permettent d'obtenir une molécule n'est surement qu'un chemin parmi tant d'autres permettant de créer la même molécule. Dès lors, associer le terme moléculaire avec le terme complexe fait sens puisque qu'il n'est pas suffisant de décrire une molécule pour en comprendre son origine et son évolution.

La diversité des informations et processus nécessaires à la description complète des systèmes requiert des espaces possédant les dimensions supérieures à celles que nous sommes capable de représenter, à savoir l'espace en trois dimensions ou, plus fréquemment encore, l'espace en deux dimensions des rapports et articles scientifiques. Dès lors, il est nécessaire de réduire les dimensions en simplifiant, projetant ou en ne décrivant qu'une partie des processus et composant pour pouvoir discuter et interpréter le système. Ceci est un processus courant qui permet de pouvoir diffuser la connaissance, tout en ayant conscience que cette représentation simplifiée ne représente pas forcément de façon fidèle le système complexe étudié.

Enfin, qu'est-ce qu'un mélange ? Un mélange est l'opposé de pur, c'est-à-dire que des composants différents sont présents dans un même système. Un mélange peut également être homogène ou non-homogène, mais aussi être présent sous différentes formes physiques (i.e. solide, liquide ou gazeux). Un mélange peut être simple ou complexe : simple quand interchanger n'importe quelle molécule avec une autre ne change en rien la description globale du système, et complexe quand certaines combinaisons donnent des résultats différents à l'échelle globale. Néanmoins, la simplicité ou la complexité apparente ne sont-elles pas dépendantes de la propriété globale observée ? Par exemple, si l'on considère la composition isotopique du xénon dans l'air terrestre, le mélange est simple : interchanger n'importe quelle molécule n'a aucun effet sur les propriétés globales. Cependant, il s'avère que ~20% de ce xénon proviendrait des comètes [2] et que donc ce mélange serait nécessairement complexe puisqu'il fait intervenir deux sources différentes qui ont été mélangées intiment depuis leur mise en contact. De plus, on peut recréer la composition isotopique du xénon dans l'air terrestre en prenant 20% d'air pauvre en xénon lourd et 80% d'air enrichi en xénon lourd, et le résultat

de ce mélange « synthétique » sera indifférenciable du mélange « naturel ». L'effet du mélange a donc tendance à diluer, voire même à effacer, les effets de source lorsque l'on s'intéresse à des propriétés ponctuelles. La seule façon de contrer cet effet de mélange serait de remonter le temps, ce qui est impossible : le mélange est la seule chose que nous avons à disposition, et c'est à nous, scientifiques, de mettre en place les protocoles et mesures nécessaires pour faire en sorte de pouvoir aller chercher dans les mélanges les propriétés caractéristiques uniques permettant de pouvoir discuter de leur origine ou de leur évolution par exemple.

Dès lors, caractériser un mélange moléculaire complexe revient à étudier ses caractéristiques dans son ensemble dans le but de pouvoir comparer et classifier ces mélanges entre eux, et ainsi peut être de pouvoir discuter de leur origine, de leur évolution ou encore de leurs processus et composants caractéristiques. La définition des moyens analytiques à mettre en place est alors dépendante des échantillons et de la problématique scientifique à laquelle on souhaite répondre.

0.1. La complexité par excellence : la vie

La vie est un processus qui, selon l'état actuel des connaissances, n'est observable que sur Terre. Ce processus est plutôt ancien (~3 à 4 milliards d'années) et est un système complexe dans le sens où il est impossible de se contenter d'un unique composé ou processus pour décrire l'ensemble du système. Par exemple, lister et quantifier l'ensemble des composés qui forment une pomme ne permet pas de produire une pomme puisque mélanger exactement ses composés à partir des substances pures ne permet pas d'obtenir une pomme. Seuls des processus biochimiques permettent, à partir de composés bien définis, de créer une pomme, et selon un unique chemin. Ainsi, la complexité intrinsèque du mélange, du fait de son processus d'évolution et de création est ce qui caractérise la pomme, et par extension, la vie dans son ensemble.

L'ensemble des processus à l'origine de la vie ne sont pas connus, seuls les processus qui permettent sa reproduction sont accessibles. Cette question de (des ?) l'origine(s) est le sujet de nombreuses recherches et hypothèses et il existe un paradoxe important dans cette question : « avoir les briques ne fait pas la maison, mais sans brique, aucune maison n'est possible ». Dès lors, de nombreuses recherches se focalisent sur la recherche de briques caractéristiques, les « biomarqueurs », qui seraient à eux seuls ce qui permet de détecter et confirmer la vie telle que nous la connaissons.

Mais est-ce que la vie peut-elle être aussi simplement réduite à un ensemble limité de briques élémentaires qu'il suffirait de détecter quelque part pour se dire que l'on a découvert de la vie ailleurs ? Gerald Joyce a défini en 1992 pour la NASA la définition de la vie, encore utilisée à ce jour :

La vie est un système chimique auto-entretenu capable de subir une évolution darwinienne.

L'avantage de cette définition est que l'on réduit le système « vie » à un mélange moléculaire complexe qui fonctionne hors équilibre et qui est capable d'adaptation. Dès lors, il suffirait de trouver les briques élémentaires qui caractérisent le système chimique « vie », celles qui permettent de fonctionner hors équilibre et celles responsables de l'adaptation. Et c'est ici que se situe un problème majeur : (1) on sait lister un grand nombre de molécules faisant partie de la vie (plus de 113 000 composés dans la base de données HMDB (https://hmdb.ca/) par exemple), (2) on sait définir ce qui fait que la vie fonctionne hors équilibre, par exemple, la compartimentation cellulaire, et (3) on sait définir comme la vie s'adapte, par des erreurs lors de la reproduction. Mais peut-on définir une liste restreinte d'observable à partir de ces vastes

définitions ? Est-il possible de s'affranchir de la caractérisation moléculaire pour pouvoir définir la vie ?

Bartlett et Wong [3], en continuité avec des études antérieures de la NASA et notamment de leur définition de la vie, ont proposé une sorte de définition unifiée de la « vye », concept où la « vie » telle que nous la connaissons est une partie de la « vye » en plus de toutes les « vies » extra-terrestres inconnues, comme présenté en Figure 1. Cette « vye » s'articule autour de quatre grands piliers : la dissipation, l'autocatalyse, l'homéostasie et l'apprentissage. Il est aisé de concevoir de multiples exemples qui regroupent quelques-unes de ces définitions, des « sous-vye », mais seule la « vye » regroupe les quatre à la fois. La complexité des systèmes vivants est incomparable à cet égard de n'importe quel autre processus non-vivant, abiotique, du fait par exemple de sa capacité à pouvoir favoriser des chemins réactionnels produisant seulement des formes énantiomères pures. Il n'existe aucun processus chimique totalement abiotique permettant de créer autant de molécules énantiomères pures, et de pouvoir les recycler à l'infini en une boucle synergique incluant l'ensemble de la vie.

La complexité du système vie nous empêche de définir proprement de grands observables qui nous permettraient de pouvoir détecter la vie. Mais cette même complexité nous empêche également de remonter à nos propres origines, le chemin évolutif de la vie telle que nous la connaissons nous étant caché du fait de multiples chemins possibles. Ainsi, en l'absence de clefs universelles, il convient de caractériser globalement les échantillons analysés, mais également de s'intéresser à quelques briques constitutives que l'on pourrait trouver intéressantes en se basant sur la connaissance de la vie telle que nous la connaissons. Effectuer seulement la caractérisation de quelques briques, et la vision du système est perdue ; il est primordial de conserver le lien entre le système et ses composants, au risque de ne pouvoir classifier et comparer des échantillons entre eux.

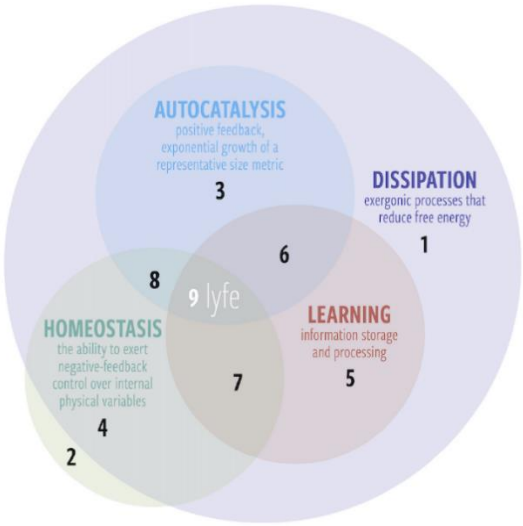


Figure 1 – Diagramme de Venn des quatre piliers de la « vye ». Les « sous-vye » (régions 1-8) sont n'importe quel système qui présentent quelques-uns des piliers, quand seulement la « vye » les présente tous. Extrait de [3]

0.2. Objectifs de cette thèse

Cette thèse n'a pas pour ambition de rechercher et fournir une quelconque théorie sur l'origine de la vie, son évolution, ou contraindre n'importe quel observable à cet effet. Les différents objectifs de ce travail sont plutôt de définir, développer et/ou retravailler des méthodes instrumentales et de traitement de données dans le but de pouvoir extraire le plus d'informations possible des données issues des instruments analytiques. En effet, analyser la matière organique complexe d'échantillons nécessite des techniques et méthodes adaptées à la diversité présente dans de tels échantillons. Or, ces méthodes et techniques ne sont pas nécessairement courantes ou alors nécessitent des ajustements pour être adaptées aux

spécificités (ex : polarité et caractère acido-basique de chaque molécule) présentes dans de tels échantillons. Ainsi, pour la caractérisation moléculaire de tels échantillons, seuls des techniques de spectrométrie de masse couplées à des techniques chromatographiques sont capables de fournir des informations sur la diversité stœchiométrique et moléculaire constitutive des échantillons organiques complexes.

Des analyses en spectrométrie de masse Orbitrap sont effectuées depuis 2008 à l'institut de planétologie et d'astrophysique de Grenoble (IPAG). Cet instrument possède des performances résolutives de 100 000 à *m/z*=400 Da, pour une précision inférieure à 2 ppm. Bien que cet instrument soit utilisé depuis plusieurs années, de nombreuses interrogations subsistent : est-ce que les protocoles d'acquisitions utilisés sont optimaux ? est-ce que l'on est capable de définir une systématique d'attribution qui permette de valider les attributions sur une gamme de masse étendue ? Bien que les caractéristiques de l'Orbitrap suffisent pour décrire ces analyses comme étant effectuées à haute résolution, des appareils encore plus résolutifs et précis existent : les *Spectrométrie de masse à résonance cyclonique ionique* (FT-ICR), présentant des performances résolutives supérieures au million à *m/z*=200 Da pour une précision de quelques centaines de ppb. Les FT-ICR possèdent également la capacité d'analyser les différentes phases des échantillons via des analyses *Désorption-Ionisation Laser* (LDI), option qui n'est pas disponible avec le LTQ Orbitrap XL. On se pose alors la question de la pertinence des données de notre Orbitrap comparés à ces instruments : quid de la résolution spectrale quant aux échantillons complexes ? est-ce qu'analyser la phase soluble uniquement a un sens ?

La spectrométrie de masse ne permettant d'obtenir que les informations stœchiométriques des échantillons, il convient d'ajouter une méthode de chromatographie en amont pour permettre d'ajouter une dimension supplémentaire, qui nous permet alors de pouvoir potentiellement remonter à la structure des molécules. Cependant, les méthodes chromatographiques non ciblées pour analyses des échantillons complexes sont très spécifiques et peu documentées. Quelles sont les colonnes et conditions chromatographiques qui pourraient être adaptées à nos échantillons ? Est-il possible de définir une systématique de développement de méthode chromatographique d'intérêt pour du personnel non-spécialiste afin de lui permettre d'utiliser et d'adapter les méthodes chromatographiques à ses échantillons ? Le traitement des données issues du couplage entre une chromatographie et un spectromètre de masse n'est pas trivial : est-ce que des logiciels sont déjà disponibles ? quelles seraient les étapes pour en développer un ? est-ce que le traitement chromatographique est à minima comparable à tout ou partie des données issues de l'infusion directe ?

Du fait de l'utilisation d'échantillons complexes, on s'attend à une diversité de plusieurs milliers de formules chimiques, possédant plusieurs centaines d'isomères : est-ce qu'il est possible d'observer des isomères avec les méthodes sélectionnées ? Est-ce que la diversité observée est comparable ou incluse dans les données disponibles en infusion directe ? Est-ce que ces composés ont leur formule brute identique à des composés listés dans des bases de données ? Peut-on estimer la diversité moléculaire probable à partir d'une comparaison à une base de données ? Quid de l'identification, normalement basée sur l'analyse individuelle de standards ? Il est physiquement impossible de posséder l'ensemble des standards purs pour identifier les signaux observés. Néanmoins, l'identification n'est que l'étape ultime de la caractérisation moléculaire : est-il possible d'effectuer des annotations, et à quel degré de confiance se placer ? Est-il possible d'établir et de réduire la liste des composés probables en se basant sur l'annotation ?

Pour tenter de répondre à l'ensemble de ces questions, le travail est partitionné en cinq chapitres :

(1) Astrophysique de laboratoire et analyses moléculaires ;

Les objets d'étude en planétologie et astrophysique sont physiquement éloignés et presque impossibles à échantillonner. Il est alors utile de chercher à modéliser ces objets pour pouvoir analyser leur composition en laboratoire avec des techniques différentes de ce qui est réalisable avec des observations astronomiques. En fonction des objets étudiés, divers types de synthèses existent et quelques grands résultats sont résumés ici. Pour étudier ces échantillons, seule la spectrométrie de masse et son couplage à la chromatographie permettent d'accéder à la complexité moléculaire. Ainsi, on présente le principe de fonctionnement des spectromètres de masse, ainsi que le principe de fonctionnement de la chromatographie.

(2) Développement de méthodes de mesures ;
Les méthodes de mesures sont la clef de voûte permettant de s'assurer de la qualité des résultats fournis et leur optimisation est cruciale dans ce processus, et ces méthodes s'incluent alors dans une chaine analytique globale d'analyse des échantillons. Dans le cadre de cette chaine analytique, l'ensemble des maillons relatifs à l'analyse moléculaire se doit d'être optimal. Ainsi, les procédures d'acquisition en Orbitrap ont fait l'objet d'une étude systématique pour déterminer la meilleure combinaison possible en ce qui concerne les scans et micro-scans, particularité de ce modèle d'Orbitrap. Une fois les données acquises en ESI-Spectrométrie de masse, le processus de traitement des données, et plus particulièrement leur attribution, est effectué. Du fait de la précision et de la résolution limitée de l'instrument, il est nécessaire de s'assurer que les attributions effectuées soient cohérentes avec la nature polymérique de l'échantillon. Pour ce faire, un protocole systématique de validation et nettoyage des formules stœchiométrique est proposé et utilisé pour l'ensemble des attributions effectuées avec ce type de données. Comme l'ESI-Orbitrap ne permet que les mesures en phase soluble, un projet d'analyse en LDI-FT-ICR a été réalisé au laboratoire COBRA à Rouen dans le but de comparer la phase soluble, insoluble et totale. La source LDI, au contraire de la source ESI, produit les ions radicalaires en plus des ions moléculaires. Dès lors, la systématique de validation proposée précédemment doit être ajustée en conséquence et ce travail d'ajustement est présenté à cette occasion. Enfin, du fait du couplage avec la chromatographie, il apparait nécessaire de devoir effectuer des identifications de molécules. Traditionnellement, cela nécessite des standards purs qui sont injectés individuellement. Du fait de la multitude de signaux et isomères, cela reviendrait à posséder des milliers de standards purs. On propose alors une méthode de prédiction des temps de rétention qui permet de réduire la liste des candidats possibles, avec la possibilité par la suite, si nécessaire, de confirmer ces prédictions par l'injection d'un nombre limité de molécules pures.

(3) Développement d'un logiciel de traitement des données issues de la chromatographie ;
Traiter les données issues du couplage entre la spectrométrie de masse et la chromatographie est un défi du fait de la taille des données à traiter et de l'objectif à atteindre. En effet, traiter plusieurs millions de points dans le but d'extraire l'ensemble des signaux nécessite de faire des choix de traitements qui ont un impact sur les algorithmes à utiliser ainsi que sur les résultats finaux. Quelques logiciels libres d'accès sont disponibles, mais présentent des défauts sur leur configuration et également sur leur traitement des données en masse. Ainsi, on présente dans cette partie le développement d'un logiciel de traitement des données issues du couplage entre spectrométrie de masse haute résolution et chromatographie, ainsi que les choix algorithmiques et de traitement des données effectués dans le but d'extraire le maximum d'informations exploitables possible.

(4) Application à des échantillons issus d'expériences d'astrophysiques de laboratoire ;

L'ensemble des développements proposés aboutissent à leur utilisation pour des échantillons organiques complexes d'intérêt pour la planétologie. Dans le cas de ce travail, quelques échantillons d'analogues d'aérosols d'atmosphère d'exoplanètes sont analysés avec les outils développés précédemment. Depuis l'analyse en infusion directe en ESI-Orbitrap seul pour observer la diversité et les différences entre échantillons, jusqu'à la comparaison entre Orbitrap et ICR, ainsi que l'analyse comparative des différentes phases, les analyses stœchiométriques sont une étape nécessaire à elle seule puisqu'elle permet d'apporter un regard global sur la diversité moléculaire à l'échelle de la composition stœchiométrique. Par la suite, le couplage avec la chromatographie permet d'ajouter la dimension temporelle à cette diversité, et ainsi de résoudre les éventuels isomères qui possèdent les propriétés physico-chimiques en adéquation avec la colonne et les conditions chromatographiques. Les résultats de l'analyse en chromatographie sont alors comparés aux analyses en infusion directe ainsi qu'à une base de données de composés biochimiques dans le but de déterminer si des composés séparés peuvent être d'intérêt pour des discussions astrophysiques.

(5) Perspectives futures – Ouverture, améliorations de la méthodologie pour application à d'autres échantillons naturels et synthétiques (ex : application aux données Hayabusa2)
L'ensemble de la systématique développée n'a de sens que si elle est appliquée à de multiples échantillons de source et de nature différentes, naturels et synthétiques, dans le but de construire un ensemble de données. L'intérêt d'un tel jeu de données peut être multiple en fonction des informations souhaitées, et peut aller de la simple recherche ciblée de composés et isomères communs ou uniques à l'ensemble des échantillons, jusqu'à des comparaisons entre matière naturelle et synthétique dans le but d'apporter des hypothèses concernant les contraintes de formation et/ou d'évolution de la matière naturelle à partir des contraintes connues lors de la génération de la matière synthétique. La mise en place initiale de cet axe de travail est proposée dans un projet LabEx porté par quelques chercheurs de l'équipe Planéto où un poste d'ingénieur de recherche est demandé à cet effet. En complément de ce projet est l'implication de l'équipe Planéto, et plus particulièrement du groupe Orbitrap, dans l'analyse des données issues de la mission Hayabusa 2, où l'IPAG a en charge l'analyse des données issues de la chromatographie, et où le logiciel de traitement des données développé lors de cette thèse sera utilisé pour analyser ces données, en interaction avec l'ensemble des autres groupes travaillant sur la matière soluble.

1. Astrophysique de laboratoire et analyses moléculaires

En planétologie et astrophysique, la disponibilité à l'échantillonnage des objets étudiés est limitée voire impossible. Dès lors, il est nécessaire, pour compléter les données issues des observations astrophysiques, d'effectuer des simulations en laboratoire dans le but de modéliser ces objets et analyser les échantillons qui en découlent. Du fait de la diversité des objets, plusieurs expériences d'astrophysique de laboratoire sont effectuées par divers groupes à travers le monde. Les échantillons peuvent être analysés sur n'importe quel instrument dans le but de les caractériser le plus finement possible. Dans le cadre de ces travaux, et comme discuté précédemment, accéder à la complexité moléculaire d'échantillons organiques complexes requiert l'utilisation de spectromètres de masse et de chromatographie.

Dans ce chapitre, après quelques généralités sur les observations et modélisations en astrophysique, nous introduisons les expériences d'astrophysique de laboratoire, ainsi que quelques appareils et résultats associés. On discutera également rapidement du choix d'échantillons analysés dans le cadre de ces travaux. Puis, la spectrométrie de masse de type Orbitrap est décrite et discutée, ainsi que la source d'ionisation de type électrospray qui lui est associée. Ensuite, pour s'assurer de la validité des données Orbitrap, des analyses FT-ICR ont été menées dans le cadre d'un projet avec le laboratoire COBRA à Rouen. Dès lors, le principe de fonctionnement des FT-ICR est présenté et discuté vis-à-vis de l'Orbitrap. En complément, des analyses en ionisation laser ont été dans le but d'analyser non seulement la phase soluble, mais également la phase insoluble et la phase totale. Il sera également introduit quelques notions générales de traitement des données issues de la spectrométrie de masse, notamment *via* l'utilisation du logiciel Attributor. Par la suite, la chromatographie sera introduite et les différentes grandeurs caractéristiques d'intérêt définies.

1.1. Observations et simulations

1.1.1. Diversité moléculaire dans les milieux astrophysiques

Observer l'environnement est un processus naturel effectué par l'homme dès que possible, lui permettant d'obtenir des informations et éventuellement de s'adapter aux contraintes ainsi identifiées. En astrophysique, observer les grands objets célestes est effectué par divers instruments à travers le monde, mais également via des missions spatiales par l'intermédiaires de sondes et rovers. Toutes les données accumulées nous indiquent que les mélanges moléculaires plus ou moins complexes sont partout, depuis les nuages moléculaires dans l'espace interstellaire jusqu'aux atmosphères d'exoplanètes en passant par les disques protoplanétaires et les astéroïdes, comme schématisé en Figure 2. En fonction des techniques utilisées, seuls quelques molécules simples sont alors détectées et des indices laissent alors penser à la présence de mélanges et structures plus complexes, comme par exemple la présence d'aérosols et de nuages pour certaines exoplanètes.

Les données d'observations de l'espace interstellaire permettent par exemple d'établir une liste de plusieurs centaines de composés ayant par exemple jusqu'à quelques carbones, hydrogènes, oxygènes et/ou azotes [4]. Plus proches de la Terre, d'autres données d'observations montrent également une diversité importante dans l'atmosphère de Titan, visité par la sonde Cassini-Huygens [5] ou encore sur l'astéroïde 67P Churyumov-Gerasimenko visité par la sonde Rosetta [6]. Quelques représentations illustrant la complexité de l'atmosphère de Titan sont présentées en Figure 3. Même si l'ensemble des données disponibles indiquent une diversité importante sur tous les objets observés, tous les systèmes ne présentent pas la même complexité apparente, la complexité d'une étoile étant différente de cette d'un nuage moléculaire, étant également différente de la complexité d'une atmosphère ou d'un astéroïde.

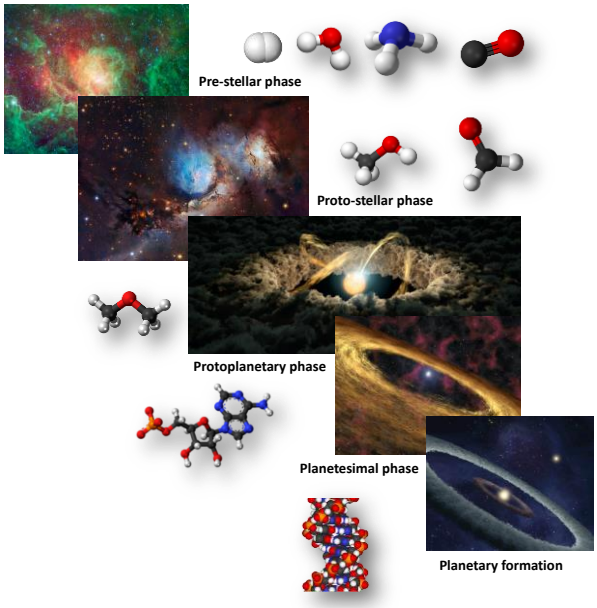


Figure 2 – Représentation schématique de la création d'un système planétaire avec son évolution chimique associée. Adapté à partir de [7]

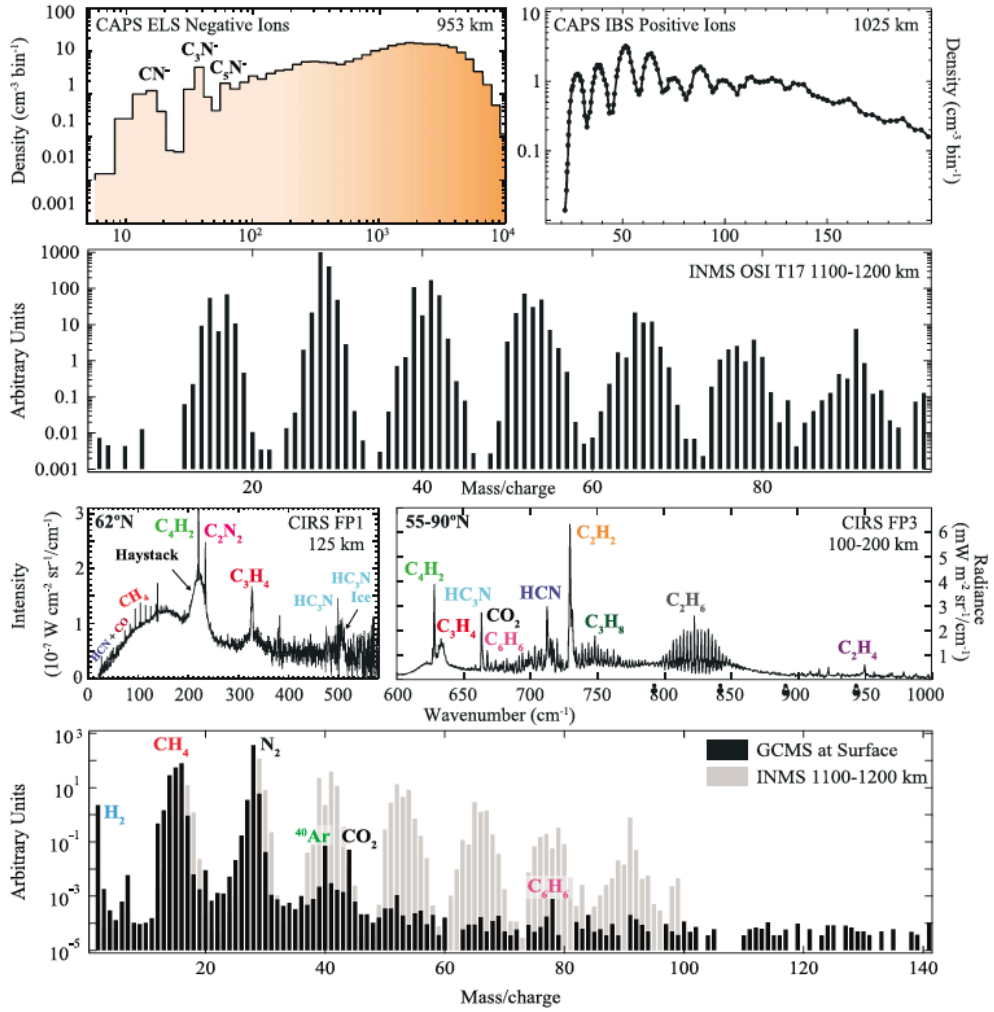


Figure 3 – Quelques représentations de mesures in-situ et à distance de la composition de l'atmosphère de Titan par la sonde Cassini-Huygens. Extrait de [5]

1.1.2. Modélisation

La description globale de chaque objet nous permet de les caractériser et de les classer afin de potentiellement décrire leur formation ou leur évolution. Ce processus descriptif est souvent supporté par des expériences de modélisations, en laboratoire ou numérique. Les simulations numériques peuvent, par exemple, présenter un ensemble de réactions ayant lieu dans un espace donné, et alors générer une liste complète de molécules que l'on doit s'attendre d'observer. Ce genre de travail de simulation a par exemple été effectué par Véronique Vuitton et al. pour Titan [8] où un ensemble de réactions chimiques est décrit. Cet ensemble de réaction établit une liste de composés et abondances auxquelles on peut s'attendre. Or, une liste de composés et abondances n'est autre qu'une liste de masses exactes associées à des intensités, soit directement des spectres de masses qu'il est possible de comparer à des données issues d'analyses en spectrométrie de masse provenant d'observations et analyses réelles, comme présenté en Figure 4 concernant des analyses dans l'atmosphère de Titan issues de la mission Cassini-Huygens.

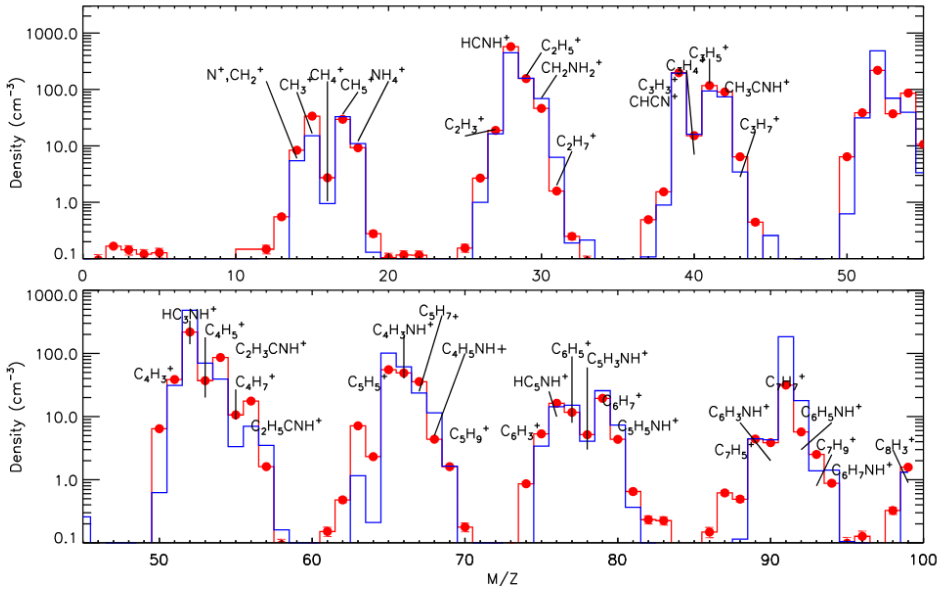


Figure 4 – Comparaison entre le spectre de masse observé (en bleu) et le spectre de masse reconstitué à partir de simulations numériques (en rouge). Extrait de [8]

1.1.3. Astrophysique de laboratoire

Les observations et modélisations, décrites précédemment, permettent d'avoir une idée de la composition des objets observés. Cependant, même si les informations accessibles de cette manière sont importantes, elles ne permettent pas d'accéder à la complexité moléculaire des objets. Ainsi, on peut utiliser ces informations générales issues des observations et modélisations pour contraindre des expériences de simulation en laboratoire : c'est ce que l'on peut appeler « l'astrophysique de laboratoire ».

Ces expériences ont pour principal intérêt de simuler un environnement astrophysique simplifié et d'observer son évolution en fonction des conditions appliquées. La variété et la complexité des simulations effectuées reposent souvent sur la complexité du mélange initial et des conditions appliquées par la suite : ces expériences permettent de proposer de nouvelles hypothèses expliquant l'apparente déplétion en azote des comètes [9], de décrire l'évolution de la matière organique complexe soumise à un flux UV intense [10], ou encore de proposer des processus de synthèse abiotiques de composés d'intérêt pour la vie [11]. Pour toutes les disciplines, il convient de se rappeler que toute simulation est par essence incorrecte, mais que c'est souvent la seule façon de pouvoir étudier des problèmes. Ainsi, c'est uniquement par la combinaison entre expériences de laboratoire et observations des objets que des hypothèses et

des discussions supportées par les expériences de laboratoire peuvent être utilisées. Ainsi, ces expériences diverses et variées ne sont en aucun cas une simulation fidèle de l'objet et des processus chimiques en action dans le système observé, mais permettent de proposer des hypothèses sur les contraintes et processus qui pourraient se dérouler dans les systèmes ainsi simulés.

Classiquement, deux grands types d'expériences d'astrophysique de laboratoire sont effectuées : des expériences en phase solide et des expériences en phase gazeuse. Ces expériences, du fait de la différence de phase, ne simulent pas les mêmes environnements et ne sont donc pas comparables. Trois types d'expériences de laboratoire sont discutées à titre d'illustration par la suite : la simulation de résidus de glaces interstellaires par Grégoire Danger [12–15] du PIIM à Marseille, la simulation d'aérosols pour l'atmosphère de Titan et de Pluton par Nathalie Carrasco [16–18] du LATMOS à Paris, et une autre expérience de simulation d'aérosols pour l'atmosphère de Titan et d'exoplanètes par Sarah Hörst de l'université Johns Hopkins dans le Maryland (Etats-Unis).

1.1.3.1. *Simulation de glaces interstellaires*

Un mélange est nébulisé sur une fenêtre refroidie à 78K, et se dépose alors sous forme de couches successives. En même temps que cette déposition, la fenêtre est irradiée avec une lampe au deutérium, qui produit une lumière monochromatique dans l'UV. Le temps de synthèse est de l'ordre de 72h, et la chambre est alors réchauffée doucement jusqu'à température ambiante. Lors de ce processus, la glace s'évapore et laisse un résidu organique qui peut être récupéré et analysé par diverses techniques analytiques. Un schéma simplifié du dispositif est disponible en Figure 5. Le mélange peut être plus ou moins complexe, mais est classiquement effectué en utilisant un mélange d'eau, de méthanol et d'ammoniaque. L'influence du ratio entre les différents mélanges sur la composition résultante a fait l'objet d'une étude dans la thèse d'Aurélien Fresneau [14], avec des données acquises sur l'Orbitrap à l'IPAG.

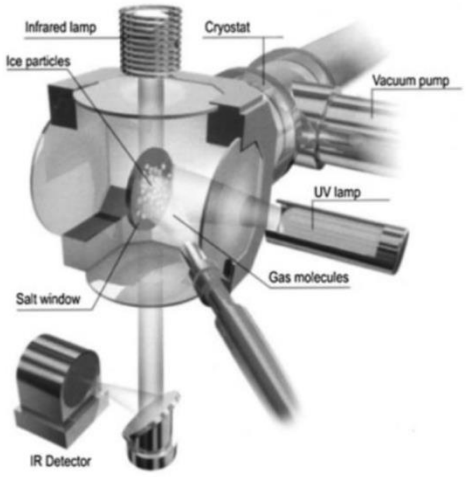


Figure 5 - Schéma de la synthèse de glaces interstellaires. Extrait de [14]

Ce type de glaces a fait l'objet de multiples autres études qui ont révélées par exemple la présence de sucres [19,20] et acides aminés [21] a été révélée par chromatographie en phase gazeuse à deux dimensions, comme illustré en Figure 6 pour quelques sucres. La puissance d'identification de la chromatographie en phase gazeuse provient de son couplage en spectrométrie de masse avec une source à impact électronique. En effet, la source à impact électronique permet de casser les molécules et d'analyser les fragments qui sont produits. Ces fragments et leur intensité relative sont l'empreinte digitale de la molécule, et il est alors possible d'effectuer des identifications directement. Néanmoins, si plusieurs molécules arrivent

en même temps, cette identification n'est plus aussi facilement réalisable. C'est pourquoi la puissance résolutive de la chromatographie en phase gazeuse en deux dimensions est critique ici puisque cela permet d'effectivement séparer les composés qui co-éluent sur la première dimension en utilisant une seconde dimension orthogonale à la première. D'autres analyses ont été effectuées, comme par exemple des analyses en spectroscopie qui indiquent la présence de fonctions d'esters, amides et acides carboxyliques [12].

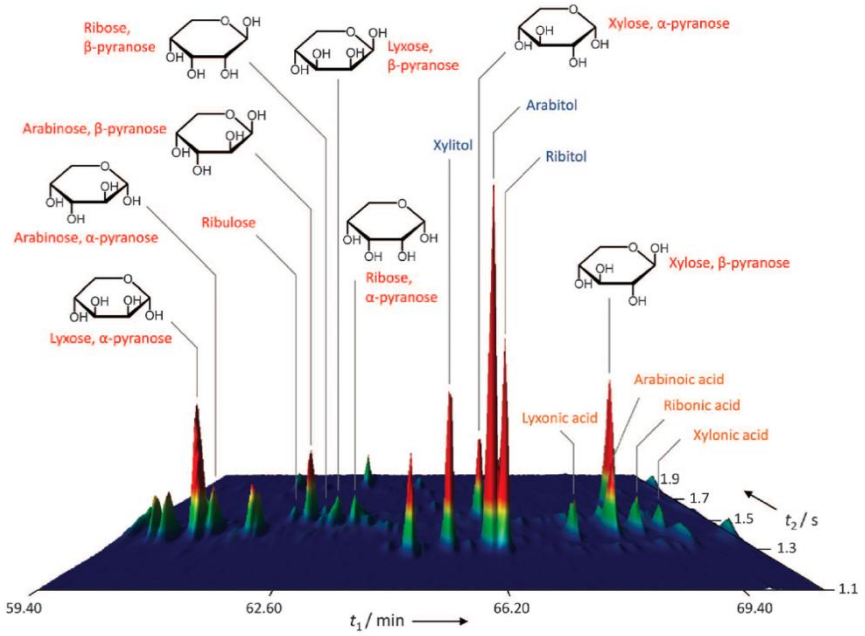


Figure 6 – Chromatogramme en GC-2D de quelques sucres détectés dans l'analyse d'analogues de glaces interstellaires. Extrait de [20].

1.1.3.2. *Simulation d'aérosols pour l'atmosphère de Titan et de Pluton*

Un mélange gazeux est injecté dans la chambre réactionnelle dans laquelle est appliquée un champ radiofréquence. En plus, les particules sont maintenues en lévitation par l'application d'un champ électrique, grossissant en microgravité, et tombent au fond de la chambre lorsque l'interaction du champ avec la particule devient instable. L'ensemble des particules déposées au fond de la chambre peuvent alors être récupérées et utilisées dans diverses analyses. Un schéma de principe de ce système, nommé PAMPRE, est présenté en Figure 7. Diverses compositions de gaz peuvent être utilisées, et ce dispositif a été majoritairement utilisé pour synthétiser des analogues d'aérosols pour l'atmosphère de Titan [16,17,22,23], et plus récemment pour l'atmosphère de Pluton [18].

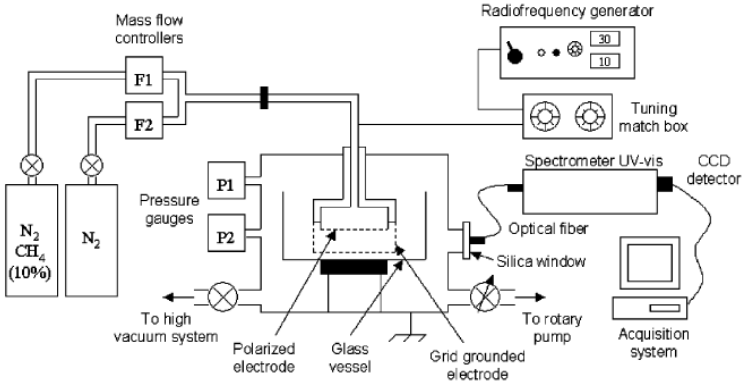


Figure 7 – Schéma de principe de l'expérience PAMPRE. Extrait de [16].

De nombreuses analyses ont été effectuées sur ce type d'échantillon, comme par exemple des acides aminés [23] révélés par ESI-Orbitrap puis confirmés par la suite par GC-MS. Une analyse comparative de la phase soluble et insoluble a également été effectuée en FT-ICR et révèle une différence en termes de composition élémentaire et d'insaturation entre la phase soluble et insoluble, comme illustré en Figure 8. Ces travaux indiquent aussi un faible taux de recouvrement entre phase insoluble et soluble en ce qui concerne les attributions stœchiométriques [17].

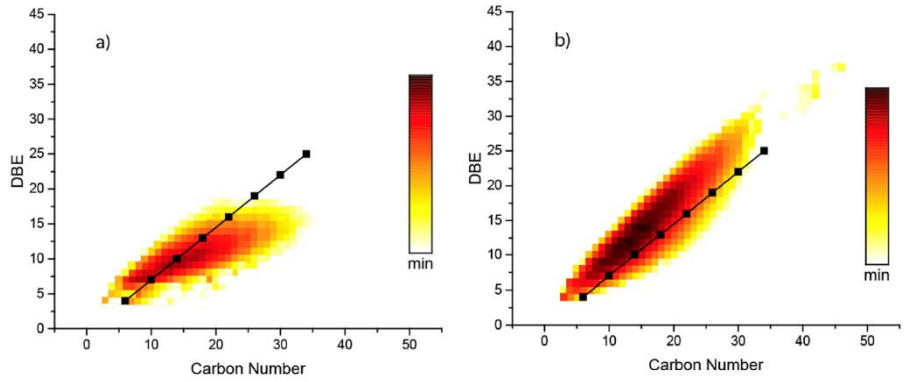


Figure 8 – Comparaison du degré d'insaturation de la phase soluble (a) et insoluble (b) analysée par LDI-ICR pour les espèces possédant entre 6 et 9 azotes. Extrait de [17].

1.1.3.3. *Simulation d'aérosols pour l'atmosphère de Titan et d'exoplanètes*

Un mélange gazeux est injecté dans la chambre réactionnelle dans laquelle est appliquée une décharge électrostatique. À la différence de l'expérience du LATMOS présentée précédemment, cette décharge produit un plasma où le gaz est introduit à une température contrôlable et dans lequel les réactions se produisent. Les particules produites sont projetées sur l'ensemble des parois de la chambre et sont récupérées par la suite. Un schéma synthétique de la chambre de réaction, PHAZER, est présenté en Figure 9. Diverses compositions de gaz peuvent être utilisées, et ce dispositif a été utilisé pour synthétiser des analogues d'aérosols pour l'atmosphère de Titan [24] et récemment pour des atmosphères d'exoplanètes [25,25–27].

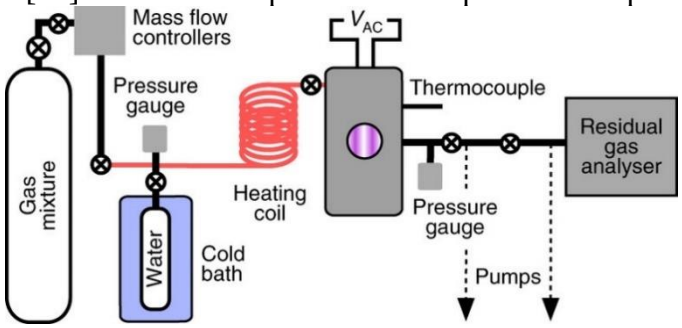


Figure 9 – Représentation schématique de la chambre PHAZER. La partie centrale, entre le « heating coil » et le thermocouple est là où se situe la production d'échantillon. Extrait de [25].

Les analogues d'atmosphères d'exoplanètes sont très récents, et les analyses effectuées sont peu nombreuses. Des analyses non destructives telles que des observations en microscopie à force atomique [25] (AFM) ont été par exemple effectuées et ont mis en évidence le couplage entre la composition initiale du gaz réactif et des processus de croissance des grains, comme illustré en Figure 10. L'analyse de ces mêmes échantillons a également indiqué une variation importante de la couleur des échantillons [25], indiquant *a priori* des variations d'insaturations importantes entre échantillons, et donc une diversité moléculaire différente. Des analyses ESI-Orbitrap ont également été effectuées et confirment la diversité de compositions moléculaire entre échantillons, avec des échantillons qui sont par exemple très solubles dans le méthanol et d'autres qui ne le sont absolument pas. Ces mêmes analyses indiquent de potentielles molécules

d'intérêt biochimiques dans ces échantillons, qui doivent cependant être confirmées par des analyses complémentaires permettant d'effectuer des identifications [26].

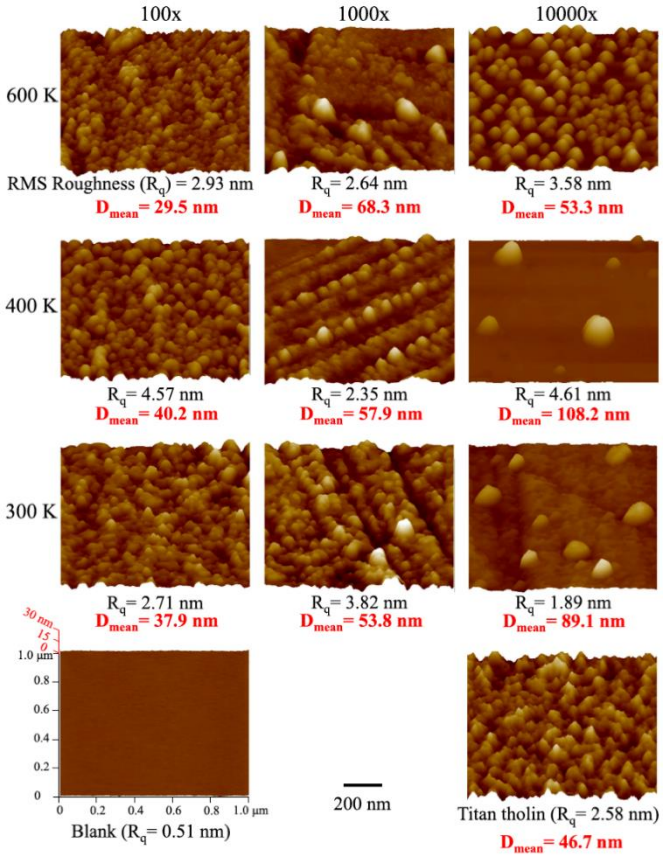


Figure 10 – Analyses en AFM des diverses expériences effectuées, où la taille moyenne des grains ainsi que la dispersion observée sont calculées pour chaque échantillon. On note alors une variation importante des tailles de grains entre différentes expériences, mettant en évidence des processus de croissance différents en fonction de la composition initiale du mélange réactif. Extrait de [25].

1.2. Concept de chaine analytique et description complète d'un échantillon

Caractériser un échantillon au laboratoire nécessite souvent de multiples analyses effectuées sur plusieurs instruments. En effet, chaque technique ne permet d'accéder qu'à une petite partie de l'information contenue dans l'échantillon, souvent du fait des étapes indispensable à sa préparation mais également du fait des limites instrumentales inhérentes à chaque technique. Ainsi, dans le but de caractériser un échantillon de la façon la plus complète possible, il est nécessaire de définir une succession d'analyses et d'étapes de préparation qui permettent d'allier au mieux la quantité d'échantillons disponibles aux quantités nécessaires pour chaque analyse.

Un grand nombre de techniques analytiques de pointe requiert des systèmes couteux en énergie et en espace physique occupé. Chromatographie haute performance, spectrométrie de masse haute résolution, résonance magnétique nucléaire, spectroscopie infrarouge à transformée de Fourier, microscopie de force atomique ou encore microscopie électronique sont autant de techniques qui sont capables de fournir des informations précieuses du moment que vous avez accès physiquement à l'échantillon. Dans le domaine spatial, nombre d'objets d'intérêts ne sont pas accessibles à l'échantillonnage direct : astéroïdes, surface et atmosphères planétaires ou encore comètes sont autant d'objets d'intérêts qui sont majoritairement hors de portée. Quelques missions telles que Hayabusa 1&2, Mars Sample Return ou encore Osiris Rex ont été, sont ou vont être effectuées pour rapporter des échantillons de quelques objets. Mais cela ne concerne que des objets proches, dans notre système solaire et reste négligeable face à la quantité d'objet à explorer dans notre système solaire et au-delà.

On peut différencier directement deux types d'analyses : les analyses destructives et les analyses non-destructives. Ainsi, un parcours analytique simple inclus l'ensemble des techniques non-destructives avant toute technique destructive. C'est ce qui est, par exemple, effectué pour les analyses d'échantillons disponibles en très petites quantités, tels que les échantillons issus de missions spatiales de retour d'échantillons par les processus de curation (Hayabusa 1 et 2 ou StarDust) où un ensemble d'étapes d'observations non destructives sont effectuées de façon systématique pour classifier et décrire physiquement les échantillons avant une distribution aux équipes dans les laboratoires pour un ensemble d'analyses complémentaires. La classification basée sur la destructivité d'une technique d'analyse n'est plus suffisante et doit alors inclure l'ensemble du processus analytique pour définir une chaine analytique complète permettant de rationnaliser au maximum consommation d'échantillon et information extraite. Quelques étapes d'un processus qui pourrait être appliqué à l'analyse d'un échantillon inconnu sont présentées en Figure 11 à titre d'illustration.

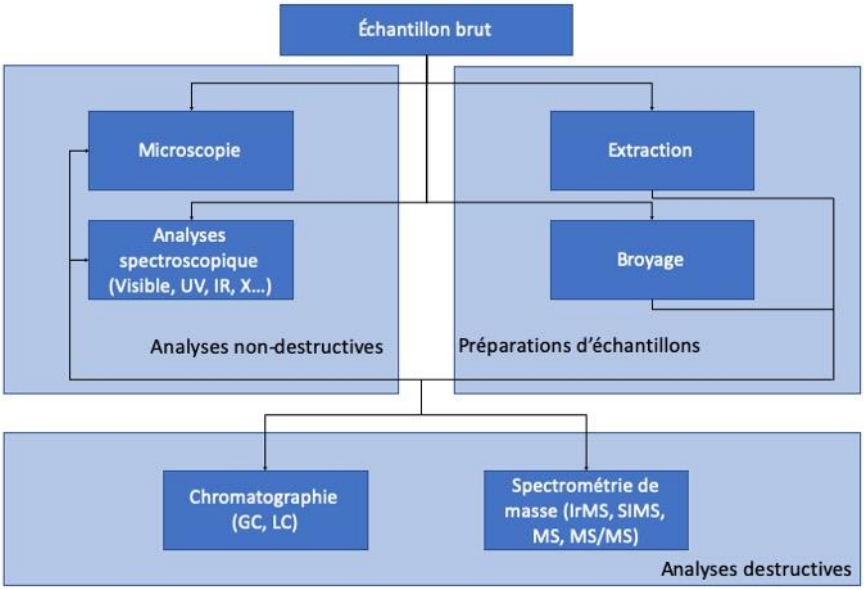


Figure 11 – Illustration de quelques étapes d'un processus analytique visant à extraire le maximum d'informations d'un échantillon à caractériser.

Dans certains cas, des analyses nécessitent les informations fournies par d'autres analyses afin de pouvoir sélectionner des conditions expérimentales compatibles avec l'échantillon. Par exemple, la chromatographie nécessite de choisir une chimie de colonne, cette dernière permet de mieux séparer certaines familles que d'autres. Ainsi, il est important d'avoir une idée des familles en présence pour adapter au mieux les conditions chromatographiques, information qui peut être fournie par de la spectroscopie ou des analyses en spectrométrie de masse par exemple. Dès lors, un processus d'analyse n'est pas une série d'expériences et d'analyses indépendantes les unes des autres, mais est bien un processus intégré où chaque partie apporte de l'information d'intérêt nécessaire pour effectuer les autres analyses ainsi que pour leur interprétation.

Si l'on se concentre plus spécifiquement sur la caractérisation moléculaire de la chaine analytique, l'information en masse permet d'ouvrir une fenêtre sur la complexité de l'objet observé. En effet, une masse exacte représente parfaitement une seule et unique formule stœchiométrique, pour peu que la résolution et la précision de l'instrument soit suffisante. Cependant, une formule stœchiométrique ne permet pas de remonter à une molécule, mais à un ensemble de molécules ayant des structures différentes : des isomères. Pour illustrer ce problème, on présente en Figure 12 un spectre de masse présentant la formule stœchiométrique $C_6H_{14}N_4O_2$ et pour lequel on représente trois isomères parmi les 125 possibles. Ces trois

isomères ont des structures et fonctions très différentes, l'un étant une brique de base pour la vie, l'autre un poison et le dernier une sorte de carburant.

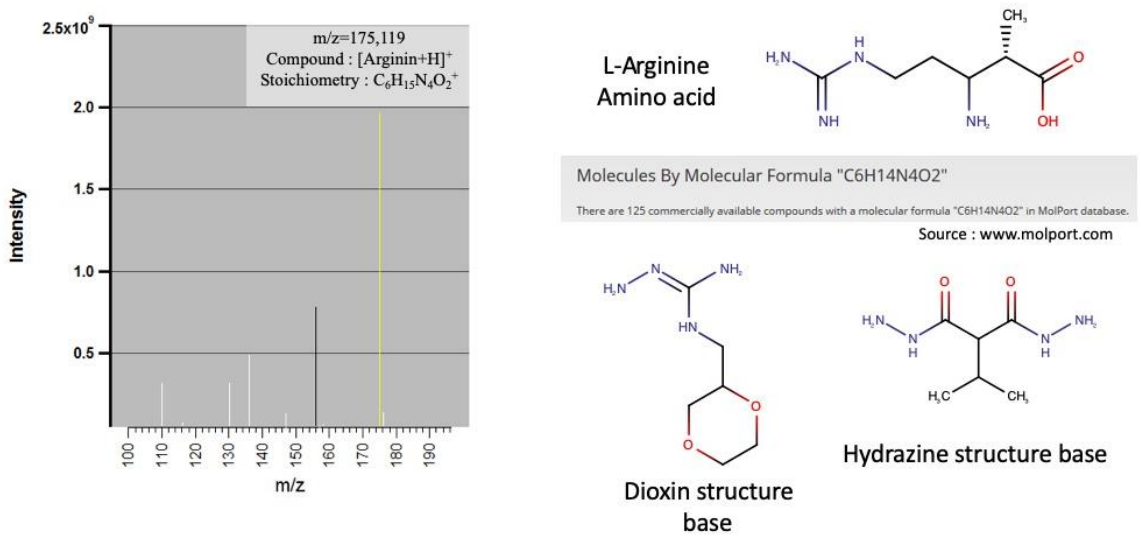


Figure 12 – Illustration du problème masse-isomères avec la formule stœchiométrique $C_6H_{14}N_4O_2$: représentation de trois isomères parmi 125 possibles.

Cette perte de dimension entre la réalité et le passage en masse est un problème, problème qui ne peut être uniquement résolu que par l'ajout d'une dimension orthogonale à la spectrométrie de masse : la chromatographie. Ainsi, l'utilisation de la spectrométrie de masse couplée à la chromatographie permet d'accéder à la caractérisation moléculaire d'échantillons complexes, et seulement ces techniques le permettent. En effet, les autres techniques de science analytique ne permettent d'accéder qu'au plus à la fonction des molécules (*i.e.* toute la spectroscopie) ou à des caractéristiques globales de l'échantillon (exemple : caractérisation physique par microscopie, constantes optiques, mesures physico-chimiques) qui ne fournissent pas d'informations directes et univoques sur les caractéristiques moléculaires de l'échantillon.

1.3. Caractérisations chimiques : du fonctionnel au moléculaire

Une chaine analytique nécessite d'être construite de manière logique, pour obtenir de l'information générale sur l'échantillon avant d'aller sur de l'information de plus en plus spécifique. Les étapes de description physique de l'échantillon ne sont pas discutées dans ces travaux. En ce qui concerne la caractérisation chimique d'un échantillon, on s'intéresse d'abord aux informations fonctionnelles avant d'aller sur la description moléculaire des échantillons.

Les informations fonctionnelles sont des informations sur les fonctions chimiques principales portées par les molécules présentes dans l'échantillon. Ce type d'information est souvent fourni par des analyses spectroscopiques telles que des analyses en infra-rouge. Acides carboxyliques, nitriles ou cétones voire tout simplement s'il y a de la matière carbonée par la présence de liaisons C-H sont autant d'informations importantes qui établissent le tableau initial de l'échantillon pour aller chercher ensuite les spécificités de l'échantillon en se basant sur ces informations initiales pour guider les techniques et méthodes à utiliser si nécessaire.

1.3.1. La spectrométrie : stratégies analytiques

Les premiers spectres de masses de l'histoire ont été obtenus à partir de composés gazeux simples tels que N_2, CO, CO_2 et O_2. Ces spectres ont été effectués par Joseph John Thomson en 1912 en utilisant un dispositif expérimental faisant appel à la fois aux champs magnétiques et aux champs électrostatiques. De nos jours, les spectromètres de masse sont multiples et possèdent des résolutions variant de quelques milliers à plusieurs millions. Néanmoins, ils sont

toujours basés sur la même série de principe, comme présenté en Figure 13 : ioniser les molécules, les placer sur une trajectoire connue et analyser leur réponse.

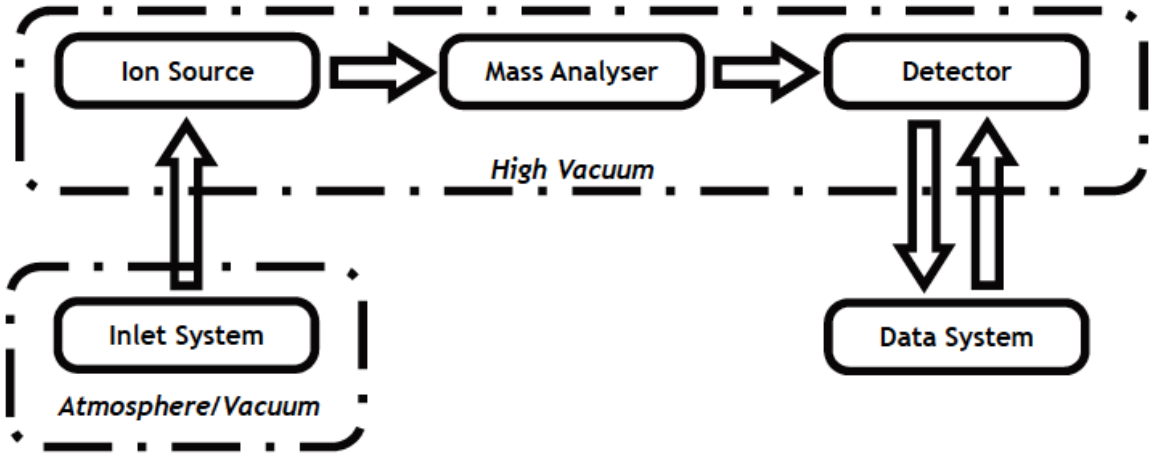


Figure 13 – Schéma de principe d'un spectromètre de masse, reproduit à partir de [28].

En fonction de la complexité de l'échantillon et de l'information que l'on cherche à obtenir, on choisit le type de spectromètre de masse qui corresponde à l'application souhaitée.

Ainsi, on différencie de façon simpliste deux grands types d'applications générales ainsi que le type de spectromètre qui convient :

- Industrie pharmaceutique : caractérisation et quantification de principe actif, généralement couplé en chromatographie et utilisant des spectromètres de masse peu résolutifs tels que les filtres quadripolaires placés en série (triples quadripôles). Les échantillons sont souvent un mélange de quelques composés, voire l'analyse d'un composé supposé pur.
- Analyses environnementales et biologiques : caractérisation et quantification de composés, couplé ou non à de la chromatographie. A la différence de l'industrie pharmaceutique, les échantillons sont souvent un mélange d'un nombre important de composés, traditionnellement appelés « échantillons complexes ». Du fait de cette complexité, plusieurs stratégies sont disponibles : réduire la complexité en utilisant la chromatographie, ou avoir suffisamment de résolution en spectrométrie pour observer l'ensemble de la complexité sans confusion. Du fait de cette variabilité d'applications, tous les spectromètres de masse peuvent être considérés, du filtre quadripolaire au spectromètre à transformée de Fourier et au temps de vol.

L'application considérées dans le cadre des travaux effectués sur des échantillons extraterrestres (météorites) ou synthétiques est comparable aux analyses environnementales et biologiques. Ainsi, nos échantillons sont réputés complexes et nécessitent une définition précise de l'information que l'on souhaite obtenir dans le but de sélectionner la stratégie analytique adéquate. Deux types de stratégies sont possibles, pouvant être subdivisées par la suite : les analyses ciblées et les analyses non-ciblées.

Une analyse ciblée est définie comme une analyse qui a vocation à identifier et/ou quantifier un jeu de composés connus, et seulement ceux-ci. Ce genre d'analyse est souvent effectuée en couplage chromatographique, principalement en phase gaz du fait de la haute résolution chromatographique disponible avec ce type d'instrument.

A l'inverse, une analyse non-ciblée est définie comme une analyse qui a vocation à voir le maximum de composés possibles contenus dans l'échantillon, aux limites instrumentales près. Ainsi, ce genre d'analyse peut être effectué en spectrométrie de masse seule ou couplée à la chromatographie, souvent liquide pour s'affranchir des contraintes fortes imposées par la chromatographie en phase gazeuse. Les analyses en spectrométrie de masse seules sont appelées « analyses en infusion directes » et sont aussi importantes que les analyses couplées à

la chromatographie. En effet, bien que dans le cas d'une analyse couplée à la chromatographie on puisse obtenir des informations structurelles en plus de l'information stœchiométrique fournie par la spectrométrie de masse, les limites de détections sont différentes entre une analyse en infusion directe et une analyse en chromatographie. Ainsi, cette capacité de l'analyse en infusion directe d'avoir une limite de détection très faible permet d'accéder à un ensemble de molécules potentiellement plus larges que l'ensemble de molécules accédées à travers une analyse en couplage chromatographique.

1.3.2. Utilité de la haute résolution

Un échantillon complexe est un mélange de plusieurs composés, pouvant aller de plusieurs dizaines à plusieurs milliers de molécules, isomères non inclus. Cette diversité est un problème pour l'analyse puisqu'une identification nécessite un pouvoir séparateur suffisant pour séparer le signal généré par un composé, du signal généré par un autre composé proche, que le signal soit résolu en temps ou en masse. La résolution chromatographique (*i.e.* temporelle) est un problème de chimie et de géométrie de colonne, alors que la résolution en masse est un problème intrinsèque à la définition des formules stœchiométriques, qui est défini mathématiquement par la résolution d'équations diophantiennes dont les inconnues sont les coefficients stœchiométriques :

$$masse = \sum \upsilon_i m_i$$

Équation 1 - Équation diophantienne définissant la masse à partir de sa formule stœchiométrique, i étant un élément du tableau périodique, m sa masse de l'atome et υ son coefficient stœchiométrique associé.

Le problème ici est la définition générale de la masse. En effet, si l'on considère deux groupements simples tels que N et CH_2, et que l'on définit la masse des éléments comme des nombres entiers (H=1u ; C=12u ; N=14u), alors les deux groupements ont une même masse de 14u, rendant impossible de les différencier. Dès lors, il apparait nécessaire de déterminer la masse exacte des éléments et du composé analysé en utilisant non leur masse entière, mais leur masse décimale. Ainsi, si l'on considère les mêmes deux groupements précédents, mais que l'on considère cette fois-ci la masse exacte des éléments (H=1,007(825) ; C=12,000(000) ; N=14, 003(074)), la masse exacte de CH_2 est alors de 14, 015(650), valeur très différente de la masse de N. On remarque également que, pour cet exemple, il aurait suffi de seulement deux décimales pour permettre de différencier les deux groupements. Cette précision, entre aucune décimale et plusieurs décimales, est directement liée à la résolution que le spectromètre de masse est capable de fournir. Il devient ainsi nécessaire d'avoir une résolution suffisamment élevée pour permettre de différencier les formules stœchiométriques d'un échantillon complexe.

Cette utilisation de la masse exacte révèle l'intérêt d'utiliser les décimales des masses mesurées. Cependant, comme les valeurs de masses nominales sont grandes par rapport aux décimales intéressantes, il faut transformer les valeurs : c'est ce qu'on appelle le « défaut de masse » qui est calculé comme suit :

$$Défaut\ de\ Masse = masse - arrondi(masse)$$

Équation 2 – Définition du défaut de masse

L'avantage de cette représentation, à la différence du spectre de masse, est que l'information concernant la stœchiométrie des molécules détectées est directement visible sur le graphique. En effet, comme présenté en Figure 14 pour quelques atomes, chacun a un défaut de masse unique, certains étant positifs et une grande majorité étant négatifs. Comme la matière organique comporte beaucoup de carbone et d'hydrogène, pour quelques atomes d'azote et d'oxygène, les défauts des masses des molécules organiques sont majoritairement positifs pour des masses inférieures à 500 Da.

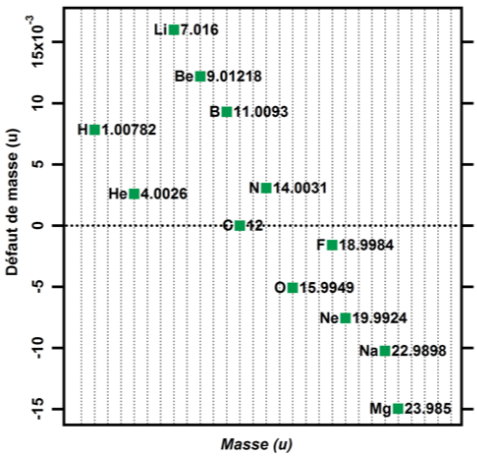


Figure 14 -Diagramme du défaut de masse en fonction de la masse pour quelques atomes.

Également, du fait de l'invariance des défauts de masse, l'ajout ou le retrait d'un ou plusieurs groupements identiques à partir d'une même molécule va engendrer des alignements représentatifs dans les graphes. Par exemple, ajouter des CH_2 à une molécule va faire augmenter son défaut de masse de +0,01564 Da par groupement ajouté. Cette additivité des groupements permet d'effectuer des représentations graphiques où des lignes sont tracées dans le DMvM représentant une variation d'un groupement unique, comme présenté à titre d'illustration en Figure 15.

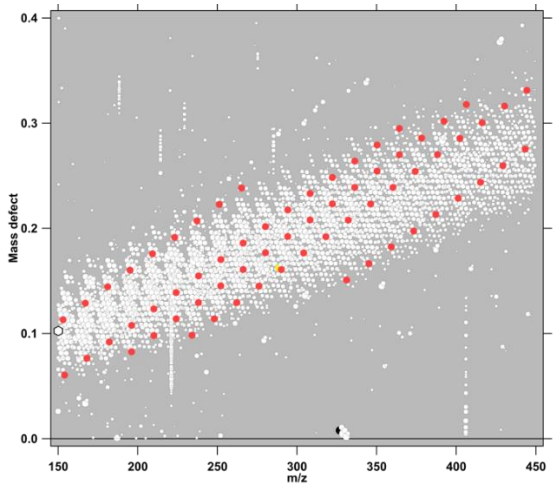


Figure 15 – Représentation de cinq familles qui ne varient qu'en CH_2.

D'autres problèmes sont présents, notamment en ce qui concerne la résolution mathématique des équations diophantiennes. Des astuces de traitement de données existent pour résoudre ces problèmes, et sont introduits par la suite.

1.3.3. Spectromètres de masse à haute résolution

Pour permettre l'analyse d'échantillons complexes avec une résolution suffisante, seuls les instruments à transformée de Fourier sont actuellement capable de fournir une résolution suffisante sur une gamme de masse allant jusqu'à plusieurs centaines de Daltons. Ces instruments utilisent soit des champs magnétiques intenses, soit des champs électrostatiques intenses pour fournir un pouvoir séparateur suffisant pour analyser la diversité moléculaire d'un échantillon complexe. Ainsi, FT-ICR (Fourier Transform Ion Cyclotron Resonance) et Orbitrap sont respectivement basés sur les champs magnétiques et les champs électrostatiques. Comme n'importe quel spectromètre de masse, ils fonctionnent sur le même principe : ioniser les molécules, les placer sur une trajectoire connue et analyser leur réponse.

1.3.3.1. *L'Orbitrap*

L'IPAG possède un LTQ-OrbitrapXL, installé en 2008. Cet Orbitrap est la première version commerciale de ce type d'instrument, vendu par ThermoScientific.

L'Orbitrap est basé sur l'application de champs électrostatiques dans la trappe orbitale pour effectuer l'analyse en masse. Une représentation schématique d'un LTQ-OrbitrapXL est disponible en Figure 16. Les nouvelles versions de l'Orbitrap présentent des modifications significatives en amont de la trappe orbitale. L'analyse en masse des ions injectés est basée sur le mouvement de précession des ions autour de l'électrode centrale de la trappe orbitale.

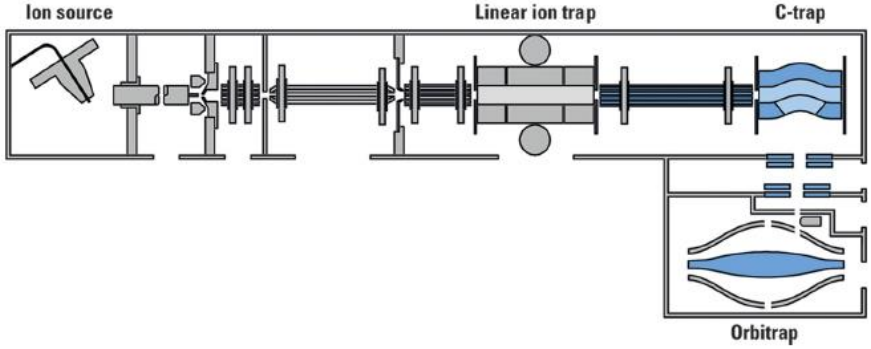


Figure 16 – Représentation schématique d'un Orbitrap, reproduit à partir de [29] .

Les ions ont une trajectoire complexe dans la trappe orbitale. Les ions sont injectés par la C-trap tangentiellement au champ électrique. Les ions sont alors entrainés par le champ électrostatique généré par l'électrode centrale, et retenus par le champ généré par les électrodes externes. Ainsi, les ions sont stabilisés en orbite dans la cavité entre les électrodes. Néanmoins, du fait de la forme particulière de l'électrode centrale, les ions créent un anneau autour de l'électrode centrale, comme représenté en Figure 17. C'est le mouvement axial (selon l'axe de l'électrode centrale Z) qui génère le signal détecté par les électrodes externes.

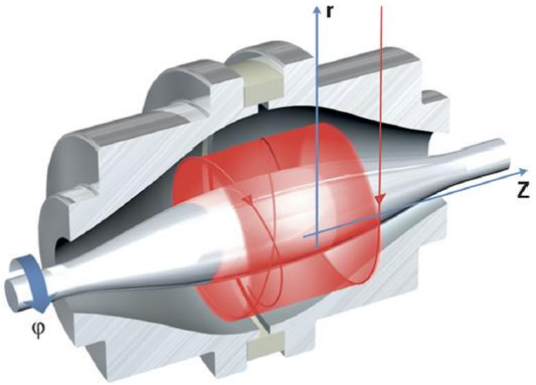


Figure 17 – Représentation de l'anneau d'ions (en rouge) en orbite dans la cavité d'une trappe orbitale, reproduit à partir de [29] .

Une trappe orbitale est un détecteur pulsé. Les ions sont stockés et injectés par la C-trap avant analyse et détection dans la trappe orbitale. Le temps de stockage des ions dans la C-trap avant injection est contrôlé de façon indirecte par la fonction AGC (Automatic Gain Control) qui s'assure de stocker un nombre d'ions proche de la valeur entrée. Ainsi, si beaucoup d'ions sont présents, le temps de stockage est court. Si peu d'ions sont présents, le temps de stockage est plus long et est de maximum 500ms.

1.3.3.2. *L'ICR*

Pour pouvoir discuter de la résolution de l'Orbitrap, il convient d'effectuer des analyses avec un instrument ayant une résolution plus élevée tel qu'un FT-ICR. Comme l'IPAG n'en

possède pas, un projet a été rédigé pour aller effectuer des analyses sur l'ICR 12T de Rouen, encadré sur place par Isabelle Schmitz-Afonso et Christopher Rüger.

Un ICR, dont une représentation schématique est disponible en Figure 18, utilise des champs magnétiques allant de 3T à 21T pour analyser en masse les ions. Cette analyse est effectuée au niveau de la cellule ICR en utilisant le mouvement cyclotron des ions dans la cellule. Le mouvement magnétron et de trapping, les deux autres types de mouvement naturel des ions dans une cellule ICR, sont réputés négligeables et n'interférant pas dans l'analyse.

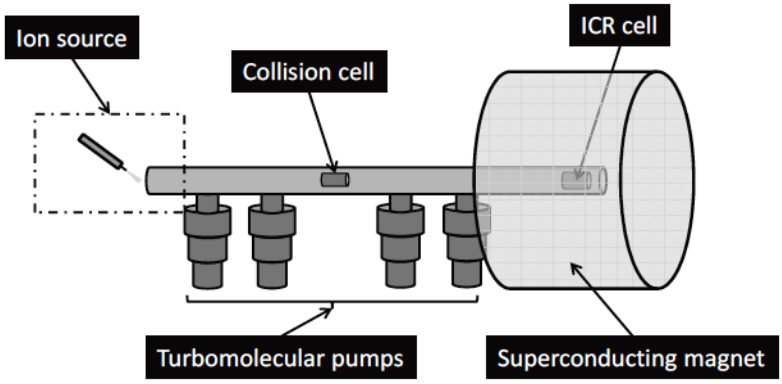


Figure 18 – Représentation schématique d'un FT-ICR, reproduit à partir de [28].

L'analyse des ions est effectuée au niveau de la cellule ICR, en utilisant le principe du mouvement cyclotron des ions dans la cellule. L'ensemble de la séquence analytique est schématisée en Figure 19. Quand les ions arrivent dans la cellule, leur trajectoire est proche de l'axe, et sur une orbite trop faible pour être détectée. Ainsi, la cellule doit exciter les ions pour les mettre sur une orbite plus large, détectable. Le diamètre de l'orbite est directement dépendant du ratio masse/charge des ions dans la trappe. Cette excitation rend également le faisceau d'ions spatialement cohérant, permet de casser les clusters ion-molécules et enfin, lorsque la détection est terminée, d'expulser les ions hors de la cellule.

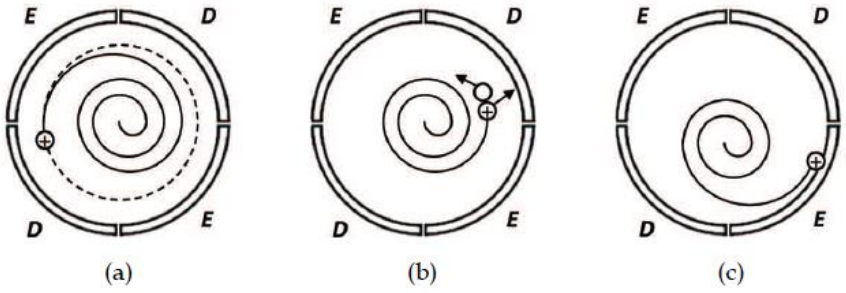


Figure 19 – Procédure d'excitation des ions dans la trappe, reproduit à partir de [28]. (a) accélération des ions pour les placer sur une orbite supérieure, propice à la détection ; (b) accélération pour casser les clusters ion-molécule ; (c) expulsion des ions de la cellule. E sont les électrodes d'excitation et D sont les électrodes de détection.

La cellule ICR est un détecteur pulsé, ce qui nécessite de stocker les ions avant de les injecter dans la cellule. Ce temps de stockage avant injection est un paramètre ajustable et critique de l'analyse. En effet, injecter trop d'ions va créer des effets de champs et faire perdre en qualité l'analyse finale. Injecter trop peu et le signal ne sera pas assez intense pour obtenir une sensibilité adéquate. L'ICR permet également d'ajuster la durée de l'analyse dans la cellule, *i.e.* la durée d'un transient, afin d'ajuster la résolution du spectre. Plus la durée d'un transient est longue, plus la résolution sera élevée. Néanmoins, cette durée ne peut pas être trop grande puisque les ions dans la cellule perdent de l'énergie au cours du temps, réduisant d'autant l'intensité du signal mesuré.

1.3.3.3. *Points communs et différences*

Que ce soit pour un ICR ou un Orbitrap, la cellule de détection est composée de plusieurs sections, soit respectivement huit et deux sections. Lorsqu'une particule chargée oscille dans l'espace, cette particule génère un champ électrique autour d'elle. Ce champ peut alors être détecté par l'utilisation d'électrodes placées autour de la cavité dans laquelle oscille la particule. Une unique électrode générera un signal continu, indiquant seulement qu'une particule chargée est présente. C'est par l'utilisation différentielle de plusieurs électrodes, indiquant si la particule est proche ou éloignée de l'électrode en question, que l'on est capable de déterminer un signal proportionnel à la fréquence d'oscillation, et potentiellement de pouvoir remonter à la masse de la particule qui génère ce champ. Le signal détecté est alors un spectre résolu en temps, aussi appelé « transient ». Ce spectre est converti en spectre de fréquences en appliquant une transformée de Fourier, qui est ensuite calibré en masse pour obtenir un spectre résolu en masse.

L'utilisateur d'un FT-ICR a accès au transient et a l'entière liberté d'appliquer les transformations et traitements du signal qu'il souhaite pour obtenir le spectre en masse final. Les opérations peuvent aller d'un simple traitement de bruit, à des traitements pour calculer la phase des ions dans la trappe et augmenter la résolution du spectre de masse final. Au contraire des FT-ICR, l'utilisateur d'un Orbitrap n'a pas accès aux transients, mais seulement au spectre de masse final. Plusieurs opérations sont effectuées entre la détection et la génération du spectre final, comme des opérations de traitement du bruit. Cet ensemble de transformation n'est pas accessible et est effectué par défaut. Du fait de cette différence avec les FT-ICR, Thermo Scientific a modifié la terminologie en renommant les transients en micro-scans et le spectre final, résultat d'un ou plusieurs micro-scans, en scan. Cette différence micro-scans et scans, ainsi que leur impact sur les résultats en fonction de leur combinaison est développé par la suite.

Une autre différence majeure entre ICR et Orbitrap est la résolution des spectres de masse. Historiquement, les ICR sont capables d'obtenir des résolutions de plusieurs centaines de milliers, voire quelques millions, alors que les Orbitrap sont limités à un peu plus d'une centaine de milliers. Une comparaison rapide de la différence de résolution entre Orbitrap et ICR est disponible en Figure 20. Au premier ordre, cette différence est due majoritairement au fait que la résolution des ICR est directement proportionnelle au champ magnétique. Dès lors, il suffit d'augmenter la taille des aimants pour obtenir une meilleure machine du point de vue de la résolution. Cependant, augmenter la taille des aimants à un coût en liquide cryogénique, ainsi qu'un coût en infrastructure du fait du champ magnétique puissant de l'appareil nécessitant une zone d'exclusion autour de l'appareil. Ces quelques contraintes rendent ainsi l'exploitation des FT-ICR onéreuse. A l'inverse, augmenter le champ électrostatique ne permet pas une meilleure résolution en Orbitrap. De plus, les champs électrostatiques ne nécessitent pas de fluides cryogéniques, et n'imposent pas un environnement contrôlé autour de la machine.

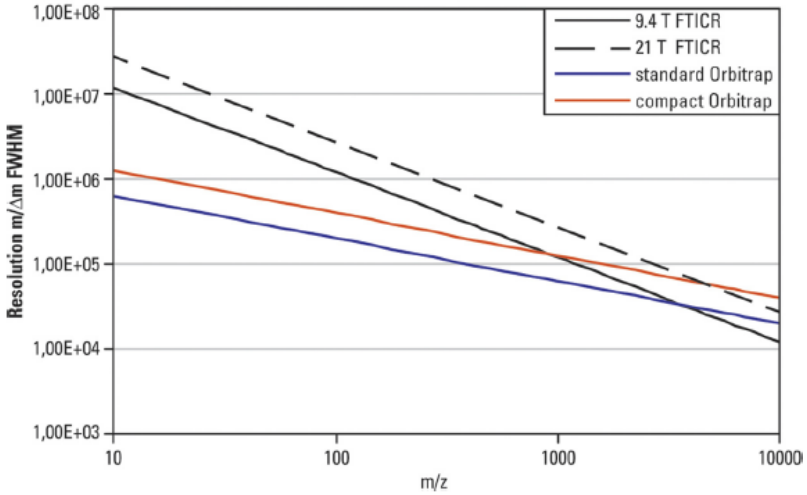


Figure 20 – Comparaison de la résolution des Orbitrap et des FT-ICR en fonction de la masse. Reproduit à partir de [29].

Une différence se trouve également dans l'évolution de la résolution avec le ratio masse/charge. En effet, si l'on exprime mathématiquement la résolution des deux instruments, équations mathématiques présentées en Équation 3, on s'aperçoit alors que l'évolution de la résolution est proportionnelle à $(m/z)^{-1/2}$ pour un Orbitrap alors que la résolution est proportionnelle à $(m/z)^{-1}$ en FT-ICR.

$$R_{Orbitrap} = \frac{1}{2\Delta w_{50\%}}\left(\frac{kz}{m}\right)^{1/2}$$

$$R_{ICR} = \frac{1{,}274 * 10^7 z B_0 T_{aqn}}{m}$$

Équation 3 – Expressions de la résolution en Orbitrap et en ICR, extraits respectivement de [30,29].

Cette différence de proportionnalité explique le fait que la résolution des FT-ICR soit moins bonne à hautes masses que la résolution des Orbitrap, comme observé sur la Figure 20. Cependant et jusqu'à présent, la résolution des FT-ICR est inégalée pour les molécules de faible masse (m/z<1000 Da), particulièrement lorsque de multiples hétéroatomes sont présents dans les échantillons analysés. Des développements récents concernant les Orbitrap permettent d'atteindre un million de résolution[31], mais cette technologie est pour l'instant peu répandue.

1.3.4. Sources d'ionisations et problématique des échantillons complexes

La source d'ionisation est responsable de la production des ions avant d'entrer dans le spectromètre de masse. Différents types de sources existent, ayant des caractéristiques différentes à prendre en compte notamment en ce qui concerne l'état physique de l'échantillon. Ainsi, certaines sources n'acceptent que des échantillons en phase liquide alors que d'autres peuvent analyser toutes les phases physiques.

Dans le cadre de nos travaux, seules deux types de sources sont utilisées et décrites ici : la source électrospray (ESI – ElectroSpray Ionisation) et la source d'ionisation par désorption laser (LDI – Laser Desorption Ionisation). D'autres types de sources existent, telles que les sources MALDI, APPI ou encore APCI. Alors que les sources ESI, APPI ou APCI n'acceptent seulement que des échantillons ayant été dissous dans un solvant, les sources LDI ou MALDI peuvent analyser toute sorte d'échantillon. Ainsi, l'information apportée par ces sources sera différente étant donné la fraction d'information accédée par la source. Il existe également des différences fondamentales entre sources acceptant seulement des liquides. En effet, comme présenté en Figure 21, l'information accessible dépend de la polarité des molécules ainsi que de leur masse. En fonction du type d'échantillon et de l'information que l'on cherche à obtenir, certains choix de sources sont meilleurs que d'autres. Par exemple, les sources APPI et APCI sont tout à fait adaptées aux échantillons peu polaires (par exemple, hydrocarbures et molécules aromatiques pour des échantillons géologiques tels que charbon, pétrole ou schistes) alors que la source ESI est bien adaptée aux échantillons plus polaires, possédant de nombreux groupements capables d'échanger des protons, tels que groupements carboxyliques, amines ou encore thiols.

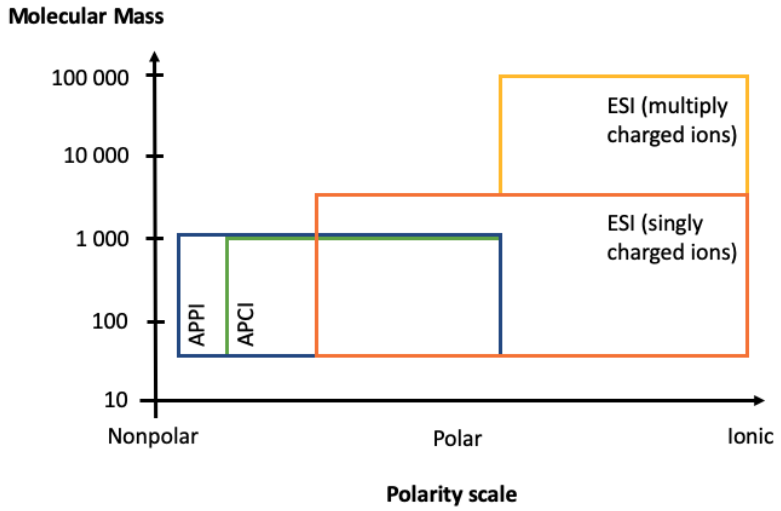


Figure 21 – Représentation graphique des domaines de polarités accessibles en fonction de la masse en utilisant les sources acceptant uniquement des échantillons sous forme liquide. Reproduit et adapté depuis [32].

1.3.4.1. *La source ESI*

Dans le cas de l'analyse de nos échantillons, que ce soit de la matière solubilisée de météorites ou d'échantillons synthétiques, les molécules présentes possèdent des groupements polaires. Ainsi, nous utilisons une source ESI pour analyser ces échantillons. Cette source d'ionisation est à pression atmosphérique et à température ambiante. Une représentation schématique est fournie en Figure 22a. L'échantillon arrive à travers un capillaire métallique sur lequel est appliqué un champ électrostatique de plusieurs kV. De manière coaxiale à ce capillaire est injecté un débit réglable de gaz, typiquement de l'azote, utilisé pour focaliser le flux de solution en sortie de capillaire, mais également pour aider à la désolvatation des gouttelettes ainsi formées.

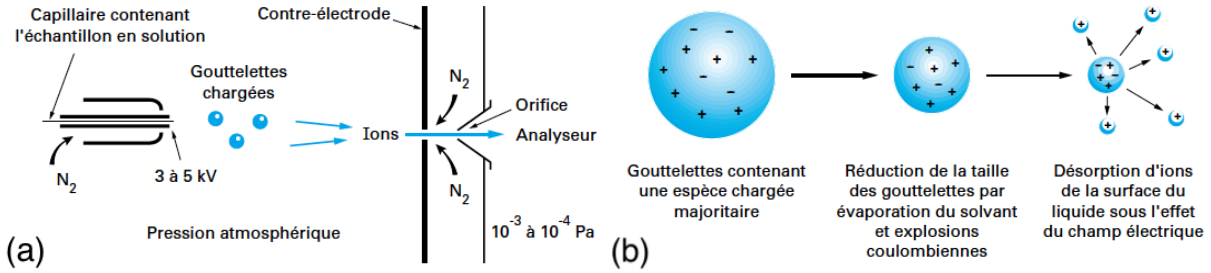


Figure 22 – (a) Schéma de principe d'une source électrospray, de l'injection de l'échantillon à son entrée dans le spectromètre de masse ; (b) Processus de création des ions chargés à partir de leur solution. Reproduit et adapté depuis [33].

En effet, sous l'effet du champ électrostatique et du flux de gaz, le liquide en sortie de capillaire est nébulisé en un nuage de fines gouttelettes de solvant, chargées et contenant les analytes. Sous l'effet de l'évaporation du solvant, les goutteles sont de plus en plus petites, ce qui rapproche les charges les unes des autres. Plus les charges sont proches, plus elles ont tendance à se repousser. Lorsque suffisamment de solvant est évaporé, une explosion coulombienne se produit et génère des gouttes plus petites. Ce processus s'enchaine jusqu'à ce que l'analyte soit totalement désolvaté et chargé comme présenté en Figure 22b. Cette technique d'ionisation créé des ions mono-chargés et multichargés, mais des molécules neutres peuvent également entrer dans le spectromètre, ainsi que des clusters ions-solvants. C'est alors le rôle des optiques ioniques placées juste après l'entrée du spectromètre de masse de sélectionner les composés chargés, de casser les clusters et de transférer les ions restant jusqu'à l'analyseur en masse.

Par construction de la source électrospray, le solvant (ou le mélange de solvants) doit répondre à des contraintes particulières. En effet, il doit (1) être suffisamment volatil pour

permettre une désolvatation rapide des composés avant d'entrer dans le spectromètre de masse ; (2) permettre le processus d'ionisation en participant au retrait ou à l'addition de protons sur les composés à analyser ; (3) être compatible avec l'analyse en spectrométrie de masse, c'est-à-dire qu'il ne va pas apporter d'interférences en ajoutant des espèces étrangères à l'échantillon capable de créer des adduits stables avec les molécules de l'échantillon. Ainsi, on choisira un solvant ou un mélange de solvant tels que le Méthanol, l'Acétonitrile, l'Eau, éventuellement tamponnés avec des composés volatils de faible masse tels que l'acide formique, l'acide acétique et leur équivalent ionique utilisant par exemple l'ion ammonium en contre ion. On évitera à tout prix les solvants non polaires et non protiques (ex : alcanes ou solvants chlorés), ainsi que les tampons peu volatils et possédant des hétéroatomes lourds tels que les tampons phosphates, l'acide chlorhydrique ou la soude, les tampons utilisés en biologie tel que le DMSO et autres.

1.3.4.2. *La source LDI*

Une source LDI, pour *Laser Desorption Ionisation*, est une source qui permet l'analyse d'échantillons sous toutes formes physiques. Son utilisation principale est néanmoins l'analyse d'échantillons solides, non accessibles par des actions de dissolution dans un solvant. L'échantillon est placé sur une plaque métallique et on tire des impulsions laser. La puissance et la longueur d'onde du laser va déterminer si un plasma se crée à la surface de l'échantillon, permettant d'ioniser les molécules de surface. Comme présenté en Figure 23, les ions créés sont alors attirés par un champ électrostatique à l'entrée du spectromètre de masse, permettant l'analyse en masse des ions ainsi créés.

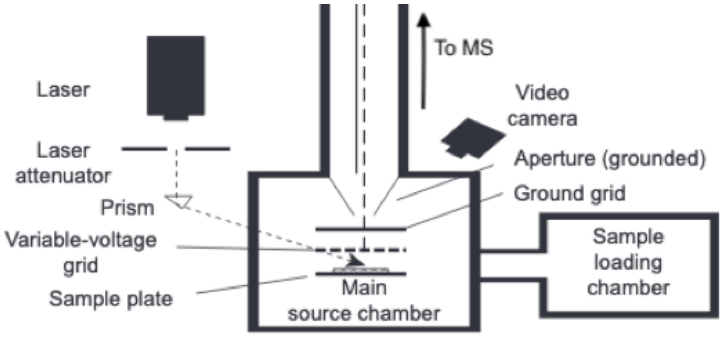


Figure 23 – Représentation schématique d'une source LDI. Le support d'échantillon est souvent mobile pour permettre une analyse spatiale de l'échantillon. Adapté de [34].

Un problème majeur de cette analyse est, comme présenté en Figure 24, la création non seulement d'ions, mais également de particules, de clusters, de fragments et de radicaux. À cette création à priori non contrôlée de particules chargées ou non, s'ajoute une potentielle chimie dans le plasma, c'est-à-dire que l'énergie déposée en surface peut être suffisamment importante pour engendrer des réactions chimiques et ainsi altérer la composition moléculaire résultant de l'analyse. Cette chimie peut se voir facilement en LDI avec par exemple la création de fullerène, espèces totalement et uniquement constituées de carbone. Tout l'enjeu de l'utilisation de cette technique est le réglage du laser, pour avoir suffisamment d'ionisation sans générer un plasma trop énergétique. Ainsi, chaque échantillon doit être testé et la puissance du laser ajustée en conséquence.

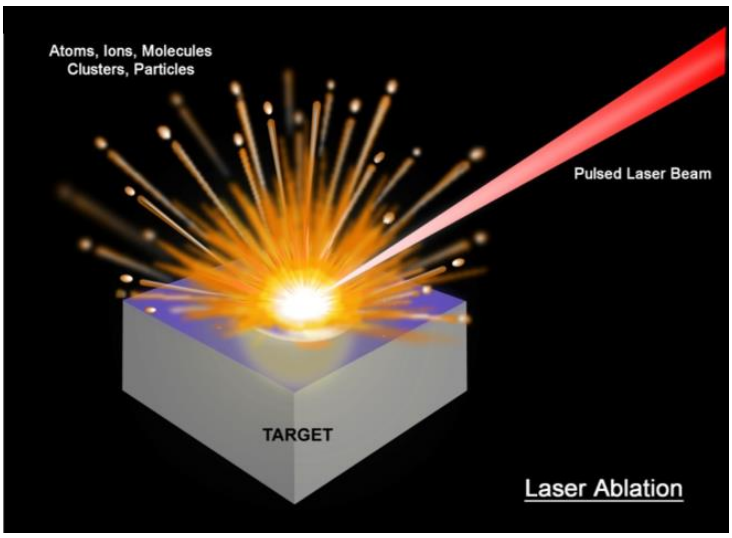


Figure 24 – Représentation schématique du processus d'ionisation en LDI. Extrait du cours de Spectrométrie de masse donné par F.-R. ORthous Daunay, IUT Grenoble 1.

Un problème spécifique pour l'application à des échantillons complexes est l'effet de matrice. En effet, pour un mélange simple de quelques composés, on associe à un composé un coefficient de réponse, incluant un éventuel ratio d'ionisation induit par la source et une constante de réponse induite par le détecteur. Ce coefficient de réponse permet finalement, si l'on détermine ces constantes, de faire des analyses quantitatives. Dans le cas d'analyses d'échantillons complexes, une compétition à l'ionisation se joue entre les espèces concentrées et/ou faciles à ioniser par rapports aux espèces minoritaires et/ou peu ionisables. Dès lors, les ratios de concentrations apparentes donnés par une telle analyse sont faussés par cet effet de matrice, qui favorise les espèces concentrées et /ou facilement ionisables par rapport aux autres. Ainsi, l'information accessible dans une analyse effectuée de la sorte doit prendre en compte cette différence de ratio d'ionisation inhérente à la source. Cela impactera, finalement, la sensibilité de la méthode aux espèces peu concentrées, nécessitant d'éventuelles adaptations des protocoles d'acquisitions pour tenter d'observer ces espèces peu concentrées.

1.3.5. Traitement des données en spectrométrie de masse

En fonction de la stratégie analytique choisie, *i.e.* analyse ciblée ou non-ciblée, les besoins en traitement de données ne sont pas les mêmes. Ainsi, traiter des données de façon non-ciblée est plus complexe qu'une analyse ciblée où ce que l'on cherche est déjà connu et est faisable facilement avec les logiciels commerciaux et de pilotage des instruments, tel qu'Xcalibur™ (Thermo Scientific, Allemagne) pour les Orbitrap.

Pour effectuer le traitement des données lors d'une analyse non-ciblée, les logiciels commerciaux adaptés aux échantillons complexes ne sont pas disponibles, ou sont extrêmement spécialisés tels que les logiciels de protéomique (*Proteome Discoverer*, Thermo Scientific, Allemagne), inadaptés à l'analyse de petites molécules non issues d'échantillons biologiques. C'est pour cela que le logiciel Attributor est développé à l'IPAG par F.-R. Orthous-Daunay en utilisant IGOR Pro (WaveMetrics, USA). Ce logiciel permet de traiter les données issues des Orbitrap, mais également des FT-ICR. Les fonctionnalités de base pour l'utilisateur sont décrites dans la thèse d'Aurélien Fresneau [14], et plusieurs articles sont basés sur les traitements effectués avec ce logiciel tels que les analyses d'analogues de glaces cométaires par Grégoire Danger et al [12,13,15] ou dans des analyses d'analogues d'aérosols d'atmosphères d'exoplanètes [Vuitton et al, 2020, Accepté].

Quelques fonctionnalités doivent être décrites :

- Réduction des données : Après l'acquisition des données, il est nécessaire de convertir avant toute chose les données en centroïdes. Ce traitement automatique

détecte les sommets des signaux en masse et réduit les artefacts dus à la transformée de Fourier (*ringing* par exemple). C'est sur ce jeu de données que se base l'ensemble des analyses.

- Hi-Res Calibration : une calibration interne basée sur une liste de molécules connues peut être réalisée dans le but de corriger les erreurs de calibration provenant de l'instrument. Ces erreurs peuvent entrainer de mauvaises attributions, et ainsi fausser les interprétations des résultats. Cette calibration est basée sur des polynômes de degrés 3 à 10 qu'il faut sélectionner en fonction de la forme de la courbe des erreurs observées. Cette calibration dépendant intégralement des molécules fournies, leur sélection doit faire l'objet d'une attention particulière.
- Attribution : En effet, deux façons d'attribuer sont possibles :

Attribution 1 : atomes	Attribution 2 : groupes
- C	- C
- H	- CH_2
- N	- CH_3/CH_5
- O	- NH
	- O

Ces deux façons d'attribuer des signaux en masse ont tous deux le même objectif : résoudre l'équation diophantienne relative à la masse :

$$masse = \sum \upsilon_i m_i$$

Équation 4 – Équation liant la masse d'une molécule à la somme du produit de ses coefficients stœchiométriques υ associés à la masse m du groupement considéré.

Cette équation peut ainsi prendre comme masse m des atomes ou des groupements d'atomes, sans aucun changement sur les mathématiques qui lui sont liées. L'avantage de l'utilisation de groupements d'atomes au lieu d'utiliser des atomes uniques est que l'on peut ajouter des informations chimiques dans la résolution de cette équation. En effet, il est tout à fait possible d'obtenir une formule stœchiométrique ne respectant pas la valence des atomes en utilisant une attribution avec des atomes uniques, alors que l'utilisation d'une matrice de groupements d'atomes appropriée sera capable de ne pas proposer de solution ne respectant pas les règles de valence des atomes. La détermination d'une matrice appropriée n'est pas l'objet de ce travail, et l'on utilisera les matrices déjà utilisées par le groupe. Ces matrices seront indiquées lorsque nécessaire, ainsi que les limites imposées relatives au nombre de groupements possibles puisque cela a un impact direct sur les résultats des attributions. On note également que les signaux analysés en spectrométrie de masse sont nécessairement des signaux ioniques, et donc qu'il faut prendre en compte la masse des électrons ajoutés ou retirés lorsque l'on considère les masses exactes des molécules. En pratique, cela revient au même que l'Équation 4 où, au lieu de considérer les atomes uniquement, on considère les atomes et les électrons dans le calcul, et avec la possibilité pour le coefficient stœchiométrique lié aux électrons d'avoir une valeur négative.

Dans Attributor, deux façons d'attribuer existent :

- Fasttribution : cet outil d'attribution calcule pour chaque signal la formule stœchiométrique ayant la plus faible erreur en fonction des groupements et/ou atomes, ainsi que des limites introduites par l'utilisateur.
- Graphttribution : cet outil d'attribution recherche les distances connues (CH_2, NH ou O par exemple) dans le DMvM et produit un graphe à partir d'une molécule qui permet de se déplacer de proche en proche selon les

variations demandées. Ainsi, un spectre peut entièrement être attribué à partir d'une unique molécule, préférentiellement située à basse masse, si l'on est capable de lier l'ensemble des variations d'un point à un autre par un réseau de variations de groupements connus.

- Représentation des données attribuées : représenter des milliers de formules brutes est un défi qu'il faut relever. Plusieurs représentations sont classiques : le DBE, pour « Double Bond Equivalent » et les représentations de type Van Krevelen.
  Le DBE est calculé comme indiqué en Équation 5 :

$$DBE = 1 + C + \frac{N}{2} - \frac{H}{2}$$

Équation 5 – Définition mathématique du « Double Bound Equivalent » pour une formule ne contenant que CHNO.

Cette transformation mathématique de l'espace des attributions permet de quantifier l'insaturation des formules stœchiométriques, que ce soient des liaisons insaturées ou des cycles. Si des isotopes du carbone, de l'azote ou de l'hydrogène sont inclus, ils doivent être ajoutés dans cette formule selon les mêmes ratios que leur isotope déjà inclus. Le DBE est une autre façon d'interpréter la valence des molécules, et l'invariance du DBE au nombre d'oxygènes par exemple indique que d'ajouter un oxygène dans une structure ne change pas son insaturation globale, que cet oxygène soit simplement ou doublement lié. Le DBE a également un intérêt pour l'attribution puisqu'un ion moléculaire aura nécessairement un DBE fractionnaire alors qu'une molécule neutre ou un ion radicalaire aura nécessairement un DBE entier.
Les représentations de type Van Krevelen sont des représentations faisant intervenir des ratios atomiques, comme présenté en Figure 25. Traditionnellement, on représente le ratio *H/C* en fonction du ratio *O/C [35]*. Cela a été introduit comme un moyen visuel pour classifier l'origine et la maturité des kérogènes et pétroles. Ces représentations sont par nature non linéaires et leur interprétation complexe. Cependant, des domaines peuvent être délimités sur ces représentations, et ainsi permettre une classification d'espace entre échantillons. Dans le cadre de cette étude, on utilisera plutôt les ratios atomiques en fonction de la masse, qui permettent des visualisations moins complexes et plus parlantes des attributions.

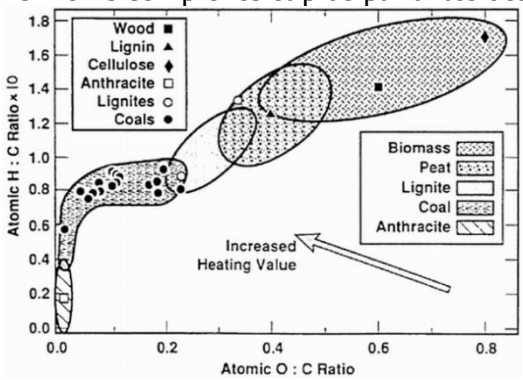


Figure 25 – Représentation de van Krevelen qui représente les différentes régions permettant de classifier les différents degrés de maturation en se basant sur la composition stœchiométrique de l'échantillon. Extrait de [35].

1.4. Méthodes expérimentales pour la chromatographique en phase liquide

Cette partie est en grande partie adaptée du livre « *Introduction to modern liquid chromatography* » par Snyder et al [36]. Seules les sources autres seront mentionnées lorsque nécessaire.

1.4.1. Petit point d'histoire de la chromatographie en phase liquide

La chromatographie telle qu'effectuée de nos jours est le fruit d'un long développement ayant commencé avant 1900. Ces différents travaux ont été réalisés avec des objectifs différents de ceux recherchés avec une séparation analytique, avec par exemple la démonstration que la migration de pétrole dans le sol engendre une stratification de la qualité du pétrole. C'est peu après 1900 que Mikhail Tswett a inventé la première technique de séparation sur colonne, en démontrant la faisabilité de la séparation et la récupération des composants de plusieurs extraits végétaux. Du fait de la séparation apparente des couleurs, avec des pigments oranges, verts et jaunes, cette technique a été appelée « chromatographie » pour représenter la séparation spatiale des couleurs. Ces travaux ont été redécouverts dans les années 1930 et ont engendré, entre autres, des développements tels que la chromatographie sur papier (1943), la chromatographie sur couche mince (1930-1950) ou encore la chromatographie de perméation sur gel (1960).

Les premiers équipements commerciaux apparaissent au début des années 1960, avec les équipements fournis par Waters Associates et DuPont. Les premiers livres et cours au sujet de cette technique séparative sont donnés à partir de 1971, et se sont développés depuis.

De 1960 à nos jours, les instruments et colonnes chromatographiques ont également évolués, permettant des analyses plus rapides et de meilleures résolutions, comme présenté en Figure 26a-f. Ces performances améliorées ont néanmoins un coût technologique, qui est une augmentation significative de la pression que doit supporter l'instrumentation comme représenté en Figure 26g. L'ensemble des théories relatives à la chromatographie ont cependant été réalisées avant l'apparition des premiers appareils. En effet, la dépendance de la qualité de la séparation à la taille des particules et à la pression appliquée a, par exemple, été publiée en 1941.

De la même manière, la chromatographie en phase gazeuse a été développée dans les années 1950, et les développements initiaux des colonnes ont été bénéfiques aux deux types d'instrumentation, sans distinction. Cependant, vu que la chromatographie en phase gazeuse n'est applicable qu'aux molécules volatiles et thermiquement stables, soit typiquement des molécules ayant un point d'évaporation inférieur à 300°C et une température de dégradation supérieure à 300°C, cette technique n'est pas directement applicable par exemple à la séparation d'échantillons biologiques ou de molécules à haut point d'ébullition. Ainsi, il est estimé que plus de 75% des molécules connues ne sont pas analysables en chromatographie en phase gazeuse. Ainsi, la puissance de la chromatographie en phase gaz n'est pas dans son universalité, mais dans son pouvoir séparateur et résolutif inégalable. En effet, les colonnes capillaires de plusieurs dizaines voire quelques centaines de mètres pour un diamètre interne micrométrique permettent d'obtenir des efficacités de colonnes très supérieures à celles obtenues sur leur contrepartie en phase liquide. La chromatographie en phase gazeuse est donc la méthode de choix en ce qui concerne les analyses et la quantification de molécules bien connues, compatibles avec les conditions de températures imposées. C'est pourquoi, à l'inverse, la chromatographie en phase liquide est repondue pour les analyses non-ciblées et pour les analyses d'échantillons fragiles tels que les échantillons biologiques par exemple.

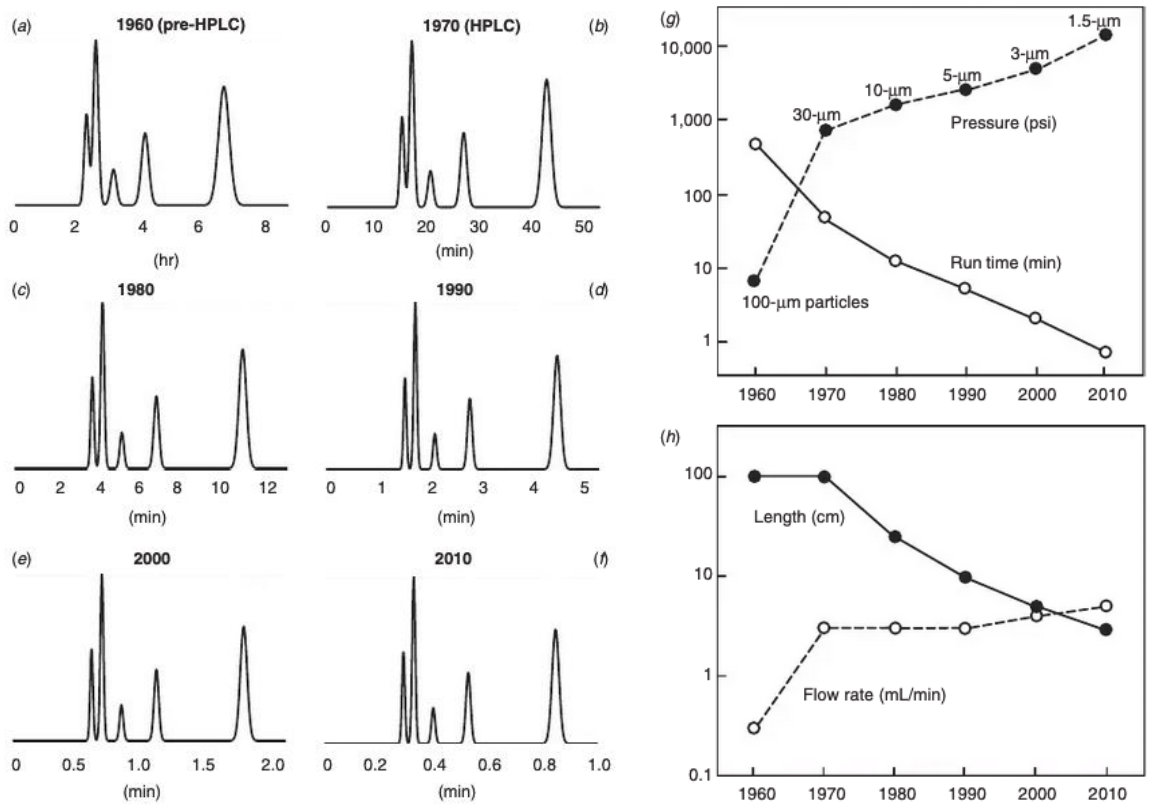


Figure 26 – Chromatogrammes illustrant l'évolution de l'HPLC au cours du temps. Échantillon : cinq herbicides. Conditions : 50% méthanol-eau, température ambiante. Les chromatogrammes a-f ont été simulés en se basant sur des données publiées. Les représentations g et h indiquent les détails des séparations a-f. Adapté et reproduit depuis [36].

1.4.2. Les colonnes en chromatographie liquide

En chromatographie, le principe de fonctionnement est toujours le même. Une phase mobile est appliquée à l'entrée d'un support portant de ce qu'on l'on appelle la phase stationnaire. L'application d'une pression suffisante crée alors un débit sur le support. Si l'on injecte dans le flux de phase mobile un mélange de molécule, alors ce dernier va circuler dans le support avec le fluide, et se retrouver physiquement proche de la phase stationnaire. Si les molécules et la phase stationnaire ont de l'affinité, alors les molécules vont s'adsorber à la surface de la phase stationnaire tout en étant à l'équilibre avec la phase mobile dans son environnement immédiat. Cet équilibre entre molécules, phase stationnaire et phase mobile est représenté en Figure 27. On notera que l'équilibre entre phase mobile et phase stationnaire est également présent, notamment lorsque la phase stationnaire est, par exemple, sensible au pH de la phase mobile du fait de la présence de groupes acido-basiques à la surface.

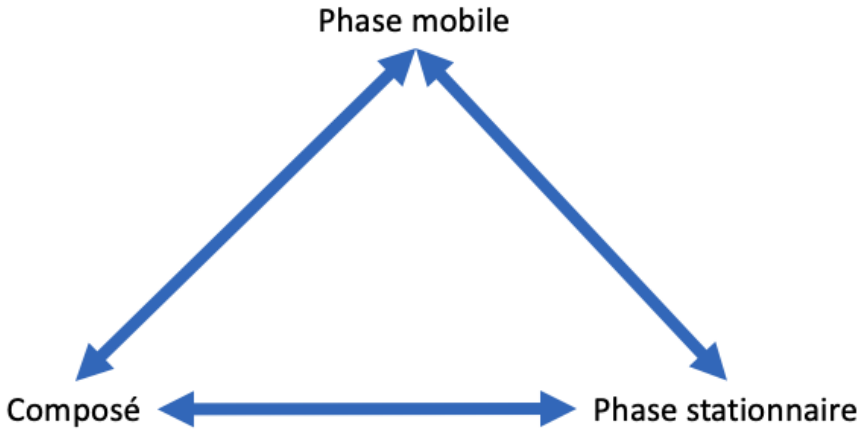


Figure 27 – Schématisation de l'équilibre entre phase stationnaire, phase mobile et composé en chromatographie.

Ainsi, lorsque l'ensemble des composés a été élué, c'est-à-dire que l'ensemble des composés sont sortis de la colonne après avoir subi de multiples équilibres entre phase stationnaire et phase mobile, il faut pouvoir détecter ces composés mélangés dans la phase mobile. Historiquement, les premières chromatographies récupéraient des fractions temporelles avant de les analyser sur un autre appareil. De nos jours, la détection est presque toujours faite en temps réel et en ligne, ce qui implique lors du développement de prendre en compte les éventuelles particularités du détecteur. Une représentation schématique de l'ensemble du processus analytique est présentée en Figure 28.

Basé sur ce processus, si l'ensemble de l'instrumentation capable de délivrer un débit stable et de soutenir une pression importante est maitrisé, alors le seul critère important pour développer et utiliser une méthode en chromatographie en phase liquide est la colonne, et plus particulièrement la définition de la phase stationnaire. Pour la chromatographie en phase liquide, seules des colonnes remplies sont disponibles, ce qui n'est pas le cas pour la chromatographie en phase gazeuse : colonnes capillaires avec différentes textures de revêtement ou colonnes remplies sont disponibles pour effectuer les séparations.

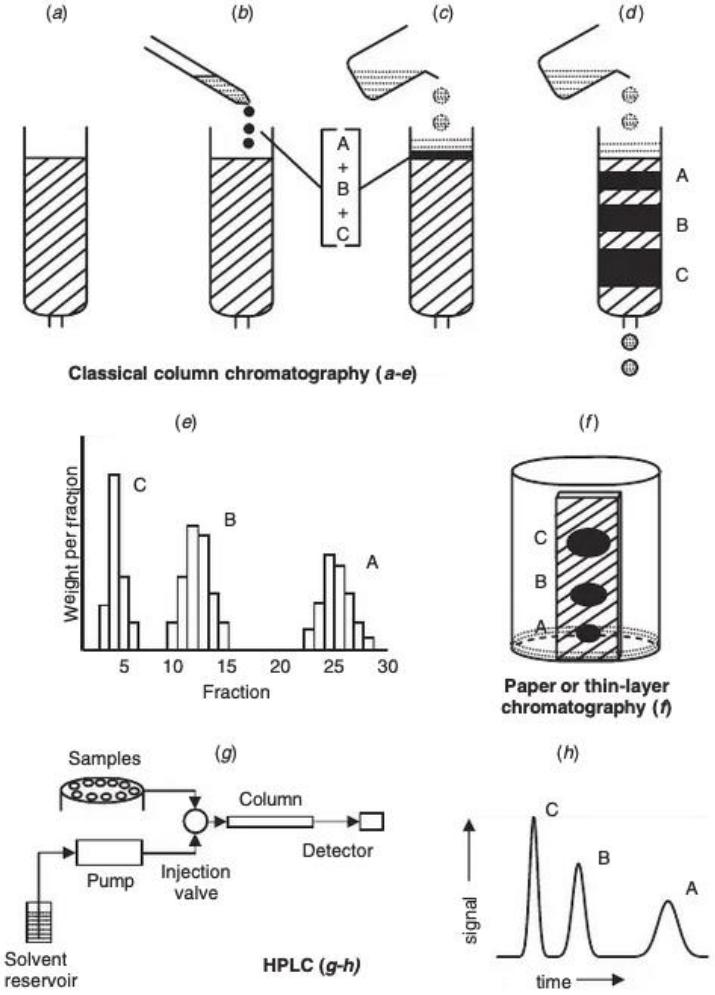


Figure 28 – Représentation schématique d'un processus d'élution en chromatographie, pour l'ensemble des principes chromatographiques (sur colonne, sur support ou en chromatographie en phase liquide). (a)-(d) représente schématiquement le processus d'élution étape par étape, du dépôt à la séparation effective sur la phase stationnaire. (e) représente la quantification massique de composés présents dans la fraction récupérée en sortie de colonne, (f) représente l'application à la chromatographie sur couche mince et (g)-(h) représentent la chromatographie en phase liquide et le résultat obtenu après détection. Adapté de [36]

Comme le composé est en équilibre entre la phase mobile et la phase stationnaire, on peut définir une constante d'équilibre tel que suit :

$$K = \frac{[C]_{\varphi_{stationnaire}}}{[C]_{\varphi_{mobile}}}$$

Équation 6 – Définition de la constante d'équilibre d'un composé partagé entre phase mobile et phase stationnaire

Cette constante n'est affectée que par :

- La structure du composé considéré : la présence de groupements acido-basiques (*i.e.* possibilité de liaisons hydrogènes et ioniques), la longueur des chaines carbonées, la présence de cycles aromatiques, l'encombrement stérique…
- La nature de la phase stationnaire : mêmes considérations que pour la structure du composé considéré ;
- La nature de la phase mobile : pouvoir solvant sur le composé, différence de nature avec la phase stationnaire ;
- La température : changement de la constante de solubilité du composé considéré avec la phase mobile et la phase stationnaire.

Ainsi, choisir une colonne nécessite d'avoir une connaissance, ou au moins une idée, du type de composés à séparer dans l'échantillon à analyser. En effet, pour obtenir une séparation des composés, ceux-ci doivent pouvoir interagir avec la phase stationnaire. Plus un composé a d'affinité avec la phase stationnaire, plus il sera retenu sur la colonne et sera donc élué à un temps de rétention élevé. Cependant, pour que cela fonctionne, il faut que l'interaction avec la phase mobile soit suffisamment petite pour que le composé soit retenu sur la colonne, mais suffisamment importante pour que le composé puisse finalement être élué hors de la colonne. Ainsi, on choisit traditionnellement un mélange de solvant dont on fait varier le pouvoir éluant lorsque l'on effectue une séparation avec un gradient d'élution, ou dont on ajuste le pouvoir éluant pour que l'ensemble des composés soient élués dans de bonnes conditions en mode isocratique.

Des catalogues commerciaux dédiés totalement aux colonnes chromatographiques existent, et il est inconcevable de décrire l'ensemble des variations de rétention que les colonnes peuvent apporter. Néanmoins, on peut décrire quelques grandes catégories de séparations, ainsi que les types de rétention habituellement observés sur ces colonnes :

- Les colonnes de chromatographie d'adsorption : les molécules sont directement adsorbées à la surface de la phase stationnaire. Ce type de chromatographie est très utilisé en chromatographie en phase gazeuse, et on trouve en phase liquide les colonnes à particules de silices pures. Ce type de séparation est utile pour séparer des composés apolaires à peu polaires.
- Les colonnes de chromatographie de partage : la phase stationnaire est une phase de silice ou polymérique, greffée avec des composés organiques divers et variés. Un mélange de solvant avec différentes polarités, miscibles, tels que Acétonitrile/Eau, Méthanol/Eau ou autres sont utilisés pour ces colonnes. L'un des deux solvants, en fonction de la polarité de la phase stationnaire forme alors une couche fine autour des greffons, créant une couche où un équilibre liquide/liquide peut se former. Les composés sont alors élués selon leur affinité avec la phase stationnaire et leur affinité pour le solvant qui l'entoure. Du fait de la variation théoriquement infinie des greffons organiques possible, de nombreuses références de colonnes existent, depuis les classiques colonnes C18 aux colonnes plus exotiques telles que les colonnes cyano ou phényl. Ce type de chromatographie est utilisé pour presque toutes les applications où l'on sépare des molécules organiques.
- Les colonnes de chromatographie d'échange ions : la phase stationnaire présente des groupements ioniques, à tous pH ou non. La séparation s'effectue alors principalement avec l'interaction ionique entre composé et phase stationnaire.

En fonction de la colonne et de l'objectif cherché, *i.e.* chromatographie analytique ou préparatrice, la rétention peut être dynamique ou totale, nécessitant alors des étapes de post-traitement pour éluer les composés retenus sur la colonne. Ce type de chromatographie est par exemple utilisé en hydrologie pour quantifier les ions présents dans les eaux.

- Les colonnes de chromatographie d'exclusion : la phase stationnaire est un gel présentant des pores plus ou moins gros. Les molécules qui ne passent pas par les pores sont éluée plus rapidement que les celles qui sont capables de migrer dans les pores de la phase stationnaire. Ce type de chromatographie est par exemple utilisé en chimie des polymères, pour caractériser la dispersion en taille des polymères par exemple.

1.4.3. Calcul des paramètres et impact des conditions chromatographiques

La chromatographie est une technique complexe qui est influencée par de multiples facteurs. Néanmoins, pour caractériser une méthode, seules trois grandeurs sont nécessaires et suffisantes : le facteur de rétention k, le facteur de sélectivité α, et l'efficacité N. Ces grandeurs sont définies mathématiquement dans la section 2.2. D'autres grandeurs peuvent être utilisées telle que la résolution, mais ce sont alors des combinaisons des trois facteurs précédemment cités.

Le développement d'une méthode chromatographique peut alors s'effectuer en déterminant et modifiant les trois facteurs calculés dans le but de comparer les méthodes et modifications et évaluer quelles conditions permettent de s'approcher des objectifs fixés lors de la définition analytique attendue. Il est alors nécessaire de définir la liste des paramètres chromatographiques pouvant être modifiés afin de savoir quelles conditions changer pour obtenir le résultat désiré. On peut également tenter de quantifier l'influence de leur modification sur chacun des trois paramètres calculés, et ce travail a été effectué par [36] et est adapté dans le Tableau 1.

Conditions	k	α	N
%B	++	+	-
Solvant B (acétonitrile, méthanol, etc.)	+	++	-
Température	+	+	+
Type de colonne (C18, amino, cyano, etc.)	+	++	-
pH de la phase mobile [a]	++	++	+
Concentration du tampon [a]	+	+	-
Force ionique [a]	++	++	+
Longueur de colonne	0	0	++
Taille des particules	0	0	++
Débit de phase mobile	0	0	+
Pression	-	-	+ [b]

Tableau 1 – Effet des conditions chromatographiques sur les paramètres calculés k, α et N. Une valeur ++ indique un impact majeur, + un impact mineur, - un très faible impact tandis que 0 indique un non-impact. Les valeurs en gras et rouge indiquent les conditions préférentiellement modifiées pour contrôler la valeur du paramètre correspondant. [a] Valable uniquement pour des composés ionisables (acides ou bases). [b] La pression en elle-même n'a que peu d'effet sur l'efficacité N. Cependant, par le choix de conditions pertinentes, des pressions plus élevées vont générer de meilleures séparations.

Un développement chromatographique va alors viser à contrôler la séparation des composés par la modification des conditions chromatographiques susmentionnées. Du fait du nombre important de conditions et de leur impact multiple sur la séparation chromatographique, il est nécessaire de développer une procédure de développement systématique qui permet de tester l'ensemble des conditions et de déterminer un jeu de conditions qui permet de réaliser la qualité de séparation désirée.

1.5. Traitements de données chromatographiques

1.5.1. Échantillons complexes et chromatographie

Le traitement des données en chromatographie doit être séparé, comme pour d'autres techniques, en deux catégories : les analyses ciblées et non-ciblées. Ces deux types d'informations souhaitées ont un impact majeur sur les procédures de traitement des données puisqu'elles ne vont pas faire appel aux mêmes algorithmes de traitement. Ainsi, une analyse ciblée repose sur une liste d'information à aller chercher dans le chromatogramme, alors qu'une analyse non-ciblée devra rendre compte de l'ensemble de l'information présente sans ajouter de biais autre que celui déjà apporté par la méthode.

Les analyses ciblées sont largement utilisées en analyse chimique, et l'ensemble des logiciels de contrôle des chaines chromatographiques sont équipés de moyens d'effectuer ces analyses automatiquement et en routine, comme par exemple le logiciel Chromeleon™ de Thermo Scientific qui permet de générer automatiquement des rapports d'analyse en se basant sur les informations fournies préalablement pour détecter, identifier et quantifier les signaux présents dans les données.

Les analyses non-ciblées sont l'enjeu majeur des analyses d'échantillons complexes. En fonction du type d'échantillon, des solutions logicielles sont disponibles pour effectuer le traitement des données : ce sont des solutions de traitement de données dites « supervisées ». En effet, si l'on prend l'exemple de la protéomique, les échantillons biologiques sont des échantillons réputés complexes du fait de la diversité et la complexité des molécules présentes. Du fait de la connaissance préalable des types de molécules présentes, ainsi qu'un certain degré d'invariance dans le type de résultat attendu, il est dès lors possible de superviser les recherches non-ciblées pour rendre un résultat global qui serait la composition en protéines de l'échantillon. La supervision ici est dans la connaissance du fait que les données sont des protéines, et que ce qui n'est pas une protéine n'est pas à considérer. A la différence des analyses ciblées, seule la connaissance du type de molécules attendues est ici introduite pour superviser les traitements, non une liste de composés à identifier comme on doit fournir pour une analyse ciblée. Des logiciels commerciaux existent et sont spécifiques à chaque science qui nécessite ce genre d'analyse, avec des logiciels divers et variés dédiés aux différentes sciences omiques permettant de fournir les résultats attendus par ces disciplines.

Effectuer des analyses non-ciblées et non supervisées est en revanche un problème qui n'a pas de logiciel commercial performant et reconnu par la communauté. Diverses solutions logicielles existent, une des plus connues étant MZmine 2 [37]. Cependant, chaque scientifique a un besoin particulier et développer un logiciel versatile capable de répondre à l'ensemble des demandes et attentes génère finalement une boite noire avec des options pléthoriques dont il est complexe de trouver quelles options sont critiques pour notre application particulière. Cependant, ces différentes suites sont principalement basées sur les mêmes principes : (1) traitement initial du bruit, (2) alignement des spectres, (3) détection des signaux et (4) génération des tables de données et graphiques nécessaires à la compréhension humaine. Les trois premières étapes sont critiques pour le traitement des données, et leur maitrise et compréhension totale est nécessaire pour s'assurer que le traitement de données effectué n'est pas entaché d'erreur et de biais de traitement introduit par une mauvaise configuration.

Plusieurs essais de traitement de données ont été effectués avec MZmine, et plusieurs limitations critiques sont présentes et n'ont pas réussies à être mitigées : (1) algorithmes de traitement opaques, (2) algorithme de détection des pics peu robuste, (3) représentation des données limitée et (4) attribution des formules stœchiométriques qui ne respectent aucune règle chimique.

1.5.2. Objectifs du logiciel à développer

Il apparait nécessaire, au vu des limitations discutées précédemment, de développer notre propre solution logicielle qui permet de maitriser l'ensemble de la chaine de traitement et

d'avoir une liberté de représentation des données qui ne soit pas limitée par une solution logicielle fermée. Il est ainsi décidé de développer cette solution sous Attributor, qui utilise IGOR PRO comme logiciel de programmation et de représentation des données. À la différence des solutions logicielles disponible, le code est accessible à tout moment, sans compilation ou génération de fichier exécutable. De plus, IGOR PRO étant également un logiciel de visualisation de données, il est facile d'effectuer n'importe quelle représentation des données directement dans le logiciel de traitement, et ne nécessite donc pas de logiciels tiers pour effectuer les représentations à partir des données générées.

Les objectifs de développement du logiciel sont donc les suivants :

- Traitement du bruit et alignement des spectres
- Création d'une carte *m/z* en fonction du temps
- Détection et modélisation des signaux

Les fonctions relatives à la génération des formules stœchiométriques et aux éventuels traitements spectre par spectre sont déjà incluses dans Attributor et n'ont donc pas à être développées.

1.6. Conclusion

Dans ce chapitre, nous avons présenté l'intérêt d'effectuer des expériences d'astrophysique de laboratoire, et discuté le choix d'échantillons utilisés pour effectuer les développements et essais analytiques présentés dans ces travaux.

Par la suite, après avoir rappelé que la spectrométrie de masse et la chromatographie s'intègrent dans un processus analytique global, nous avons présenté le fonctionnement d'un Orbitrap et d'un FT-ICR, deux spectromètres de masses à haute résolution. Leurs points communs et différences ont également été discutés dans un but de comparaison et de mise en évidence de leurs limites respectives. Les différentes sources d'ionisations possibles ont alors été introduites, avant de rentrer dans les détails de la source ESI et de la source LDI, toutes deux utilisées pour l'acquisition des données utilisées dans ce travail. Quelques notions générales à mettre en évidence relatives au traitement des données en spectrométrie de masse ont enfin été présentées et illustrées.

Puis, le principe général de la chromatographie est présenté, ainsi que les principaux paramètres d'intérêt qui peuvent être utilisés pour le développement en chromatographie, ainsi que l'impact de la variation des conditions chromatographiques sur ces paramètres. Pour terminer, une discussion est effectuée sur les logiciels de traitement chromatographiques disponibles, ouvrant sur une définition des objectifs de développement de notre propre logiciel de traitement des données issues de la chromatographie.

2. Développement de méthodes de mesures

On vient de définir les grands axes théoriques relatifs à la spectrométrie de masse, à la chromatographie et à leur traitement de données spécifiques. Leur mise en place pratique doit désormais être effectuée avec méthode. La méthode est importante en sciences : c'est-ce qui permet de s'assurer de la cohérence des analyses effectuées et donc, finalement, de la pertinence des résultats et conclusions avancées. Dans un laboratoire de planétologie, le personnel utilisateur des machines n'est pas forcément un chimiste spécialisé en sciences analytiques, mais a besoin des informations délivrées par ces instruments pour étudier les objets tels que météorites ou échantillons de synthèses issus d'expériences d'astrophysique de laboratoire. Le développement de systématiques pas par pas permet donc à un utilisateur non-spécialiste de pouvoir utiliser les instruments, traiter les données et de pouvoir éventuellement proposer de nouvelles méthodes adaptées à ce qu'il souhaite observer.

Dans ce chapitre, trois grands axes sont explorés : les méthodes en spectrométrie de masse, les méthodes en chromatographie et les questions d'identification moléculaire. On va décrire d'abord comment les méthodes d'acquisition des données en ESI-Orbitrap ont été revues à l'aide d'une étude systématique dont les résultats ont été publiées dans une revue à comité de lecture. Puis, on propose une méthode de validation des données et attributions issues de données acquises en ESI-Spectrométrie de masse pour les échantillons de synthèse et/ou présentant des tendances polymériques. Les données issues de LDI-ICR présentant des ions radicalaires, on propose également une méthode d'attribution adaptée prenant en compte ce type de signaux.

Dans un second temps dédié aux méthodes chromatographiques, nous proposons une systématique de développement de méthode pas à pas pour la chromatographie basée sur l'expérience passée en laboratoire au cours de ces années de thèse et avec l'encadrement d'une stagiaire de M1. Cette méthodologie se veut volontairement hybride entre description textuelle, récapitulatif dans un tableau et représentation graphique. Ces différentes façons de présenter l'étape de développement permet d'avoir plusieurs niveaux de compréhension et d'explications disponibles en fonction du besoin de l'utilisateur. Puis, une fois la méthodologie de développement détaillée, le développement de deux méthodes chromatographiques sont expliquées, depuis le choix de la colonne et des conditions chromatographiques jusqu'à leur validation avant utilisation en routine.

Enfin, identifier des molécules en chromatographie et spectrométrie de masse requiert l'analyse de standards connus. Cependant, du fait de la diversité moléculaire observée, plusieurs dizaines voire centaines de standards sont nécessaires, et ce pour chaque formule stœchiométrique attribuée. Dès lors, il convient de réduire la liste des possibles, et on propose alors un outil de prédiction des temps de rétention. Dans un premier temps, on discutera alors de la théorie de la prédiction des temps de rétention, puis nous discuterons des résultats pratiques effectués avec les méthodes chromatographiques développées.

2.1. Méthodes en spectrométrie de masse

Attribuer un spectre de masse est un processus classique sous Attributor. Le choix des groupes et leur limitation en nombre permet le plus souvent d'obtenir des attributions représentatives des données acquises. Cependant, la résolution en masse n'est pas suffisante pour résoudre des permutations atomiques ayant des masses très proches, et nécessitant alors des résolutions superlatives, non atteignables sur un Orbitrap classique, et nécessitant des FT-ICR à très haut champ pour y parvenir. Il y a alors une nécessité de développer une méthodologie permettant de valider une attribution en se basant sur la connaissance de ce type d'échantillon complexe, synthétique et réputé polymérique.

À l'autre bout de la chaine analytique, les méthodes d'acquisitions sont une étape cruciale de ce processus : elles permettent de s'assurer que les analyses effectuées peuvent être

effectuées de nouveau dans les conditions les plus similaires possibles, mais aussi de pouvoir comparer les résultats d'échantillons différents obtenus avec des méthodes similaires ou alors de pouvoir lier ou étudier d'éventuelles variations des résultats à des variations des protocoles d'acquisitions. Du fait du paramètre critique de ces protocoles, une attention particulière doit également être portée à leur optimisation dans le but d'obtenir les meilleures données possibles à partir d'un échantillon.

2.1.1. Optimisation de l'acquisition des données Orbitrap

En plus des paramètres instrumentaux classiques, tels que les réglages de la source et les réglages des optiques ioniques, la longueur des transients et leur nombre est théoriquement un paramètre critique permettant d'ajuster la résolution et la sensibilité d'un instrument à transformée de Fourier. Dans le cas de l'Orbitrap, seulement le nombre de transient est accessible de façon indirecte à travers le nombre de micro-scans effectués, et la résolution est un paramètre fixé à 100 000 par défaut dans toutes nos analyses.

Thermo Scientific, du fait de la non-accessibilité du transient et du traitement automatique du signal effectué, donne accès à seulement deux paramètres : le nombre de scans et le nombre de micro-scans constitutif d'un scan. Par défaut, le nombre de micro-scans est fixé à 1. Ainsi, une séquence analytique normale en Orbitrap se déroule comme présenté en Figure 29.

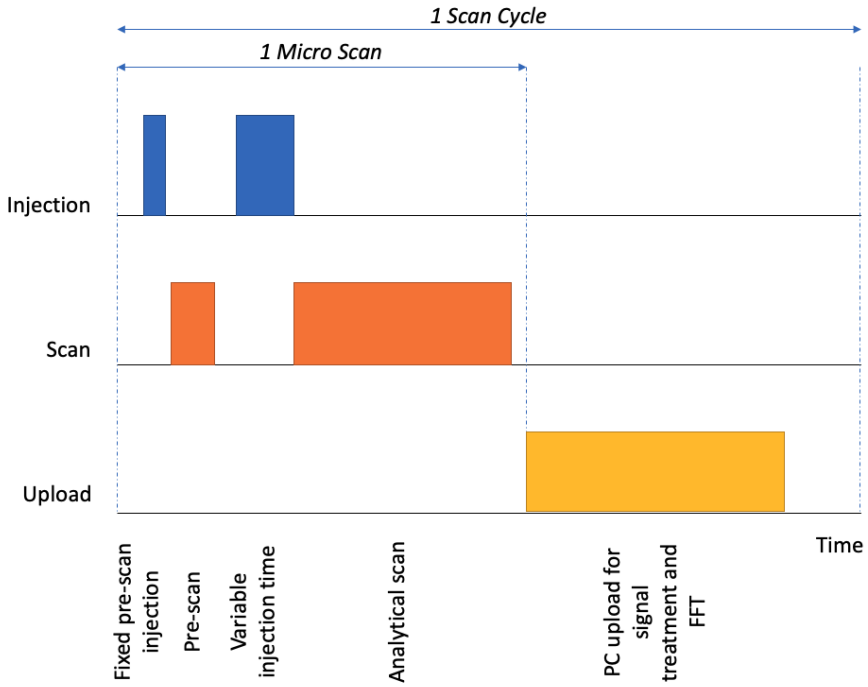


Figure 29 – Échelle de temps pour un scan en Orbitrap. Reproduit et adapté des documents de formation fournis par Thermo Scientific lors de l'installation de la machine.

Le terme « Injection » regroupe le temps de stockage des ions dans la C-Trap ainsi que l'injection effective de ces ions dans la trappe orbitale. Une injection initiale et un pré-scan sont effectués pour déterminer le temps de stockage optimal des ions pour effectuer l'analyse : le temps d'IT (Ion Time) est ainsi fixé. Les ions sont ensuite injectés durant le temps déterminé précédemment, et le scan analytique est effectué. Cet enchainement d'actions détermine un micro-scan. Il est possible d'ajouter plusieurs micro-scans à la suite, de quelques-uns à plusieurs centaines voire quelques milliers. La limite dans ce cas est la quantité l'échantillons disponibles dans la seringue puisqu'il n'est pas possible de mettre en pause l'acquisition pour recharger la seringue en cours d'analyse.

Historiquement, à l'IPAG, avant toute optimisation et analyse systématique à ce sujet, il avait été déterminé de façon empirique que 4 scans de 128 micro-scans donnaient de meilleurs résultats que l'équivalent de 512 scans de 1 micro-scans. Cette qualité supérieure est basée sur

moins de bruit et plus de molécules attribuées en général. Cette observation est à l'origine de l'analyse systématique effectuée en vue d'optimiser le protocole d'acquisition utilisé en routine en Orbitrap.

Ces travaux d'optimisation ont fait l'objet d'une publication [38] dans *Rapid Communication in Mass Spectrometry*, DOI 10.1002/rcm.8818. La version acceptée est reproduite ci-après.

Wolters Cédric (Orcid ID: 0000-0002-9710-4740)

Enhancing data acquisition for the analysis of complex organic matter in direct-infusion Orbitrap mass spectrometry by using micro-scans

Cédric Wolters[1*], Laurène Flandinet[1], Chao He[2], Junko Isa[1], François-Régis Orthous-Daunay[1], Roland Thissen[3], Sarah Hörst[2], Véronique Vuitton[1]

[1]Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
[2]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA
[3]Université Paris-Saclay, CNRS, Institut de Chimie Physique UMR8000, 91405, Orsay, France

*Author for Correspondence : cedric.wolters@univ-grenoble-alpes.fr

Abstract

RATIONALE: Acquisition quality in analytical science is key to obtaining optimal data from a sample. In very high-resolution mass spectrometry, quality is driven by the optimization of multiple parameters, including the use of scans and micro-scans (or transients) for performing a Fourier transformation.

METHODS: 39 mass spectra of a single synthesized complex sample were acquired using various numbers of scan and micro-scan determined through a simple experimental design. An electrospray ion source coupled with an LTQ-Orbitrap-XL mass spectrometer was used and acquisition was performed using a single mass range. All the resulting spectra were treated in the same way to enable comparisons of assigned stoichiometric formulae between acquisitions.

RESULTS: Converting the number of scans into micro-scans enhances signal quality by lowering noise and reducing artifacts. This modification also increases the number of attributed stoichiometric formulae for an equivalent acquisition time, giving access to a larger molecular diversity for the analyzed complex sample.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/rcm.8818

CONCLUSION: For complex samples, the use of long acquisition times leads to optimal data quality, and the use of micro-scans instead of scans-only maximizes the number of attributed stoichiometric formulae.

Introduction

Electrospray ionization Orbitrap mass spectrometers have been used for more than ten years for high resolution mass spectrometry analysis, either in direct-infusion or coupled to gas (GC) or liquid (LC) chromatography. Application areas include the analysis of natural and synthetic complex samples in "omics" sciences such as metabolomics (1,2), proteomics (3–6) or lipidomics (7,8), and the analysis of environment complex samples such as groundwaters (9,10) and atmospheric aerosols (11–13). High mass resolution is required for each of these fields of study due to its ability to reveal the molecular diversity of a sample by attributing unique stoichiometric formulae to detect signals.

For all these complex samples, direct-infusion Ion Cyclotron Resonance (ICR) mass spectrometers are unmatched as far as resolution is concerned, allowing the identification and attribution of mass separation values below an electron mass. On the other hand, Orbitrap mass spectrometers are mainly used due to their versatility, scan rate suitable for LC acquisition, and easy to setup parameters. Various applications use direct-infusion Orbitrap analysis especially in environmental science (14), nuclear physics (15) or planetary sciences (16–18) for the analysis of meteorites and astrophysical laboratory analogues. A development project for an Orbitrap is currently under way, the CosmOrbitrap (19–21), aiming at allowing the use of direct-infusion mode with a laser desorption ionization source onboard spacecraft to analyze molecular diversity in-situ on non-terrestrial objects.

Data quality is important in analytical techniques. This particular term could refer to multiple concepts, from the acquisition process and validation procedures (22) to the structure of the data itself. In this study, we focus on the second part by assessing that the sample can be attributed properly (resolution, overall signal intensity, artifacts) and that the attribution is representative of the sample. In order to have a reproducible spectrum with high signal to noise ratio, we need to run instrumental methods that increase accuracy, precision and sensitivity. While it is well-known that using transient averaging before performing the Fourier transformations in FT-ICR instruments gives high sensitivity and mass precision and accuracy, this option – called "micro-scans" by Thermo Scientific – is poorly known for direct-infusion Orbitrap mass spectrometry. The Fourier transformation is a linear mathematical application. Thus, doing the sum of Fourier transformed mass spectra is equivalent to doing the Fourier transformation of the summed transients. Given this, performing micro-scan acquisitions should be the equivalent of standard scan acquisitions. Nevertheless, for Orbitrap mass spectrometry, Thermo Scientific software performs the Orbitrap specific noise removal just before applying the Fourier transformation. This specific operation combined with the limited acquisition time nullifies the linearity of the Fourier transformation application, generating a non-equivalence of the Fourier transformation application depending on if you apply it on summed transients or on the individual transients.

Only a few works mention micro-scans for LC-coupling and mainly during MS/MS acquisitions (4–6,15). This specific parameter is also not advertised by Thermo Scientific on instrument specifications and technical sheets. In this work, we will focus on using a complex organic sample available in our lab to investigate the impact of micro-scans and scans on data quality and the number of attributed stoichiometric formulas.

Methods

Sample synthesis

We use a laboratory analogue of Titan's atmospheric aerosols, called tholin, which is representative of a complex organic sample with several thousands of stoichiometric formulae. This analogue was synthetized to mimic the formation of aerosols in Titan's atmosphere by using a mixture of N_2, CH_4, and CO in proportions of 94.98%:5%:0.02% (%vol). The apparatus, sample synthesis, and sample recovery have been described extensively by He et al (23). Several mass spectrometry studies has been carried out on this kind of sample and they have revealed the presence of rich and diverse organic matter diversity (24–28). The Orbitrap mass spectra interpretation of this specific sample is ongoing and will not be discussed extensively in this paper.

Sample preparation

1 mg of the sample was dissolved in 1mL of methanol (Carlo Elba, Milan, Italy; UPLC grade) using an Eppendorf (Hamburg, Germany) pipette and a 1.5-mL polypropylene plastic tube. The solution was vortexed for ten minutes and then centrifugated for ten minutes at 9400 G. 500 µL of the supernatant was dissolved with 500 µL fresh methanol in a new 1.5-mL polypropylene plastic tube.

This preparation was performed three times to obtain all the required volumes for carrying out all the analyses.

Instrumentation

Analyses were performed with an LTQ Orbitrap XL™mass spectrometer equipped with an IonMax™ ESI source (Thermo Scientific, Bremen, Germany). This hybrid mass spectrometer uses an ion trap coupled to the Orbitrap cell. A C-trap is placed between the two traps to store ions coming from the source and inject them in packets directly into the Orbitrap cavity. The MS functions were controlled by LTQTune software (Thermo Scientific). Data treatment and visualization were performed with XCalibur software (Thermo Scientific) and then with a home-made software package "Attributor" (29) developed with IGOR Pro (WaveMetrics, Portland, OR, USA).

For accurate mass calibration, a mixture of Caffeine, Met-Arg-Phe-Ala peptide (MRFA) and Ultramark 1621 (a mixture of fluorinated phosphazine polymers) was used (Thermo Scientific). This calibration mixture is the Thermo Scientific default calibration mixture for LTQ-Orbitrap-XL instruments.

All analyses were performed at the maximum instrument resolution (resolution better than 100,000 at *m/z* 400; mass accuracy of ±2ppm) using the AGC setting of 5.10^5 ions storage. For the study needs, we used an intermediate mass range, from *m/z* 150 to 450 as it includes the highest intensity distribution and allows a broad attribution with limited resolution issues at high masses. The sample flow rate was set at 3 µL/min and led to a mean sample storage time around 10 ms before injection into the Orbitrap. The ESI settings were 3.5 kV for source voltage, 5 (arbitrary units) for the sheath gas flow rate and 0 (arbitrary units) for the auxiliary and sweep gas flow rates. The capillary temperature was set at 275°C, the capillary voltage at 34 V, and the tube lens voltage at 70 V. All other instrument parameters were checked and transferred from the automatic instrument calibration. Micro-scans and scans were defined before each acquisition depending on the experiment needed, as described in the paragraph "Experimental design".

Micro-scans

The term micro-scans refers to the free parameter that can be changed inside the MS configuration panel in the LTQTune software. From the "LTQ-Orbitrap Operation Training Course Manual" (European Training Institute, Thermo Scientific). Several descriptions are given of the acquisition sequence used by the Orbitrap:

- ion time (IT) is "the time in milliseconds that ions are allowed to accumulate in the trap. The IT time is variable and is calculated by the AGC."
- A micro-scan is "a complete cycle combination of the prescan and the analytical scan which will include the IT time (trapping and cooling of the ions) and scan out time (time taken to eject the ions)."
- A scan is the "time taken to complete the overall experiment and is dependent upon the number of micro-scans used."

At the completion of the process, the complete scan is the combination of one or several micro-scans. Each complete/combined scan includes one Fourier transformation. In the end, carrying out micro-scans or the equivalent scans-only operation results in roughly the same overall acquisition time.

The process of using transients is well-known for FT-ICR instruments where transients (micro-scans for Orbitrap) are averaged before the Fourier transformation to produce the final spectra (30,31). Unfortunately, the Orbitrap does not give access to the transient signal nor the Fourier transformation process as opposed to ICR instrument where the complete process is open access and can be used to generate the mass spectra through dedicated signal treatments. It is also common in FT-ICR direct-infusion experiments to acquire data over a long time (28,31) using a high number of transients to increase the data quality in direct infusion, whereas this process seems seldom used for Orbitrap direct-infusion measurements.

Experimental design

To investigate the impact of scans and micro-scans, we designed a set of experiments to be able to obtain maps of numbers of stoichiometric formulae. As we aimed to investigate the impact of changing from short to long duration acquisitions and wanted to limit the number of resulting experiments, we designed an experimental plan based on asymmetric 2^n*2^m maps with n=[0;3] the number of scans (i.e. 1, 2, 4 and 8 scans) and m=[0;10] the number of micro-scans (i.e. 1, 2, 4, 8, 16, …, 1024 scans). This resulted in 38 individual acquisitions varying from 1 to 8 scans and from 1 to 1024 micro-scans. The reference analysis was chosen to be outside the map with 1024 scans using 1 micro-scan each. In terms of time, acquisitions varied from seconds to around 30 minutes.

The repeatability and reproducibility of the acquisition are also important factors that are not addressed by the experimental design realized for this study. A study on similar samples has shown 1% repeatability and 2% reproducibility of direct-infusion analyses. The repeatability and reproducibility were evaluated on a number of identical attributed stoichiometric formulae across the same sample prepared and acquired at different moments (hours, days and weeks). Variations are mostly observed at very low intensity and are due to signal to noise variation and slight differences in sample preparation as several solutions are done. Thus, we do not expect repeatability and reproducibility to have an impact in this study.

Data treatment

A home-made software package was used to extract data from the Thermo .raw files. The extraction was based on an IGOR Pro routine that uses Windows open libraries to open and read directly the data inside the Thermo-produced .raw files. After data extraction, each scan was summed to produce a single resulting mass spectrum. This mass spectrum was then loaded into Attributor, a home-made mass spectrometry software package specially developed to treat and attribute high-resolution mass spectrometry data (18). Each spectrum was converted from peak profile to centroids, normalized and internally calibrated, and peaks were assigned a molecular formula using the same routines and parameters to be able to compare the results. The internal calibration used CH_2 families with more than 10 members, i.e. series of attributed stoichiometric formulae that differ only by variation of a CH_2 group. In order to select the accurate stoichiometric formulae for the internal calibration, we selected assigned peaks whose stoichiometric formulae only included oxygen and nitrogen - 0 to up to 3. Attributions were performed for singly positive charged ions using the graph properties of the spectra, with CH_2, C, NH and O as propagation parameters. Even if ^{13}C stoichiometric formulae are detected and can be attributed, they are not considered in this study. We would also have to consider the isotopic ratios to explain the absence of ^{13}C between different experiments in addition to variation induced by the acquisition method. The propagation list is then used as is, without any post-processing treatment. Typical errors and stoichiometry in the function of mass to charge ratio representation are available in Supporting Information 1.

Results

The mass spectrum in Figure 1A highlights a striking molecular organization in the sample. We can hence distinguish a periodic pattern where maxima (or minima) are visible, separated by 13.5 u. One of the clusters constituting the pattern is highlighted with a black box in Figure 1B for illustration. This pattern structure is typical of Titan's tholins and indicates that the sample is made through a polymerization process as described in several works such as in Pernot and co-workers (24) and Hörst's thesis (25). They highlight that the major pattern is not only CH_2 as we can expect from complex organic matter but a combined role of CH_2 and HCN in the final structural composition. Given this dual polymerization pattern, each cluster exhibits a constant (C+N) value, with hydrogenation variation generating the cluster diversity (24–27).

Before doing any attribution, we check the spectra for each experiment and particularly if the micro-scans affect the signal intensity. As can be seen in Figure 1A, using the micro-scans function of the Orbitrap seems to enhance the signal after *m/z* 325 compared with the scans-only operation while at the same time, enlarging the cluster distribution at both lower and higher masses (Figure 1B). This enlargement seems to indicate a higher diversity of H content for a given (C+N) content. At first glance, we also see a difference for the minimum signal of each spectrum, at 8.10^{-3} for micro-scans and 2.10^{-3} for scans-only (not shown in the Figure here). This limitation of the dynamical range could appear as a problem for the sensitivity of the analysis, but we will demonstrate later in this work that the dynamic range is not the correct parameter to compare the two operation techniques.

As intensity *vs* m/z spectra do not show much information about a sample's molecular diversity, we use another representation that can highlight sample signals and diversity over artifacts and noise: the mass defect diagram, a variation of Kendrick diagrams using ^{12}C as a reference instead of $^{12}C^{1}H_2$. The mass defect is thus defined, as shown in equation 1, by only using the exact mass measured by the instrument:

$$Mass\ defect = Exact\ mass - round(Exact\ mass) \quad (1)$$

We report in Figure 2 the mass defect diagrams for 1024 scans with 1 micro-scan experiment (left) and 1 scan with 1024 micro-scans experiment (right). We can see different structures and shapes inside these diagrams: (1) vertical lines, (2) cones of close points at positive and negative mass defect values and (3) individual points. The vertical lines are more frequent in scan-only and are attributed to radio artifacts, which are generated by radio wavelengths that are detected by the Orbitrap amplification system. Noise points are generated by classical electronic noise and do not present a periodic pattern. The cone structures are similar in both experiments and represent ions resulting from electrospraying the sample, with the top cone attributed to (mostly) singly-charged ions and the bottom cone to (only) doubly-charged ions.

Despite the similar acquisition time of approximately 30 minutes and the same total number of analytical scans for the analyses shown in Figure 2, all different features indicate differences directly inside the Orbitrap data generation system, depending on the way in which the experiment has been set up. Differences between scans and micro-scans concerning radio artifacts and random noise could be attributed to the way in which the Thermo Scientific software handles the automatic noise removal before doing the Fourier transformation. This process appears to be more efficient when several micro-scans are summed before the noise removal than when the noise removal is applied on each scan. In extreme cases with using scans-only, such as with diluted samples, the random points and radio artifacts can spread on the entire range of mass defect and *m/z* values and hide any true sample signal (see Figure 6 in Danger et al (16) for an example of radio artifacts overlapping the real sample signal). Although this kind of issue can be treated by post-treatment procedures such as noise hard cutting, using micro-scans significantly reduces the need for post-treatment procedures.

We summarize in Figure 3 the total number of attributed stoichiometric formulae per experiment. We observe that for a given number of scans and an increasing number of micro-scans and *vice-versa*, we attribute more and more stoichiometric formulae. Nevertheless, this observation hides some disparities, with an asymmetry in the number of attributed stoichiometric formulae for equivalent acquisition times. Diagonals in the table represent acquisitions carried out during an equivalent amount of time, from a second to 30 minutes. This indicates that for an identical sample, acquiring either scans, micro-scans or a combination of both does not lead to equivalent results, driving us to the search for the best acquisition setup.

There is also interest in checking if the attributed stoichiometric formulae are overlapping between the different analyses. We report in Supporting Information 2 individual attributed stoichiometric formulas on top of the mass defect diagram for each analysis using 1 single scan for 1 to 1024 micro-scans, respectively a second to approximately 30 minutes of acquisition time. As expected, we observe an increasing data complexity with an enlargement of the sample cloud correlated with the enlargement of the attributed stoichiometric formula cloud. This indicates that all attributed stoichiometric formulae at low micro-scans are included in the highest number of micro-scans. We also observe that attributions are strictly contained into the sample cloud and that there is no attribution out of this area, consistent with other points being either artifacts, noise or ions with a higher charge. This observation highlight that the accumulation of scans increases the signal to noise ratio, leading to more detected peaks while increasing acquisition time.

Another interesting point here is that we can observe in Figure 3 an equivalent number of stoichiometric formulae for different acquisition methods. This can be important in the case of samples with low available quantity. The most significant improvement is observed between 1024 scans and 2 scans of 128 micro-scans that attribute the same number of stoichiometric formulae, but with significant variation of acquisition time: 30 minutes and 7.5 minutes, respectively. Thus, using micro-scans allows a reduction of time and volume consumption of four times compared with the scans-only protocol for an equivalent number of stoichiometric formulae.

Aside from the number of stoichiometric formulae, we also observed from the raw mass spectra a difference of dynamic range for methods using micro-scans *versus* methods using scans-only. This can also be seen in Figure 4 that represents the histogram of intensity for stoichiometric formulae, where we have a hard intensity cut around 10^{-2} using 1024 micro-scans whereas we observed intensities down to roughly 3.10^{-3} when using 1024 scans. Even if the apparent dynamic range is lower using micro-scans than using scans-only, comparing identical stoichiometric formulae between the two operations leads to the reverse conclusion. When we compare for identical stoichiometric formulae between the two methods, all identical molecules are higher than 5.10^{-1} while all detected molecules go down to 10^{-2} using micro-scans. Thus, the dynamic range is not a parameter to consider when comparing scans with micro-scans operations as the intensity of identical stoichiometric formulae is not consistent between the two operations.

When focusing on identical stoichiometric formulae between the two experiments, more than 97% of the stoichiometric formulae attributed using 1024 scans are found in the stoichiometric formula attributions using 1024 micro-scans, corresponding to 2164 identical stoichiometric formulae in both datasets. A careful inspection of the remaining 3% (i.e. 41 stoichiometric formulae) indicates that they are wrong stoichiometric formulae that would be removed using a manual post-treatment work on stoichiometric formulae (for instance too high N/C or O/C ratio) and some low intensity attributed points that can be assigned to an unfortunately badly positioned noise point. The intensity comparison of the identical stoichiometric formulae shows that we are attributing more than 1300 new mid to low intensity stoichiometric formulae using micro-scans in addition to all the stoichiometric formulae detected using 1024 scans, represented in Figure 3 by the red bars. Of the new attributed molecules, some could also be misattributed and would have been removed by a manual post-treatment step, estimated to be less than 100 signals.

There are several ways to represent stoichiometric formulae to compare data. Stoichiometric ratio representation has been considered such as H/C; N/C; O/C as a function of the ratios or the mass, but the resulting information is not conclusive as the information is stretched and/or non-linear in most of the cases. We report here that the Double Bond Equivalent (DBE) as a function of the number of carbon, nitrogen, oxygen and hydrogen representation can be used to highlight new attributed stoichiometric formulae, as presented in Figure 5. In addition to identical stoichiometric formulae, we observe that the micro-scans operation detects a wider diversity with the addition of low DBE, but also detects new stoichiometric formulae at all DBE values. This is consistent with observations made on the mass spectra, where all the peak clusters inside mass spectra are denser using micro-scans than when using scans-only as discussed for Figure 1B. This can be also explained by an increase of the signal to noise ratio using micro-scans compared with the scan-only method.

Discussion and Conclusion

In the framework of analyzing complex samples such as environmental or biological samples with several thousands of molecular formulae, it is necessary to accumulate for long enough to have the maximum possible number of formulae at the end. Nevertheless, the way used to accumulate data has a direct impact on the number of formulae obtained. Thus, converting the number of scans to micro-scans will result in the best possible acquisition methods and the maximum number of attributed stoichiometric formulae at the end. This way of using micro-scans over long periods instead of using scans-only is also identical to the data processing used in classical ICR data-processing (without specific post-treatment) where all transients are summed before the Fourier Transformation necessary to obtain the final mass spectra.

All results have been acquired using a complex organic matter sample. This means that the instrument resolving power and mass accuracy are critical as the number of probable stoichiometric formulae increases with the mass increase. The use of a high resolving power and high accuracy instrument is then mandatory to achieve sufficient separation between masses. In the case of our study, we analyze only CHNO molecules in a mass range where we know tthat he Orbitrap at its highest resolving power can attribute one and only one stoichiometric formula per mass, at the given ±2ppm mass accuracy. Based on this study, the use of micro-scans or scans operation does not affect the resolution nor the mass accuracy of the data. Micro-scans enable the detection of additional low-intensity signals by increasing their signal to noise ratio compared with the equivalent number of scans. The use of micro-scans operation results in cleaner mass spectra, in the sense that the automatic noise removal performed by the Thermo Scientific software before the Fourier transformation is more efficient using micro-scans than using the scans-only method. In the case of simple or pure samples, there is no organic diversity. In that case, the use of a limited time acquisition, i.e. several seconds to a minute using scans-only is sufficient to obtain a good mass spectrum of the analyzed sample.

Acknowledgments

This work is supported by the French National Research Agency in the framework of the Investissements d'Avenir program (ANR-15-IDEX-02), through the funding of the "Origin of

Life" project of the Univ. Grenoble-Alpes and the French Space Agency (CNES) under their Exobiology and Solar System programs. Cédric Wolters acknowledges a PhD fellowship from CNES/ANR (ANR-16-CE29-0015 2016-2021). The paper has been read by a native English speaker and many sentences have been rewritten. Authors acknowledge the reviewers for their feedbacks that helps improve the article.

Bibliography

1. Creek DJ, Jankevics A, Breitling R, Watson DG, Barrett MP, Burgess KEV. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Anal Chem.* 2011;83(22):8703–8710.
2. Guo X, Long P, Meng Q, Ho C-T, Zhang L. An emerging strategy for evaluating the grades of Keemun black tea by combinatory liquid chromatography-Orbitrap mass spectrometry-based untargeted metabolomics and inhibition effects on α-glucosidase and α-amylase. *Food Chem.* 2018;246:74–81.
3. Scigelova M, Makarov A. Orbitrap Mass Analyzer – Overview and Applications in Proteomics. *PROTEOMICS.* 2006;6(S2):16–21.
4. Kalli A, Hess S. Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer. *PROTEOMICS.* 2012;12(1):21–31.
5. Kalli A, Smith GT, Sweredoski MJ, Hess S. Evaluation and Optimization of Mass Spectrometric Settings during Data-dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers. *Journal of Proteome Research.* 2013;12(7):3071–3086.
6. Kilpatrick LE, Kilpatrick EL. Optimizing High-Resolution Mass Spectrometry for the Identification of Low-Abundance Post-Translational Modifications of Intact Proteins. *Journal of Proteome Research.* 2017;16(9):3255–3265.
7. Taguchi R, Ishikawa M. Precise and global identification of phospholipid molecular species by an Orbitrap mass spectrometer and automated search engine Lipid Search. *J Chromatogr A.* 2010;1217(25):4229–4239.
8. Rombouts C, De Spiegeleer M, Van Meulebroek L, De Vos WH, Vanhaecke L. Validated comprehensive metabolomics and lipidomics analysis of colon tissue and cell lines. *Anal Chim Acta.* 2019;1066:79–92.
9. Sun C, Shotyk W, Cuss CW, et al. Characterization of Naphthenic Acids and Other Dissolved Organics in Natural Water from the Athabasca Oil Sands Region, Canada. *Environ Sci Technol.* 2017;51(17):9524–9532.
10. Kim C, Ryu H-D, Chung EG, Kim Y. Determination of 18 veterinary antibiotics in environmental water using high-performance liquid chromatography-q-orbitrap combined with on-line solid-phase extraction. *J Chromatogr B.* 2018;1084:158–165.
11. Parshintsev J, Vaikkinen A, Lipponen K, et al. Desorption atmospheric pressure photoionization high-resolution mass spectrometry: a complementary approach for the chemical analysis of atmospheric aerosols: DAPPI-HRMS for analysis of atmospheric aerosols. *Rapid Commun Mass Spectrom.* 2015;29(13):1233–1241.
12. Riva M, Ehn M, Li D, et al. CI-Orbitrap: An Analytical Instrument To Study Atmospheric Reactive Organic Species. *Anal Chem.* 2019;91(15):9419–9423.
13. Roveretto M, Li M, Hayeck N, et al. Real-Time Detection of Gas-Phase Organohalogens from Aqueous Photochemistry Using Orbitrap Mass Spectrometry. *ACS Earth Space Chem.* 2019;3(3):329–334.
14. Urai M, Kasuga I, Kurisu F, Furumai H. Molecular characterization of dissolved

organic matter in various urban water resources using Orbitrap Fourier transform mass spectrometry. *Water Science and Technology: Water Supply*. 2014;14(4):547–553.
15. Hoegg ED, Barinaga CJ, Hager GJ, Hart GL, Koppenaal DW, Marcus RK. Isotope ratio characteristics and sensitivity for uranium determinations using a liquid sampling-atmospheric pressure glow discharge ion source coupled to an Orbitrap mass analyzer. *J Anal At Spectrom*. 2016;31(12):2355–2362.
16. Danger G, Orthous-Daunay F-R, de Marcellus P, et al. Characterization of laboratory analogs of interstellar/cometary organic residues using very high resolution mass spectrometry. *Geochim Cosmochim Acta*. 2013;118:184–201.
17. Naraoka H, Hashiguchi M, Sato Y, Hamase K. New Applications of High-Resolution Analytical Methods to Study Trace Organic Compounds in Extraterrestrial Materials. *Life*. 2019;9(3):62-73.
18. Orthous-Daunay F-R, Piani L, Flandinet L, et al. Ultraviolet-photon fingerprints on chondritic large organic molecules. *GEOCHEMICAL JOURNAL*. 2019;53(1):21–32.
19. Briois C, Thissen R, Thirkell L, et al. Orbitrap mass analyser for in situ characterisation of planetary environments: Performance evaluation of a laboratory prototype. *Planetary and Space Science*. 2016;131:33–45.
20. Arevalo R, Selliez L, Briois C, et al. An Orbitrap-based laser desorption/ablation mass spectrometer designed for spaceflight. *Rapid Commun Mass Spectrom*. 2018;32(21):1875–1886.
21. Selliez L, Briois C, Carrasco N, et al. Identification of organic molecules with a laboratory prototype based on the Laser Ablation-CosmOrbitrap. *Planetary and Space Science*. 2019;170:42–51.
22. MacDougall, Crummett WB, et al. Guidelines for data acquisition and data quality evaluation in environmental chemistry. *Anal Chem*. 198052(14):2242–2249.
23. He C, Hörst SM, Riemer S, Sebree JA, Pauley N, Vuitton V. Carbon Monoxide Affecting Planetary Atmospheric Chemistry. *The Astrophysical Journal*. 2017;841(2):L31.
24. Pernot P, Carrasco N, Thissen R, Schmitz-Afonso I. Tholinomics—Chemical Analysis of Nitrogen-Rich Polymers. *Anal Chem*. 2010;82(4):1371–8130.
25. Hörst SM. Post-Cassini Investigations of Titan Atmospheric Chemistry. University of Arizona; 2011.
26. Hörst SM, Yelle RV, Buch A, et al. Formation of Amino Acids and Nucleotide Bases in a Titan Atmosphere Simulation Experiment. *Astrobiology*. 2012;12(9):809–817.
27. Somogyi Á, Smith MA, Vuitton V, Thissen R, Komáromi I. Chemical ionization in the atmosphere? A model study on negatively charged "exotic" ions generated from Titan's tholins by ultrahigh resolution MS and MS/MS. *Int J Mass Spectrom*. 2012;316–318:157–163.
28. Maillard J, Carrasco N, Schmitz-Afonso I, Gautier T, Afonso C. Comparison of soluble and insoluble organic matter in analogues of Titan's aerosols. *Earth Planet Sci Lett*. 2018;495:185–191.
29. Orthous-Daunay F-R, Thissen R, Vuitton V. Measured mass to stoichiometric formula through exhaustive search. *Proceedings IAU Symposium S350*. (Accepted). 2019;
30. Kido Soule MC, Longnecker K, Giovannoni SJ, Kujawinski EB. Impact of instrument and experiment parameters on reproducibility of ultrahigh resolution ESI FT-ICR mass spectra of natural organic matter. *Org Geochem*. 2010;41(8):725–733.
31. Blackburn JWT, Kew W, Graham MC, Uhrín D. Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter. *Anal Chem*. 2017;18;89(8):4382–4386.

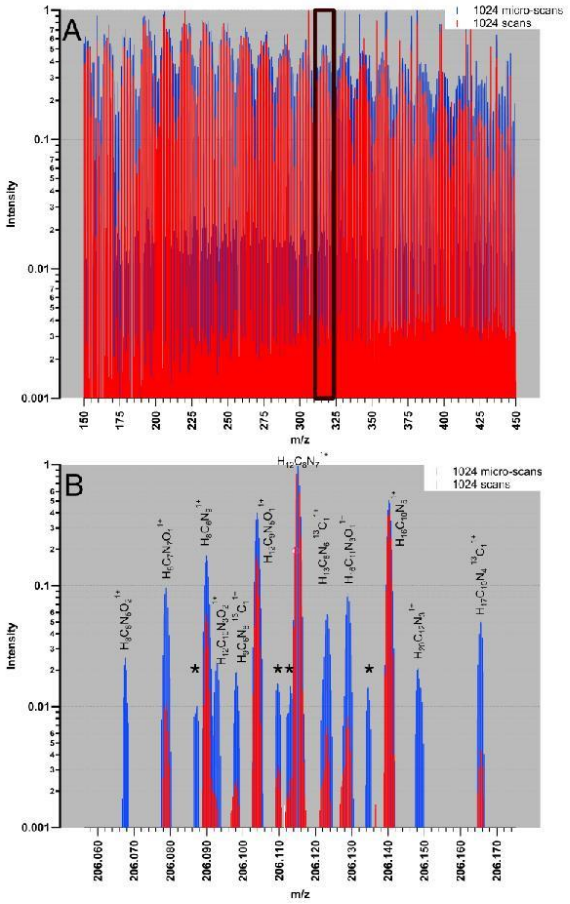


Figure 1: (A) Mass spectrum of peaks profile data for two complementary Orbitrap operations. Data are normalized around *m/z* 206.114. The micro-scans operation of the Orbitrap present a slightly highest intensity trend at high masses. (B) Zoom on the *m/z* 206.1 peak cluster where we observe with micro-scans an added diversity on both the left and the right of the cluster. This particular example is representative of all peak clusters. Detected but non-attributed peaks using CHNO^{13}C are highlighted with a "*". Two ^{13}C are also attributed but are not considered afterward as with every detected ^{13}C stoichiometric formula.

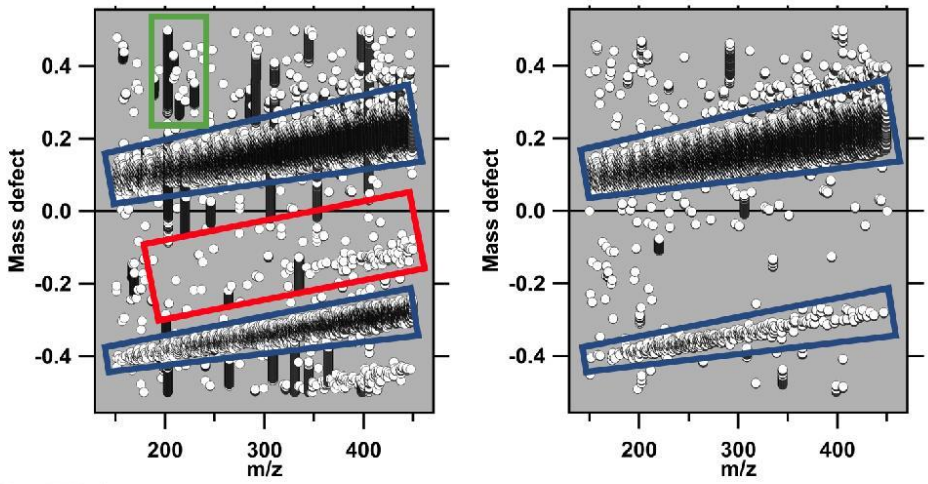


Figure 2: Left – Mass defect diagram from profile data for 1024 scans with 1 micro-scan experiment; 53,238 points. Right – Mass defect diagram from profile data for 1 scan with 1024 micro-scans experiment; 45,198 points. The blue box is sample ions, the green box is a radio artifact example and the red box is an example of random points.

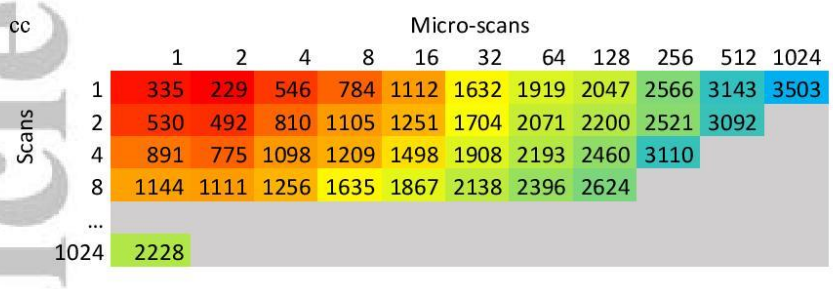


Figure 3: Total number of stoichiometric formulae attributed per analysis; color code from red to blue for the lowest to the highest number. Grey color is an experiment that has not been acquired. The average time per micro-scans or scan is 1.5s. This average time is mostly dependent on the sample concentration and the AGC target.

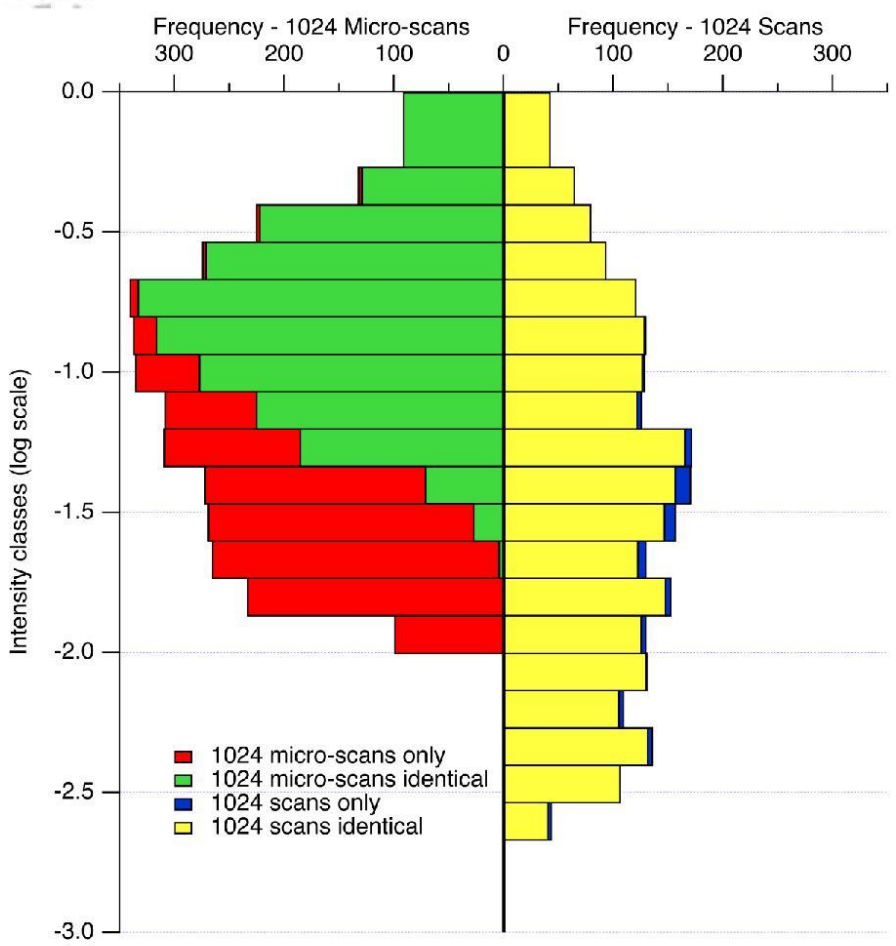


Figure 4: Intensity histogram for attributed stoichiometric formulae using 1024 scans and 1024 micro-scans. Identical stoichiometric formulae between each analysis are indicated in yellow and green, respectively.

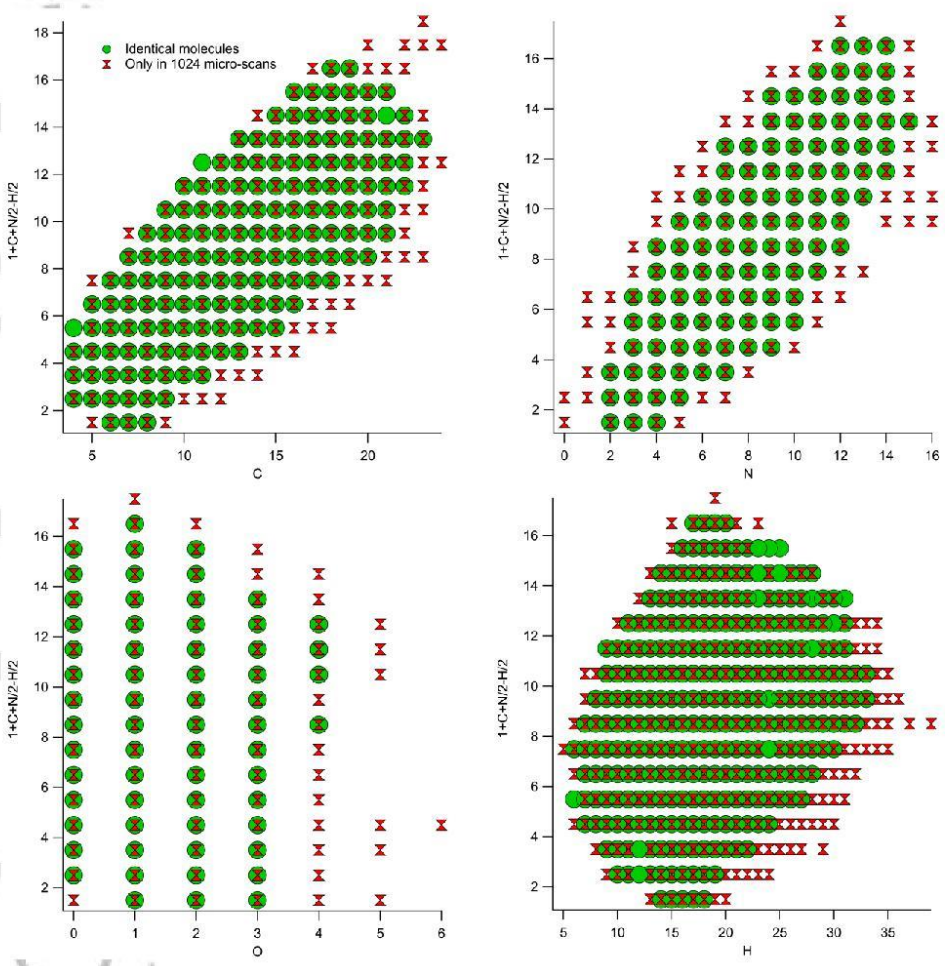


Figure 5: Double Bond Equivalent (DBE) versus the carbon, nitrogen, oxygen and hydrogen number for identical stoichiometric formulae in both operations and formulae only in 1024 micro-scans. When using the micro-scans method, the added diversity is everywhere.

2.1.2. Traitement et validation des données en spectrométrie de masse

L'attribution est un processus en plusieurs étapes qui inclus l'attribution ainsi que toutes les étapes de nettoyage des données. Ainsi, cette étape est cruciale dans le processus de traitement des données en spectrométrie de masse pour s'assurer que la liste des formules stœchiométriques en présence soit représentative de l'échantillon et contienne aussi peu que possible d'artefacts dus au traitement et à la perte de résolution de l'instrument avec la masse. Avec un traitement plus ou moins automatisé des attributions de données, il arrive toujours la question de la validité des résultats présentés et discutés. Cette partie a donc pour effet de proposer une méthode systématique du traitement des données dans le but de s'assurer, dans le cas d'échantillons polymériques, de la complète validité des attributions, depuis la calibration des données jusqu'aux attributions stœchiométriques.

Pour être compréhensible, cette partie va suivre le cheminement de traitement des données d'un échantillon d'analogue d'atmosphère d'exoplanète, composé de carbone, azote, hydrogène et oxygène seulement. Il n'y a nullement besoin d'informations supplémentaires sur cet échantillon pour effectuer les analyses.

2.1.2.1. *Exploration de l'échantillon*

Après réduction des données, il convient de prendre quelques minutes pour étudier le spectre de masse et le diagramme du défaut de masse en fonction de la masse. Cette vérification a plusieurs fonctions : (1) vérifier que l'on a bien du signal dans la zone attendue, (2) que le bruit et les artefacts ne sont pas trop présents car ils peuvent gêner l'attribution et (3) éventuellement avoir une première idée des formules stœchiométriques et des erreurs en présence par des attributions ponctuelles. Le spectre de masse et le DMvM de l'échantillon utilisé pour illustrer cette partie sont présentés en Figure 30. On note que : (1 - bleu) le spectre de masse présente des signaux réguliers et denses sur une gamme d'un peu moins de trois décades, (2 - orange) qu'il n'y a que peu de signaux hors de cette distribution dense, (3 - vert) que le DMvM présente un fuseau dense de valeurs positives, indiquant la présence de molécules organiques hydrocarbonées, et (4 - violet) qu'on ne voit que peu de lignes de points alignés verticalement, indiquant qu'il n'y a que peu d'artefacts, et surtout que ces alignements artificiels semblent être hors du fuseau principal.

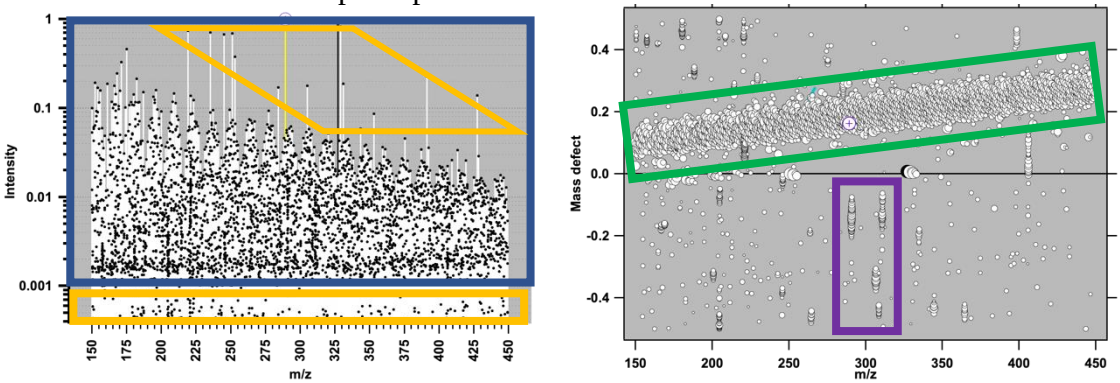


Figure 30 – Spectre de masse et diagramme du défaut de masse en fonction de la masse pour l'échantillon utilisé à titre d'illustration. Les différentes formes colorées délimitent quelques zones visuelles où de l'information est disponible pour une exploration rapide des données.

Basé sur l'ensemble de ces observations, on peut réaliser une première attribution par *Fasttribution* (voir 1.3.5) dans le but de s'assurer de la qualité de la calibration brute, et ainsi déterminer a priori s'il y a besoin d'une calibration interne ainsi que des premières limites à imposer sur les groupes d'attribution. Ainsi, comme présenté en Figure 31(a) par l'utilisation d'une matrice d'attribution sans contraintes particulières, on obtient un résultat d'attribution où une grande majorité des molécules sont à une erreur faible, avec quelques molécules qui se situent à une erreur plus élevée, comme visible Figure 31(b). Il y a également un intérêt à vérifier la distribution en hétéroatomes en fonction de la masse. Puisque l'échantillon est réputé

polymérique, on s'attend à un continuum dense d'attributions sans distribution ponctuelle d'attributions. Ainsi, si l'on prend par exemple la distribution des oxygènes en fonction de la masse disponible en Figure 31(c) et (d), on voit des attributions ponctuelles avant plus de 10 atomes et étant hors du fuseau principal. La même observation s'applique à la distribution en azote pour un nombre d'azote supérieur à 12. Ainsi, on pourrait déjà limiter les attributions a priori pour N<12 et O<10 et ainsi éviter les attributions exotiques si l'on devait se contenter de cette attribution unique.

(a)

Group	Constrains
CH2	0-100
NH	0-100
CH3	1-1
O	0-100
C	0-1000

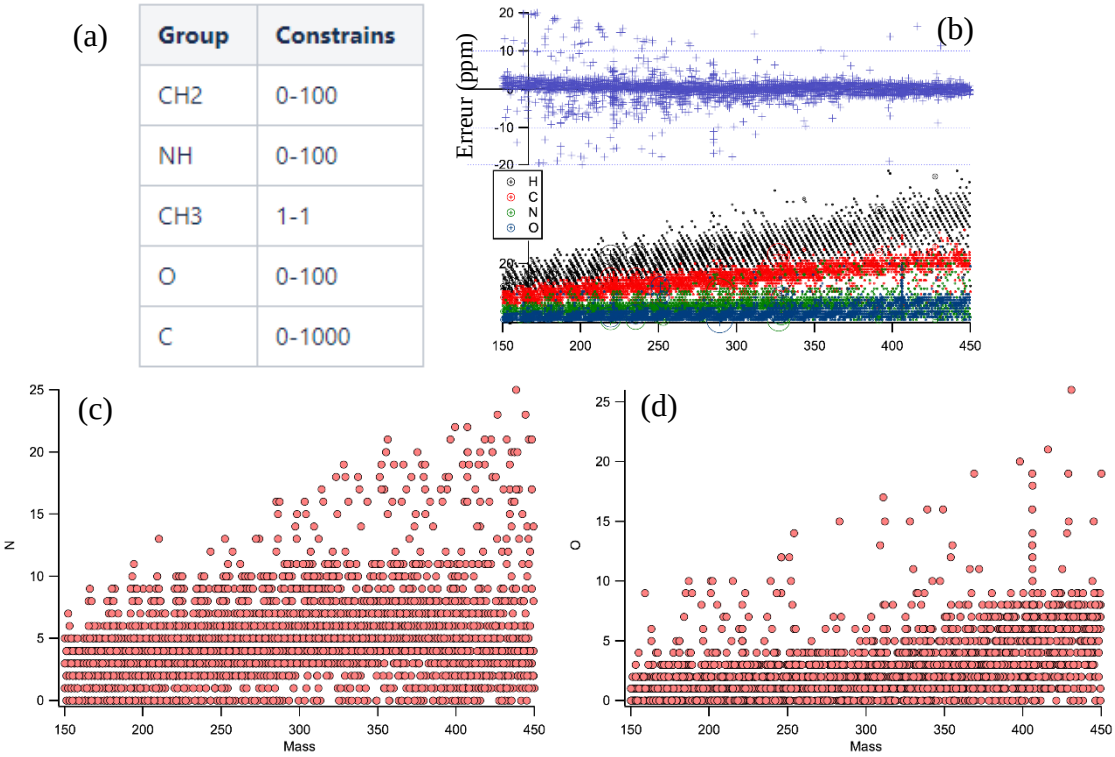


Figure 31 – Matrice d'attribution et représentations graphique des molécules attribuées par l'attribution exploratoire. Chaque point représente une molécule attribuée. (a) Matrice d'attribution utilisée (b) Représentation de l'erreur et des stœchiométries pour chaque molécule attribuée en fonction de la masse (c) Nombre d'azote en fonction de la masse (d) Nombre d'oxygène en fonction de la masse.

2.1.2.2. *Calibration interne*

Après l'étape exploratoire où l'on observe une décalibration légère, l'idée est de corriger ceci. Pour ce faire, on doit établir une liste de données sur laquelle s'appuyer pour effectuer la calibration interne. On peut par exemple sélectionner le fuseau dense de l'attribution précédente, et appliquer un polynôme sur ces données pour recalibrer le spectre, comme présenté en Figure 32.

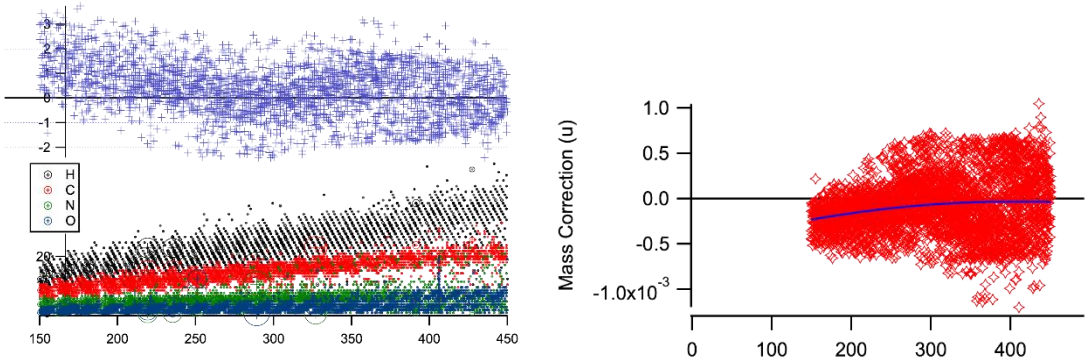


Figure 32 – Illustration de la sélection du fuseau principal et de la calibration polynômiale.

On peut alors vérifier la calibration des données en effectuant une nouvelle attribution sur les données, sans autres nouvelles contraintes que celles de la matrice initiale. Cependant, on peut également vérifier la calibration en extrayant plusieurs familles en CH_2 par *Graphttribution*. Ainsi, si l'on sélectionne par exemple toutes les familles ayant plus de douze membres et à ±2 ppm, on obtient l'attribution disponible en Figure 33, largement décalibrée à haute masse. Cette décalibration n'est pas surprenante du fait de l'augmentation des solutions possibles à haute masses et de la résolution qui n'est alors plus suffisante. Cependant, en utilisant les familles en CH_2, nous sommes ainsi capables de rétablir la calibration à haute masse et ainsi de s'assurer de la pertinence des attributions dans cette zone.

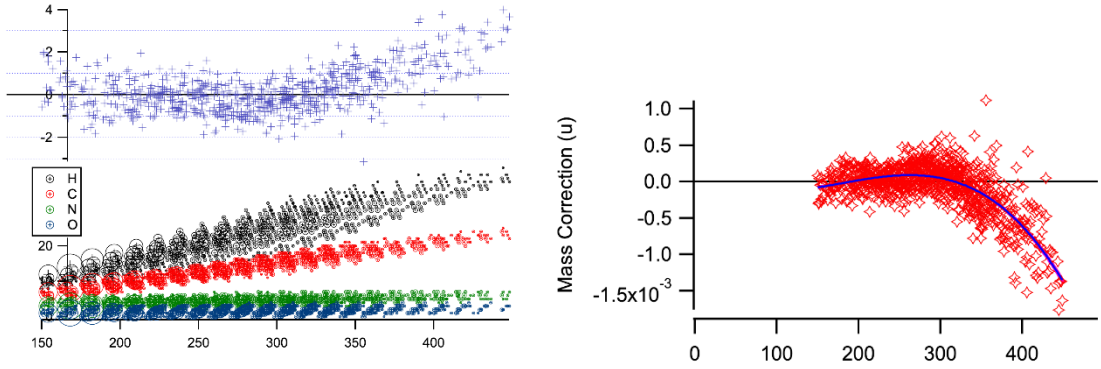


Figure 33 – Calibration en utilisant les familles en CH_2. La décalibration observée à haute masse est due à la perte de résolution de l'instrumentation qui ne permet plus d'assurer une attribution point à point correcte dans cette zone.

On vérifie une dernière fois la calibration en effectuant une *Fastattribution* sans contraintes, pour nous permettre de fixer les limites d'attribution. Ainsi, comme présenté en Figure 34, on peut observer une limite franche entre une distribution dense pour N<15 et O<11 et une distribution ponctuelle au-delà. Cette limite sera celle appliquée pour l'attribution suivante. On observe également que la calibration est correcte : centrée sur zéro, pas de déviation aux bords.

(a)

Group	Constrains
CH2	0-100
NH	0-100
CH3	1-1
O	0-100
C	0-1000

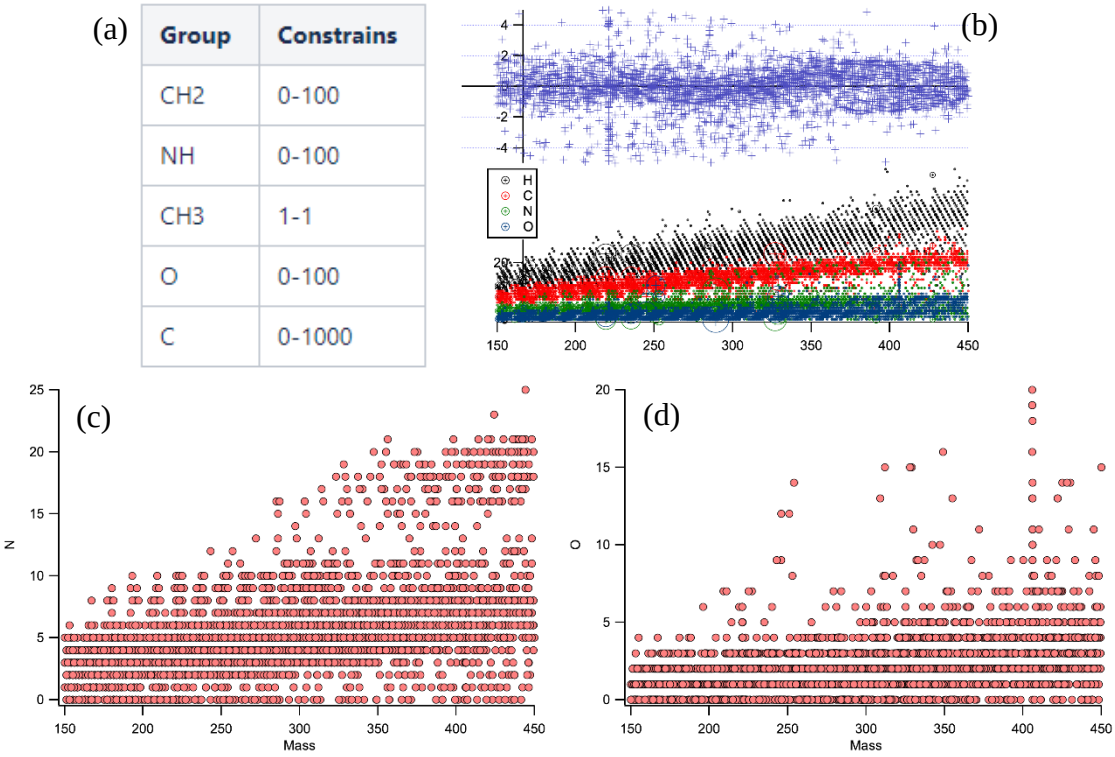


Figure 34 – Attribution après calibration finale. Chaque point représente une molécule attribuée. (a) Matrice d'attribution utilisée (b) Représentation de l'erreur et des stœchiométries pour chaque molécule attribuée en fonction de la masse (c) Nombre d'azote en fonction de la masse (d) Nombre d'oxygène en fonction de la masse.

2.1.2.3. *Attributions*

Une fois le spectre calibré, l'étape suivante consiste à effectuer une attribution de ce spectre. On considère une matrice contrainte cette fois-ci en ajoutant les limites déterminées après la calibration des données. Cette attribution et ces résultats sont fournis en Figure 35.

(a)

Group	Constrains
CH2	0-100
NH	0-14
CH3	1-1
O	0-10
C	0-1000

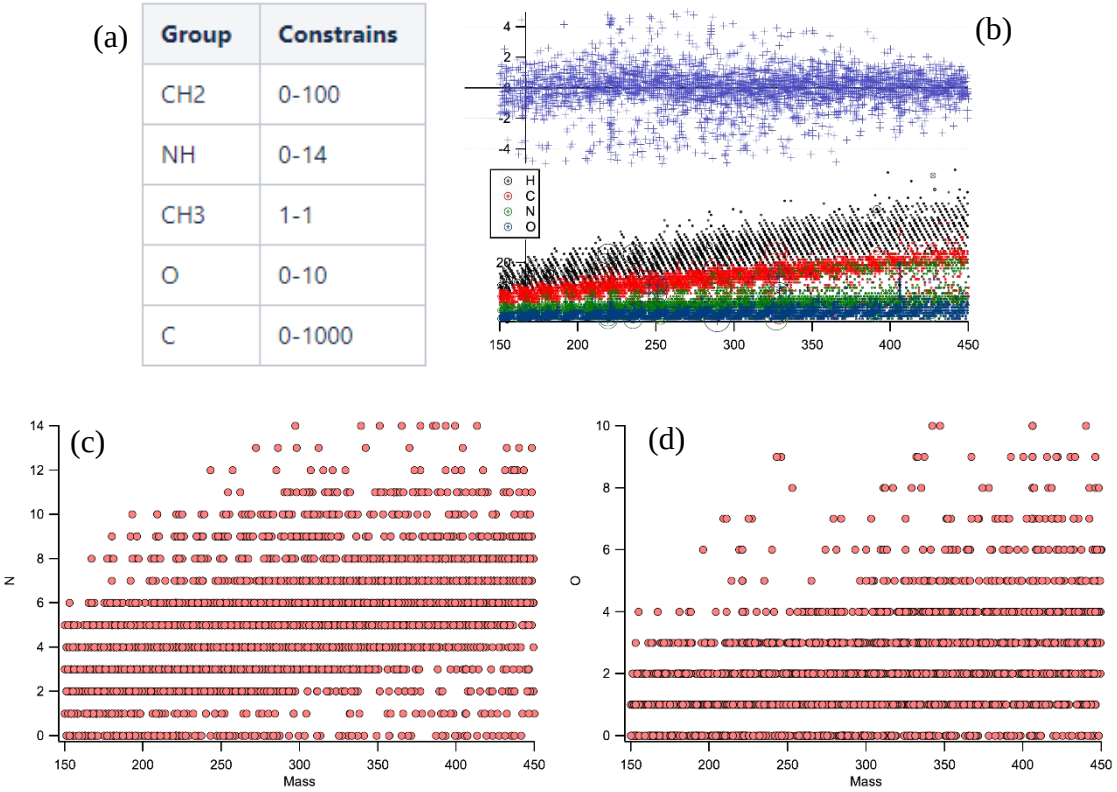


Figure 35 – Attribution contrainte finale, avant nettoyage des mauvaises attributions. (a) Matrice d'attribution utilisée (b) Représentation de l'erreur et des stœchiométries pour chaque molécule attribuée en fonction de la masse (c) Nombre d'azote en fonction de la masse (d) Nombre d'oxygène en fonction de la masse.

Le nettoyage des formules stœchiométriques s'effectue, pour des attributions en CHNO, en quatre étapes : (1) nettoyage des CH, (2) nettoyage des CHN, (3) nettoyage des CHO et enfin (4) nettoyage des CHNO. Si d'autres atomes sont présents, il faut ajouter d'autant plus d'étapes de nettoyages pour vérifier l'ensemble des familles et ainsi s'assurer de leur pertinence.

La difficulté de ces nettoyages est de trouver la bonne représentation permettant de voir graphiquement les attributions qui sont hors des tendances. Chaque famille est extraite et traitée de façon indépendante des autres familles, puis chaque famille est fusionnée à la fin pour obtenir la liste des attributions définitives.

Nettoyage des CH

Ce nettoyage est basique : aucune formule composée uniquement de CH ne peut être détectée *a priori* en ESI. Ainsi, leur suppression pure et simple est effectuée de façon systématique. Un moyen rapide est d'effectuer un clustering en effectuant le test logique (N ==0 && O==0). Cela va mettre toutes les formules en CH seulement à la valeur 1 et l'ensemble du reste de l'espace à la valeur 0. On ne conserve que les formules ayant une valeur nulle à ce test. Une illustration de ce traitement est présentée en Figure 36.

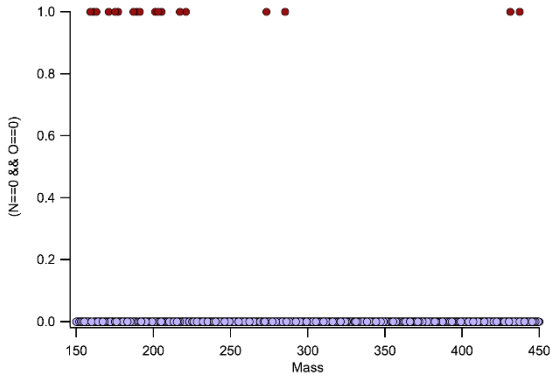


Figure 36 – Traitement logique des formules en CH (valeur 1) en fonction de l'ensemble des autres formules (valeur 0).

Nettoyage des hétéroatomes

L'étape de nettoyage des hétéroatomes est effectuée en deux étapes : la fixation du nombre d'hétéroatomes à conserver puis la vérification de l'hydrogénation des attributions. Ces étapes sont communes au nettoyage des CHN, CHO et CHNO, seule la description du nettoyage des CHN sera effectuée, puisque les deux suivantes seront la répétition du même principe et des mêmes bases de choix à effectuer.

Une représentation efficace qui permet de séparer visuellement les différentes familles qui composent la famille en CHN est la représentation N/C en fonction de la masse. Ainsi, chaque famille qui ne diffère que par l'ajout ou le retrait d'un azote est représentée sous la forme d'une hyperbole. Cette séparation visuelle ne permet cependant pas un traitement mathématique simple puisqu'effectuer du classement dans un espace hyperbolique est complexe. Ainsi, on va se limiter à une interprétation visuelle, déterminer une limite sur le nombre d'azote à conserver et supprimer manuellement l'ensemble des formules ayant plus que ce nombre ainsi fixé. Un exemple de ce traitement est disponible en Figure 37, où l'on observe la distribution en molécules avant (gauche) et après (droite) nettoyage. Ici, on fixe par exemple le nombre d'azote à 8 maximum puisque les distributions contenant 9 et 10 atomes sont discontinues et ne contiennent que quelques membres seulement. On retire également les points qui semblent hors tendance, notamment à faible nombre d'azote : en effet, en comparant leur formule stœchiométrique avec les formules voisines, ces points n'appartiennent pas à des familles en CH_2 et sont abusivement attribués sur des signaux à faibles intensités, possiblement du bruit instrumental.

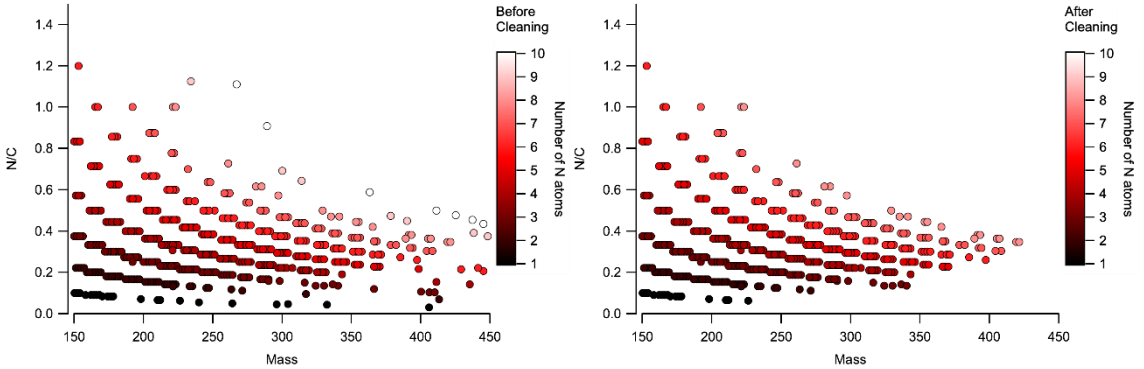


Figure 37 – Nettoyage des N de la famille en CHN ; chaque hyperbole représente une famille avec le même nombre d'azote.

Après avoir nettoyé les familles ayant un nombre d'azote anormal, il faut s'assurer que les attributions restantes aient également des attributions en hydrogène qui soient correctes. Pour ce faire, la représentation H/C en fonction du DBE est utilisée. Cependant, à la différence de la représentation précédente, les familles à nombre d'azote constant se chevauchent, rendant toute interprétation impossible. Il faut alors séparer chaque famille à azote constant et effectuer le

traitement un à un. Quatre représentations de ce traitement sont fournies en Figure 38 à titre d'illustration du processus. Ici, seules les familles à possédant deux et six azotes présentent des points hors des tendances, qui sont retirés manuellement.

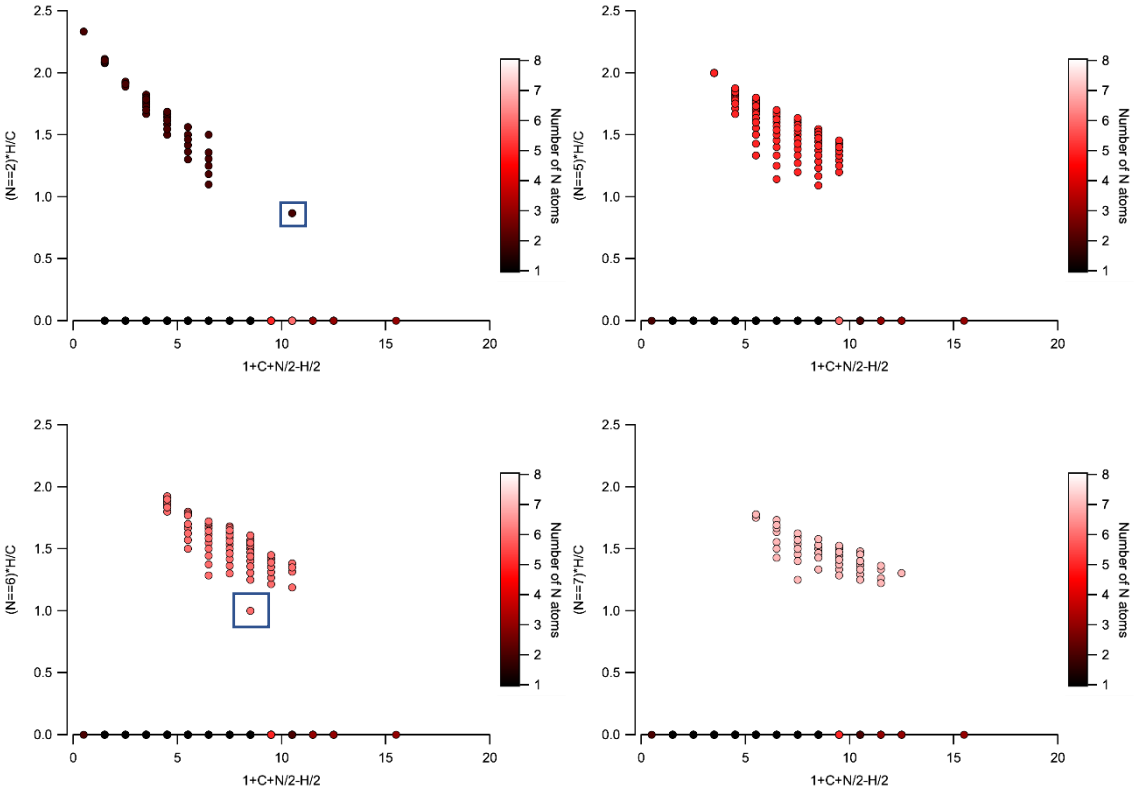


Figure 38 – Illustration du processus de nettoyage des hydrogènes pour la famille des CHN pour N=[2 ;5 ;6 ;7]. Les points retirés sont encadrés à titre d'information

2.1.2.4. *Synthèse du processus*

Le processus d'attribution, pour être valide, se doit d'être systématique et non biaisé. Voici l'ensemble du processus, mis sous forme synthétique, qui se doit d'être suivi lors de l'attribution d'échantillons qui présentent des profils polymériques.

1) Exploration de l'échantillon, s'assurer que les données soient correctes, première attribution ;
2) Recalibration grossière basée sur la première attribution ;
3) Recalibration fine en se basant sur des familles en CH2 ;
4) Attribution et détermination des limites aux groupes d'attributions
5) Attribution finale
6) Nettoyage des formules stœchiométriques
   Pour un échantillon en CHNO et analysée en ESI :
   a. Nettoyage des CH
   b. Famille en CHN
      i. Détermination du nombre maximal d'azote : N/C=f(masse)
      ii. Nettoyage de l'hydrogénation : H/C=f(DBE) et filtrer sur le nombre de N
   c. Famille en CHO
      i. Détermination du nombre maximal d'azote : O/C=f(masse)
      ii. Nettoyage de l'hydrogénation : H/C=f(DBE) et filtrer sur le nombre de O
   d. Famille en CHNO
      i. Détermination du nombre maximal d'azote :

1. N/C=f(masse)
2. O/C=f(masse)

ii. Nettoyage de l'hydrogénation : H/C=f(DBE) et filtrer sur le nombre de N et de O

7) Reconstruction finale de la liste des molécules

2.1.3. Détermination de la matrice d'attribution en LDI

L'ensemble de ce qui a été décrit dans cette partie réfère à l'analyse de la fraction soluble. Or, un des objectifs de cette thèse est de vérifier ce qu'il se passe dans la fraction insoluble des échantillons. Pour ce faire, des analyses en LDI sont effectuées. Néanmoins, afin d'attribuer les analyses LDI qui comportent des radicaux et des ions moléculaires, il faut ajuster la matrice d'attribution qui est utilisée habituellement. En effet, un radical n'aura pas perdu ou gagné un proton comparé à un ion moléculaire généré par une source ESI, engendrant une différence sur la matrice pour prendre en compte cette différence. En utilisant *Graphttribution*, il est également possible de prendre en compte les signaux comportant du ^{13}C. Cette considération n'a pas été prise avant car le nombre de molécules contenant du ^{13}C sur les analyses Orbitrap ne sont pas significatives, alors que dans le cas des analyses ICR, elles représentent plus de 20% des attributions. En plus de l'isotopie en carbone, l'isotopie de l'azote et de l'oxygène ont été essayés dans des essais non présentés ici et sont non concluants.

Cette différence relative à l'isotopie en carbone entre l'analyse ICR et Orbitrap provient de la différence de gamme dynamique effective. On rappelle que la gamme dynamique est la différence entre la molécule attribuée ayant l'intensité la plus élevée et celle ayant l'intensité la plus faible. La gamme dynamique apparente indique *a priori* que les analyses Orbitrap présentent une gamme dynamique plus importante que celle de l'ICR. Cependant, lorsque l'on change les intensités des molécules attribuées en Orbitrap par celles issues des attributions ICR, on se rend compte que les données Orbitrap ne représentent que les molécules les plus intenses des données ICR. Ainsi, les données ICR sont plus sensibles que les données Orbitrap, et la gamme dynamique apparente n'est pas une bonne manière de comparer les données issues d'instruments différents. Ce type de discussion sur la gamme dynamique est également effectué en partie 2.1.1, dans l'article rédigé au sujet de l'optimisation des acquisitions en Orbitrap.

Du fait de la présence simultanée de radicaux et d'ions moléculaires, il faut déterminer une matrice correcte qui permette d'attribuer sans générer de problèmes d'attributions majeurs. Pour ce faire, on prend arbitrairement comme échantillon d'essai l'échantillon total analysé en LDI, polarité positive et on ne considère que 60% des signaux les plus intenses pour accélérer les attributions et les différents essais. Théoriquement, les données sont structurées comme suit :

- Des ions moléculaires hydrocarbonés sans ^{13}C ;
- Des radicaux hydrocarbonés sans ^{13}C ;
- Des ions moléculaires hydrocarbonés avec ^{13}C ;
- Des radicaux hydrocarbonés avec ^{13}C ;
- Autres…

Pour explorer cet espace, deux attributions sont effectuées en utilisant *Fastattribution* et une attribution en utilisant *Graphttribution*. Les attributions utilisent les matrices présentées en Tableau 2, où les deux premières colonnes réfèrent aux matrices utilisées en *Fastattribution* et la dernière colonne réfère à la matrice utilisée en *Graphttribution*, où en plus de cette matrice sont recherchées les permutations en ^{12}C, $^{12}CH_2$, NH et O.

Group	Constrains for radical ions	Constrains for molecular ions	Constrains for both
^{12}C	0-1000	0-1000	0-1000
$^{12}CH2$	0-100	0-100	0-100
$^{12}CH3$	0-0	1-1	0-1
$^{12}CH4$	1-1	0-0	0-1
^{13}C	0-1	0-1	0-1
NH	0-15	0-15	0-15
O	0-8	0-8	0-8

Tableau 2 – Matrice d'attribution utilisée pour attribuer. On utilise la dernière colonne pour attribuer avec Graphttribution, et les deux premières colonnes pour attribuer avec Fastattribution.

Une étude détaillée des attributions issues de *Fastattribution* indique que de nombreux signaux sont attribués à la fois à un ion moléculaire et à un radial, ce qui n'était pas le cas pour *Graphttribution*. Ensuite, comme présenté en Figure 39, on observe un dédoublement du fuseau des carbones (en bleu, 500+ signaux) pour l'analyse en *Fastattribution*, dédoublement qui n'a aucun sens étant donné la nature polymérique de l'échantillon. Ce dédoublement n'est également pas dû à la présence de ^{13}C puisque l'on n'observe qu'un unique ^{13}C dans les molécules, alors que ce dédoublement révèle une variation d'une dizaine de carbones entre les deux fuseaux.

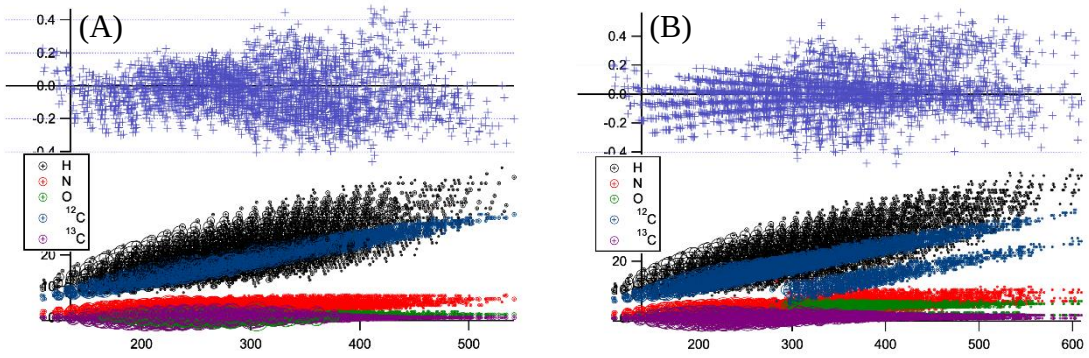


Figure 39 – Comparaison des attributions : (A) Graphttribution ; (B) Fastattribution.

L'attribution en *Graphttribution* génère quatre listes principales de molécules :

- Graine : $N_1{}^{12}C_7H_{10}{}^+$ (2200+ attributions) – ions moléculaires
- Graine : $N_2{}^{12}C_7H_{10}{}^+$ (450+ attributions) – radicaux
- Graine : $N_2{}^{12}C_9{}^{13}C_1H_{13}{}^+$ (230+ attributions) – ions moléculaires en ^{13}C
- Graine : $^{12}C_{10}{}^+$ (15+ attributions) – fullerènes, dus à l'ionisation LDI

Un point intéressant ici est que les trois premières familles sont incluses l'une dans l'autre, comme montré en Figure 40. Cela indique que les ions moléculaires génèrent pour certains des radicaux, et que les ions moléculaires les plus intenses présentent des signaux en ^{13}C. Ce résultat est attendu et le fait de l'observer confirme que les attributions effectuées sont *a priori* correctes.

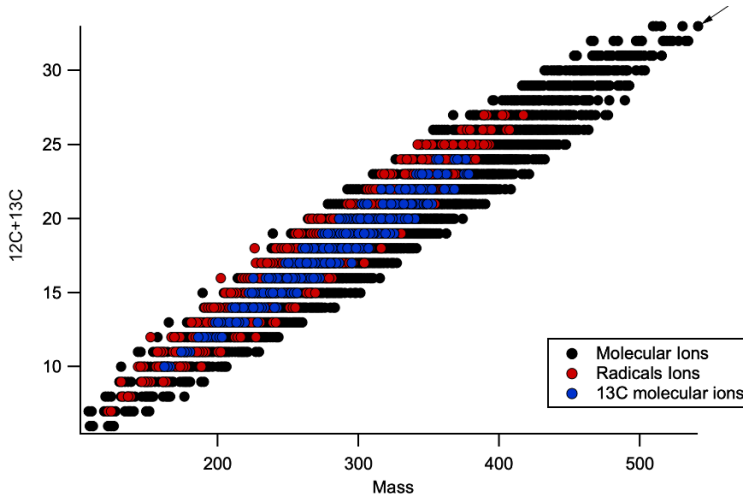


Figure 40 – Représentation des trois familles dans un diagramme du nombre de carbone en fonction de la masse.

Une comparaison des formules attribuées entre *Graphttribution* et *Fastattribution* pour une même masse, et l'étude détaillée des différences entre formules pour une même masse indique deux permutations majeures :

- $CO_5H_6 \leftrightarrow N_7$ – résolution requise : ~14 000 000 à m/z=98 Da ; 32 occurrences
  - Cette résolution est supérieure à la résolution possible avec un ICR de 12T
- $^{13}CHN_3O_4 \leftrightarrow C_{10}$ – résolution requise : ~1 950 000 à m/z=120 Da ; 552 occurrences
  - Cette résolution est supérieure à la résolution possible avec un ICR de 12T

Cette dernière permutation explique le dédoublement du fuseau de carbone en *Fastattribution*, et valide également l'utilisation de *Graphttribution* et de cette matrice d'attribution pour l'ensemble des données LDI.

Les données sont ensuite nettoyées et validées en utilisant la méthode systématique introduite en partie 2.1.2.

2.2. Développement de méthodes chromatographiques

2.2.1. Objectifs et sélections de méthode

À partir de la diversité en familles moléculaires identifiées dans la littérature pour les échantillons complexes considérés (sucres, acides organiques, acides aminés…) il semble pertinent de chercher une méthode analytique qui soit capable de séparer ces différentes familles en un minimum d'injections. Ce point est important car les échantillons sont disponibles en petites quantités (de l'ordre du milligramme) et leur utilisation doit alors être minimisée. Dans ce but, des méthodes analytiques utilisées en métabolomique paraissent pertinentes pour séparer un maximum de molécules organiques biologiques avec un minimum de perte d'échantillon.

Zhang et al [39] ont étudié quelques colonnes aux propriétés très différentes dans le cadre d'analyses non ciblées en métabolomique avec un Orbitrap. Trois chimies de colonnes ont été testées : une colonne C18 pour une analyse en phase inverse, une colonne de silice pour une analyse en phase normale et deux colonnes zwitterioniques pour des analyses en mode d'interaction hydrophile. 177 composés, répartis-en 10 catégories, ont été testés sur chacune des colonnes et leur temps de rétention ainsi que la forme du pic chromatographique ont été étudiées. Les conditions chromatographiques sont données dans le Tableau 3.

Colonne	Phase mobile A	Phase mobile B	Débit (ml/min)	Dénomination
ACE-C18	0,1% Acide formique dans H_2O	0,1% Acide formique dans Acétonitrile	0,3	C18+FA
ZIC-HILIC	0,1% Acide formique dans H_2O	0,1% Acide formique dans Acétonitrile	0,3	ZIC-HILIC+FA
ZIC-pHILIC	20mM Ammonium carbonate dans H_2O pH=9,2	Acétonitrile	0,3	ZIC-pHILIC+AC
Cogent Diamond Hydride	0,1% Acide formique dans H_2O	0,1% Acide formique dans Acétonitrile	0,4	CDH+FA
Cogent Diamond Hydride	20mM Ammonium acétate dans H_2O pH=7	Acétonitrile	0,4	CDH+AA

Tableau 3 - Conditions chromatographiques utilisées pour la comparaison des colonnes, extrait de Zhang et al [39].

Les auteurs rejettent directement la colonne C18 qui voit plus de 80% des composés éluer dans un unique pic large à un temps de rétention proche du temps mort de la colonne. Le point intéressant ici est que chacune des colonnes restantes permet de séparer des composés que les autres colonnes ne séparent pas de façon optimale. Gautier et al [40] ont mentionnés la complémentarité nécessaire de plusieurs méthodes HPLC dans le but de réussir à séparer et caractériser un maximum de composés dans les matrices complexes synthétisées en planétologie. Il est alors intéressant de sélectionner au moins deux colonnes qui ont des rétentions complémentaires, complémentarité caractérisée par le calcul du coefficient de corrélation.

Le Tableau 4 montre ces différents coefficients pour les quatre méthodes HPLC testées. On remarque deux types de valeurs : des coefficients supérieurs à 0,7 et des coefficients compris entre 0,3 et 0,7. Les deux méthodes avec des coefficients supérieurs à 0,7 sont des méthodes plutôt corrélées, qui apportent donc chacune le même type d'information. Ces associations de méthodes ne sont donc pas pertinentes dans le cadre de nos recherches. Les associations restantes ont toutes en commun de fonctionner à des pH différents, avec l'intérêt de séparer plutôt les composés à caractère basique sur une méthode et les composés à caractère acide sur l'autre méthode.

Méthode 1	Méthode 2	Coefficient de corrélation
ZIC-HILIC+FA	CDH+FA	0,92
CDH+AA	CDH+FA	0,72
ZIC-HILIC+FA	CDH+AA	0,69
ZIC-pHILIC+AC	CDH+AA	0,58
ZIC-HILIC+FA	ZIC-pHILIC+AC	0,54
ZIC-pHILIC+AC	CDH+FA	0,45

Tableau 4 Coefficient de corrélation pour les 4 méthodes HPLC testées. Les coefficients de corrélation sont calculés par régression linéaire entre les temps de rétentions des 177 composés pour les colonnes considérées.

Des recherches dans les catalogues commerciaux sur les colonnes ZIC-HILIC et ZIC-pHILIC montrent que les phases stationnaires des deux colonnes sont similaires. La différence est située sur le support qui est en silice pour l'une et en polymère pour l'autre. En première

approximation, les deux colonnes peuvent alors être considérées comme équivalentes. Ainsi, il est choisi de mettre en place une colonne ZIC-pHILIC pour effectuer à la fois les analyses avec les méthodes à pH acide et à pH basique.

2.2.2. La colonne HILIC

Une colonne HILIC, pour Chromatographie d'Interaction Hydrophile (en anglais : *Hydrophilic Interaction Chromatography*), est une colonne dont les interactions sont résumées en Figure 41 :

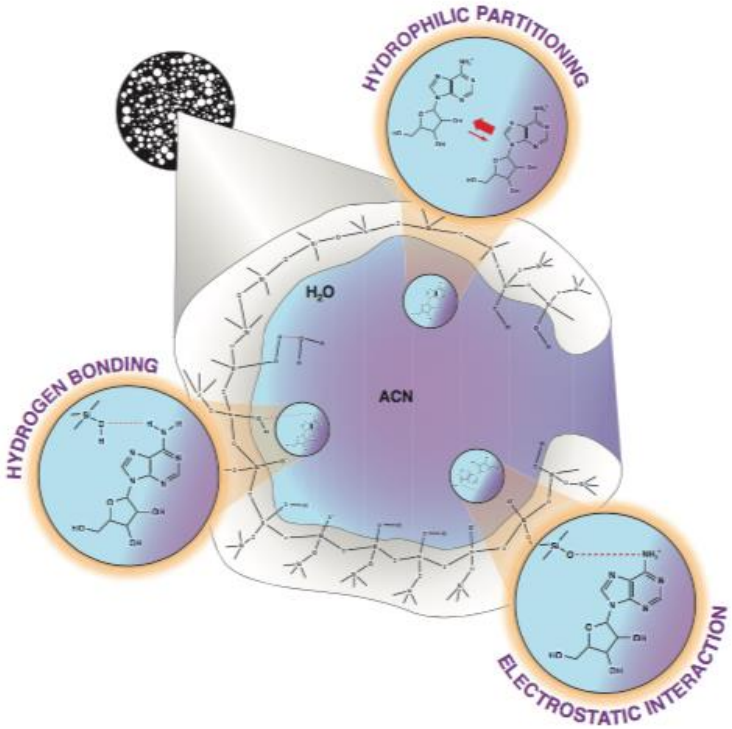


Figure 41 – Représentation schématique d'une colonne HILIC, fisant apparaitre les trois types d'interactions attendues. Extrait de [41].

Selon le type de molécule, la force relative de chacun de ces trois types d'interaction sera différente, engendrant une séparation des composés. Les trois types d'interactions sont alors relatifs à des propriétés physico-chimiques des composés, tels que, par exemple :

- Le partage hydrophile (*Hydrophilic partitionning*), est l'équilibre entre les deux phases liquides (ici, sur le schéma précédent, H_2O et ACN) des composés à séparer. Ainsi, les paramètres physico-chimiques tels que le $logD_{pH\ ou}$ le logP vont caractériser la capacité de chaque molécule à être plus ou moins retenues selon ce principe ;
- Les liaisons hydrogènes (*Hydrogen bonding*), est la capacité d'une molécule à effectuer des liaisons entre groupements polaires de la molécule et de la phase stationnaire. Ainsi, les paramètres physico-chimiques tels que le nombre de groupements donneurs ou accepteurs de liaisons hydrogènes va caractériser la capacité de chaque molécule à être plus ou moins retenue selon ce principe ;
- Les interactions ioniques (*Electrostatic interactions*), est la capacité d'une molécule à effectuer des liaisons ioniques avec la phase stationnaire. Ainsi, les paramètres physico-chimiques tels que la charge nette portée par la molécule à un pH donné vont caractériser la capacité de chaque molécule à être plus ou moins retenue selon ce principe.

Si nous sommes capables de lister, pour chaque molécule étudiée, chacun de ses paramètres physico-chimiques ayant une importance pour expliquer sa rétention sur la colonne, il est possible d'anticiper la rétention des composés en fonction de ses propriétés physico-chimiques.

2.2.3. Sélection des composés

À partir des méthodes présentées dans l'article de Zhang et al [39] et de l'évaluation de ces méthodes effectuées précédemment, les deux méthodes HILIC doivent être adaptées à notre système HPLC. Une méthode chromatographique nécessite une méthodologie systématique pour son développement : la méthode publiée est alors traitée comme référence et des essais conduits pour ajuster la méthode à nos besoins.

Pour commencer, une évaluation systématique des temps de rétention, ainsi que de leur écart par rapport aux temps attendus est effectuée pour chaque modification des paramètres chromatographiques. Ces évaluations sont effectuées à partir de sept composés dont le spectre de masse du mélange est présenté en Figure 42. Ces composés serviront de mélange test tout au long du développement. Pour être représentatifs, ces composés ont été sélectionnés selon plusieurs critères : (A) ils doivent présenter des rétentions réparties sur l'ensemble de la gamme temporelle utile en HPLC, et (B) présenter également des comportements en spectrométrie de masse spécifiques. Ces comportements peuvent être tels que : (1) l'Uracile qui a un signal très faible en ESI-MS, ce qui permet de déterminer si la méthode séparative est capable de détecter des composés à faible intensité, (2) la Glutamine et la Lysine qui ont des masses proches permettant de s'assurer que la résolution en masse de l'instrument n'est pas perdue et (3) que l'ensemble de la gamme dynamique (*i.e.* la plage d'intensités sur laquelle le spectromètre de masse est capable de définir chaque masse avec suffisamment de précision) du spectromètre de masse doit être utilisée, de 2.10^9 pour l'Arginine à 2.10^5 pour l'Uracile, soit 10^4 de gamme dynamique.

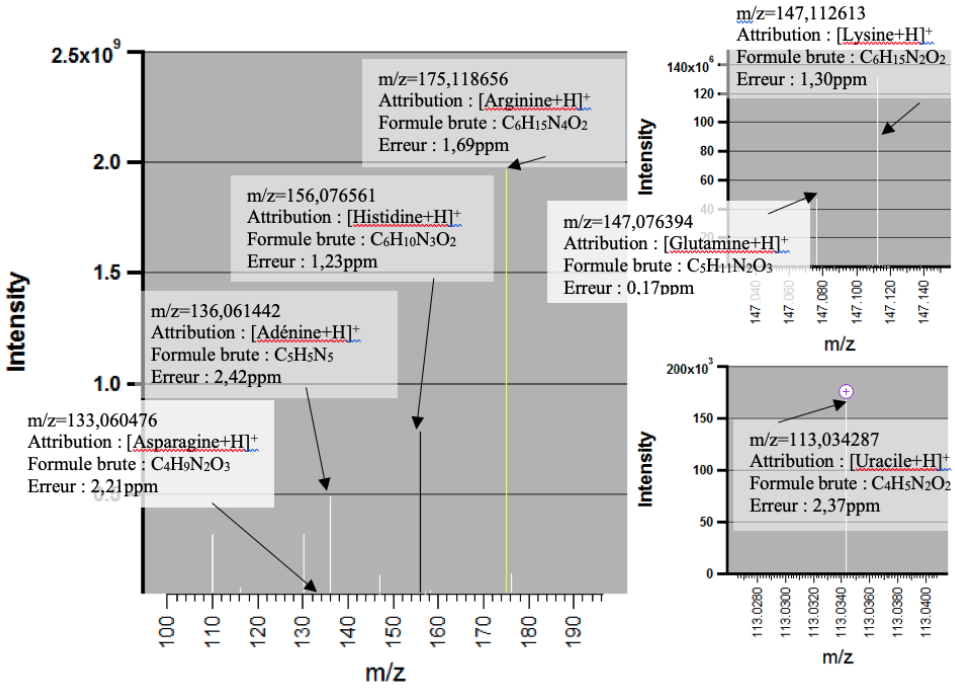


Figure 42 – Spectres de masse du mélange de test, avec zoom sur la gamme autour de 147,10 et 113,03 Spectre acquis en infusion directe, gamme de masse 75-250, T_L=70V, 4scans composés de 128µscans.

2.2.4. Ajustement de la méthode à pH 9.5

La méthode proposée par Zhang et al [39] n'indique pas le protocole de fabrication de la phase mobile Ammonium Carbonate à pH 9,2 et 20mmol.L^{-1}. Ainsi, des recherches annexes ont permis de déterminer que le tampon carbonate pouvait être préparé directement à 1,92g.L^{-1} et le pH ajusté avec de l'ammoniaque à 32%.

L'application de la méthode, sans aucun ajustement, présente des déviations des temps de rétention acceptables pour les composés les moins retenus mais présente des déviations importantes pour les composés les plus retenus, telles que présentées dans le Tableau 5.

Composé	t_R attendu (min)	t_R expérimental (min)	Déviation (%)
Uracile	8,35	8,25	-1,2%
Adénine	9,47	9,08	-4,1%
Histidine	14,42	14,02	-2,8%
Glutamine	14,50	14,56	0,4%
Asparagine	14,98	14,89	-0,6%
Lysine	23,26	20,76	-10,7%
Arginine	28,18	22,10	-21,6%

Tableau 5 – Temps de rétention pour l'application directe de la méthode basique.

Ainsi, ce caractère trop éluant pour les composés ionisés au pH choisi est problématique du fait d'une potentielle perte de sélectivité de la méthode pour d'autres composés non testés. Cette sur-élution peut être provoquée par une trop forte concentration en eau dans le gradient [36] ou par une différence dans l'application du gradient sur la colonne en fonction du temps, *i.e.* une différence de temps de délais entre les différents systèmes. N'ayant aucune indication sur les paramètres du système de la publication, un test où le gradient est coupé lorsque la concentration de 60% d'eau est atteinte est testé et donne les rétentions présentées dans le Tableau 6.

<table>
<tr><th>Composé</th><th>t_R attendu (min)</th><th>t_R expérimental (min)</th><th>Déviation (%)</th><th colspan="2">Résolution</th><th colspan="2">Sélectivité</th></tr>
<tr><td>Uracile</td><td>8,35</td><td>8,42*</td><td>-0,8%</td><td rowspan="2">>1,5*</td><td></td><td rowspan="2">1,14*</td><td></td></tr>
<tr><td>Adénine</td><td>9,47</td><td>9,06</td><td>-4,3%</td><td rowspan="2">14,90</td><td rowspan="2">2,00</td></tr>
<tr><td>Histidine</td><td>14,42</td><td>14,02</td><td>-2,8%</td><td rowspan="2">1,95</td><td rowspan="2">1,05</td></tr>
<tr><td>Glutamine</td><td>14,50</td><td>14,56</td><td>0,4%</td><td rowspan="2">1,48</td><td rowspan="2">1,03</td></tr>
<tr><td>Asparagine</td><td>14,98</td><td>14,87</td><td>-0,7%</td><td rowspan="2">25,36</td><td rowspan="2">1,97</td></tr>
<tr><td>Lysine</td><td>23,26</td><td>25,4</td><td>9,2%</td><td rowspan="2">3,20</td><td rowspan="2">1,12</td></tr>
<tr><td>Arginine</td><td>28,18</td><td>28,02</td><td>-0,6%</td><td></td><td></td></tr>
</table>

Tableau 6 – Temps de rétention, résolution et sélectivité après adaptation du gradient basique.
** le signal est trop faible en MS pour calculer la résolution et la sélectivité. Cependant, en UV@254nm, ces deux espèces sont séparées par un retour net à la ligne de base et un temps de rétention défini, d'où les valeurs affichées.*

Les temps de rétention obtenus sont équivalents à ceux obtenus par Zhang et al [39], mis à part pour la lysine. Cette variation est néanmoins considérée comme acceptable puisque la rétention des composés est plus importante qu'avant l'adaptation du gradient.

Les échantillons complexes peuvent présenter une partie insoluble, pouvant engendrer des problèmes si des particules venaient à se retrouver dans la colonne. Ainsi, pour éviter la dégradation de la colonne analytique, il est primordial d'ajouter une précolonne en amont de celle-ci. Des règles de transfert de méthode dans le but de conserver la sélectivité de la méthode avant et après le transfert existent, mais pourraient être négligées en première approximation du fait du faible volume de rétention ajouté, comme présenté dans le Tableau 7. Le volume de rétention considéré est le volume interne cylindrique de la colonne, sans prise en compte des particules.

Paramètre	Colonne analytique	Pré-colonne
Diamètre interne	4,6 mm	2,1 mm
Longueur	150 mm	20 mm
Granulométrie	5 µm	5 µm
Volume de rétention	1 084 mm^3	66 mm^3

Tableau 7 - Géométries de la colonne et de la précolonne.

Quelques tests, non présentés ici, indiquent néanmoins une variation des rétentions. Ces variations sont compensées par l'ajout d'une minute sur la pente du gradient, de 5 minutes de plateau à 60% d'acétonitrile ainsi qu'un abaissement du débit à 275µL.min^{-1}. Ces ajustements sont calculés numériquement par les principes de transfert de méthode. Néanmoins, la colonne et la précolonne ne présentent pas le même diamètre interne, ce qui nécessite une transformation mathématique. Ainsi, un calcul du diamètre interne moyen peut être réalisé, soit en pondérant chaque diamètre interne par la longueur de chaque système, soit en pondérant chaque diamètre interne par le volume de chaque système. Ces calculs donnent respectivement un diamètre interne moyen de 4,5mm et de 4,3mm. Les méthodes résultantes sont calculées à partir du calculateur développé par Guillarme et al [42], et donnent des méthodes similaires, tant concernant le débit que le gradient. Les temps de rétentions et paramètres de séparations sont présentés dans le Tableau 8 pour évaluer la qualité de la modification.

Composé	t_R (min)	Résolution avant/après transfert		Sélectivité avant/après transfert	
Uracile	9,29	>1,5*/2,96		1,14*/1,19	
Adénine	10,09		14,9/14,9		2/1,97
Histidine	15,57	1,95/2,05		1,05/1,05	
Glutamine	16,17		1,48/1,49		1,03/1,03
Asparagine	16,53	25,36/23,07		1,97/2,01	
Lysine	28,40		3,2/2,92		1,12/1,11
Arginine	31,22				

Tableau 8 - Temps de rétention avant et après transfert pour la méthode basique.
** le signal est trop faible en MS pour calculer la résolution et la sélectivité. Cependant, en UV@254nm, ces deux espèces sont séparées par un retour net à la ligne de base, d'où les valeurs affichées.*

Les résolutions et sélectivités avant et après transfert de la méthode basique étant similaires ou identiques, la méthode est alors prête à être validée et le gradient final est présenté en Figure 43.

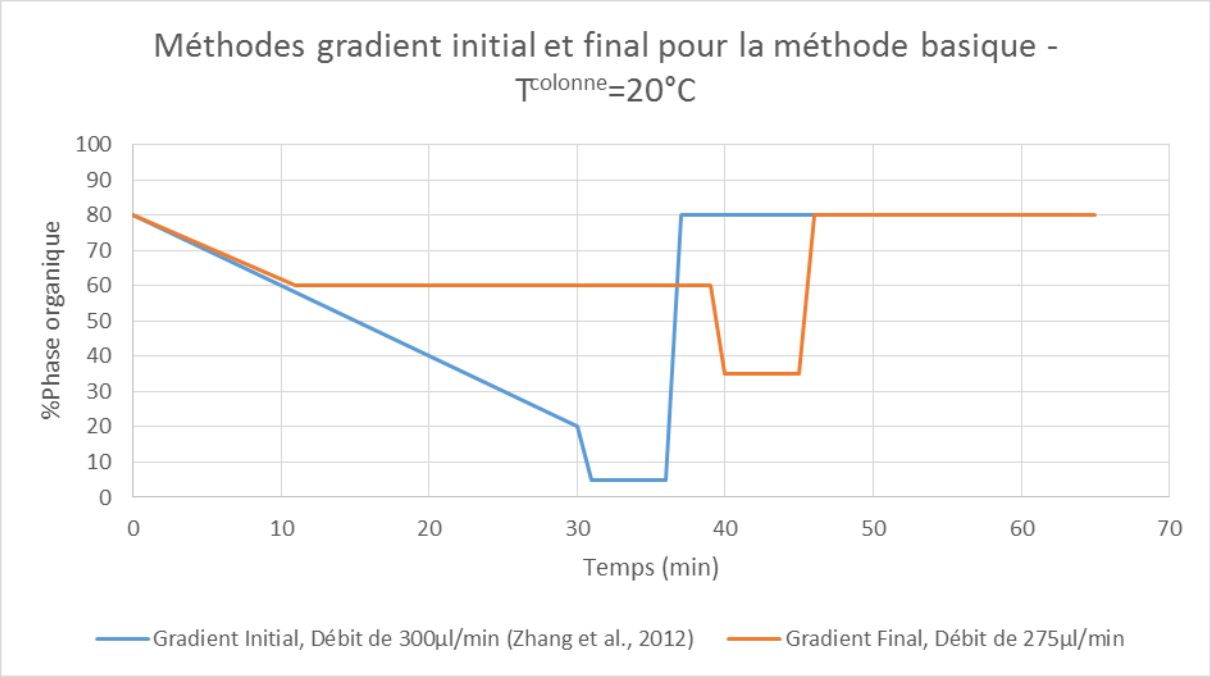


Figure 43 - Comparaison du gradient initial et du gradient final pour la méthode basique.

On peut également constater l'utilité du couplage de la chromatographie avec un spectromètre de masse en comparant une représentation 2D telle qu'elle peut être obtenue par un détecteur UV ou en mode « Total Ion Count » (TIC) à une représentation en 3D. La Figure 44 Gauche montre le chromatogramme TIC, et on remarque clairement que les trois composés autour de 16 minutes ne sont pas résolus. On note également un élargissement de pic important pour la Lysine et l'Asparagine, qui peut être expliqué par l'élution de ces composés après une dizaine de minutes de plateau isocratique. Cependant, la visualisation en trois dimensions, disponible en Figure 44 Droite, permet de s'assurer qu'il est possible de séparer par le temps et par la masse les composés testés.

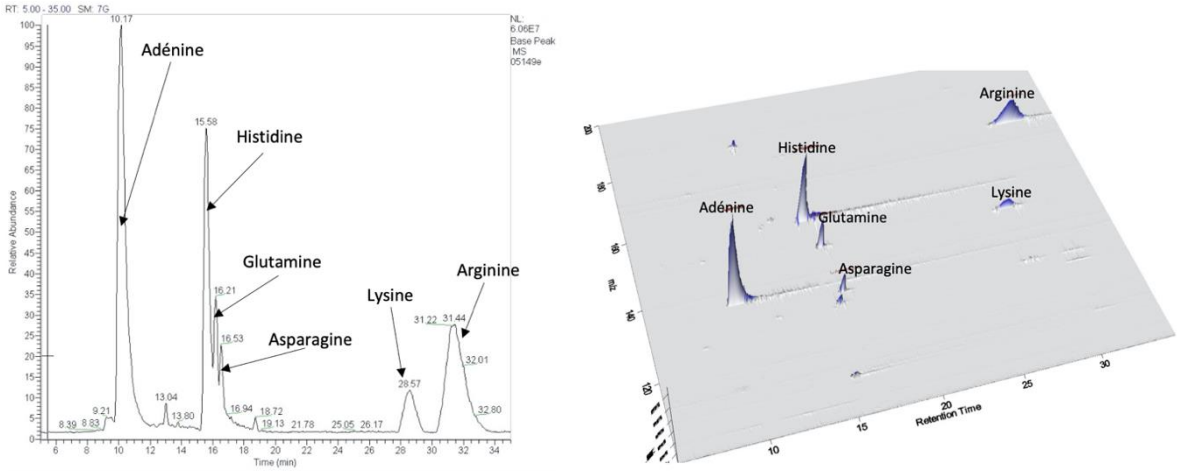


Figure 44 – (Gauche) Chromatogramme type « Base Peak », méthode basique. Zoom sur la zone 5-35 minutes et lissage Gaussien de 7 points. L'uracile n'est pas visible sur ce chromatogramme. (Droite) Représentation en 3D du chromatogramme, méthode basique. Zoom entre 5-35 minutes et entre les masses 100 et 200.

2.2.5. Ajustement de la méthode à pH 3.2

L'implémentation directe de la méthode acide de Zhang et al., 2012 sur la colonne ZIC-pHILIC ne donne pas des résultats exploitables, notamment du fait d'une rétention très élevée des composés ionisés au pH considéré comme la lysine ou l'arginine. Ceci est dû au fait du changement de la phase support (silice vers polymère) de la phase stationnaire, qui a surement

plus d'effet sur la rétention que ce qui est estimé en première approximation. Ces deux composés nécessitent quatre autres gradients complets (soit plus de 4h) à la suite d'une injection de 10µl pour qu'ils soient tous deux élués. Cette rétention trop élevée peut être due à une mauvaise force tampon de la phase mobile.

Pour solutionner ce point, l'eau tamponnée à 0,1% acide formique est remplacée par une solution d'ammonium formate à 27mmol ajusté à pH 3,2 par de l'acide formique pur. L'application directe de la méthode avec cette phase mobile donne des résultats exploitables, comme montré dans le Tableau 9.

Composé	t_R attendu (min)	t_R expérimental (min)	Déviation (%)
Uracile	11,34	9,57	-15,6%
Adénine	16,57	11,04	-33,4%
Glutamine	18,47	16,23	-12,1%
Asparagine	19,12	17,05	-10,8%
Histidine	26,58	22,86	-14,0%
Lysine	26,47	23,44	-11,4%
Arginine	28,2	24,61	-12,7%

Tableau 9 - Temps de rétention après application directe de la méthode acide.

On remarque que l'ensemble des composés présente des temps de rétention beaucoup plus faible que ceux obtenus sur la colonne ZIC-HILIC. Même si l'approximation de l'équivalence de la colonne n'est pas vérifiée, une adaptation du gradient pour obtenir des temps de rétention plus élevés est nécessaire. Ainsi, une augmentation de 5% de phase mobile organique en début et en fin de gradient est appliquée, sans modification des temps de gradient. Les résultats sont présentés dans le Tableau 10.

<table>
<tr><th>Composé</th><th>t_R attendu (min)</th><th>t_R expérimental (min)</th><th>Déviation (%)</th><th colspan="2">Résolution</th><th colspan="2">Sélectivité</th></tr>
<tr><td>Uracile</td><td>11,34</td><td>10,24</td><td>-9,7%</td><td rowspan="2">6,07</td><td></td><td rowspan="2">1,52</td><td></td></tr>
<tr><td>Adénine</td><td>16,57</td><td>13,42</td><td>-19,0%</td><td rowspan="2">8,64</td><td rowspan="2">1,52</td></tr>
<tr><td>Glutamine</td><td>18,47</td><td>18,28</td><td>-1,0%</td><td rowspan="2">1,71</td><td rowspan="2">1,06</td></tr>
<tr><td>Asparagine</td><td>19,12</td><td>19,04</td><td>-0,4%</td><td rowspan="2">10,92</td><td rowspan="2">1,39</td></tr>
<tr><td>Histidine</td><td>26,58</td><td>24,91</td><td>-6,3%</td><td rowspan="2">1,35</td><td rowspan="2">1,03</td></tr>
<tr><td>Lysine</td><td>26,47</td><td>25,51</td><td>-3,6%</td><td rowspan="2">2,28</td><td rowspan="2">1,05</td></tr>
<tr><td>Arginine</td><td>28,2</td><td>26,63</td><td>-5,6%</td><td></td><td></td></tr>
</table>

Tableau 10 - Temps de rétention, résolution et sélectivité après adaptation du gradient acide.

Les temps de rétention ainsi obtenus sont considérés comme acceptables et la méthode est prête pour l'adaptation de la méthode à la précolonne. De la même manière que pour la méthode basique, deux géométries équivalentes moyennes sont considérées pour les calculs en utilisant les mêmes paramètres que ceux utilisés pour la méthode basique, disponible dans le Tableau 7. Les gradients résultants étant presque identiques, le débit est choisi à 275µl/min tandis que le temps de gradient et le temps de plateau sont allongés de 10 minutes. Les temps de rétention obtenus sont présentés dans le Tableau 11.

Composé	t_R (min)	Résolution avant/après transfert		Sélectivité avant/après transfert	
Uracile	10,61	6,07/10,82		1,52/1,74	
Adénine	15,00		8,64/13,9		1,52/1,56
Glutamine	20,84	1,71/3,31		1,06/1,06	
Asparagine	21,74		10,92/17,13		1,39/1,42
Histidine	28,98	1,35/1,85		1,03/1,04	
Lysine	29,85		2,28/2,42		1,05/1,05
Arginine	31,11				

Tableau 11 - Temps de rétention avant et après transfert pour la méthode acide.

Même si quelques sélectivités ne sont pas comparables (Uracile par exemple), les résolutions sont meilleures après transfert. La méthode est prête à être validée et le gradient final est disponible en Figure 45.

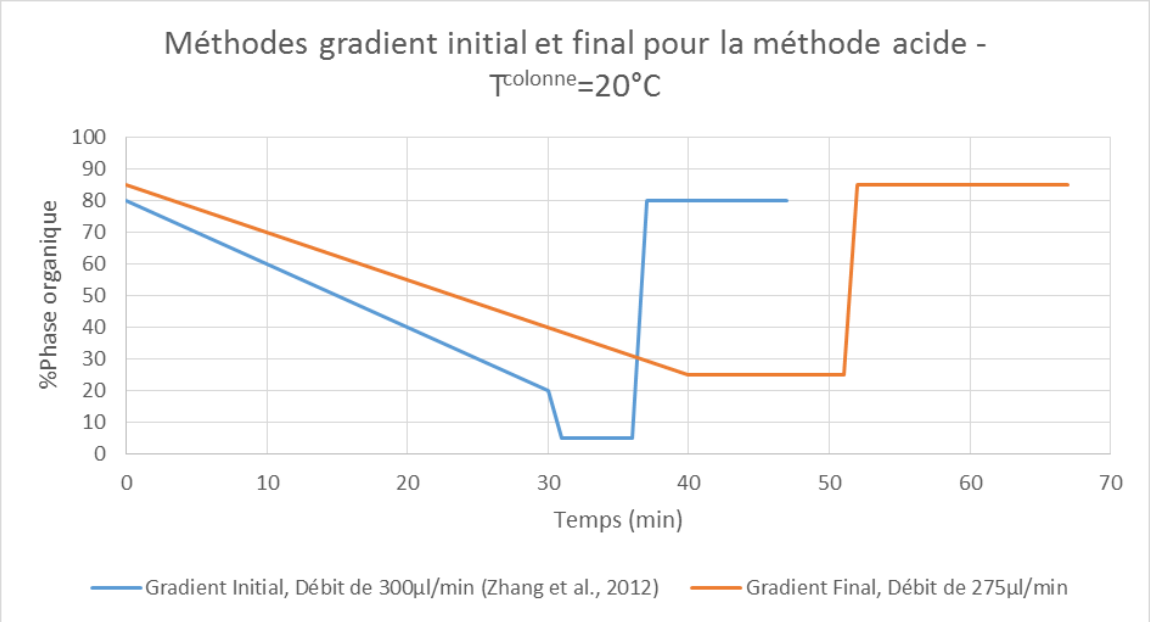


Figure 45 - Comparaison du gradient initial et du gradient final pour la méthode acide.

On peut également constater l'utilité du couplage de la chromatographie avec un spectromètre de masse en comparant une représentation 2D telle qu'elle peut être obtenue par un détecteur UV ou en mode TIC à une représentation en 3D. La Figure 46 Gauche montre le chromatogramme obtenu à l'issue des différentes modifications, et cette représentation ne permet de distinguer chacun des composés. Le point critique est de regarder la séparation entre la Lysine et l'Histidine en 3D, comme en Figure 46 Droite, la première éluant dans la traine du second sur le chromatogramme 2D. Néanmoins, la représentation en 3D permet de s'assurer qu'il est possible de séparer par le temps et par la masse les composés testés.

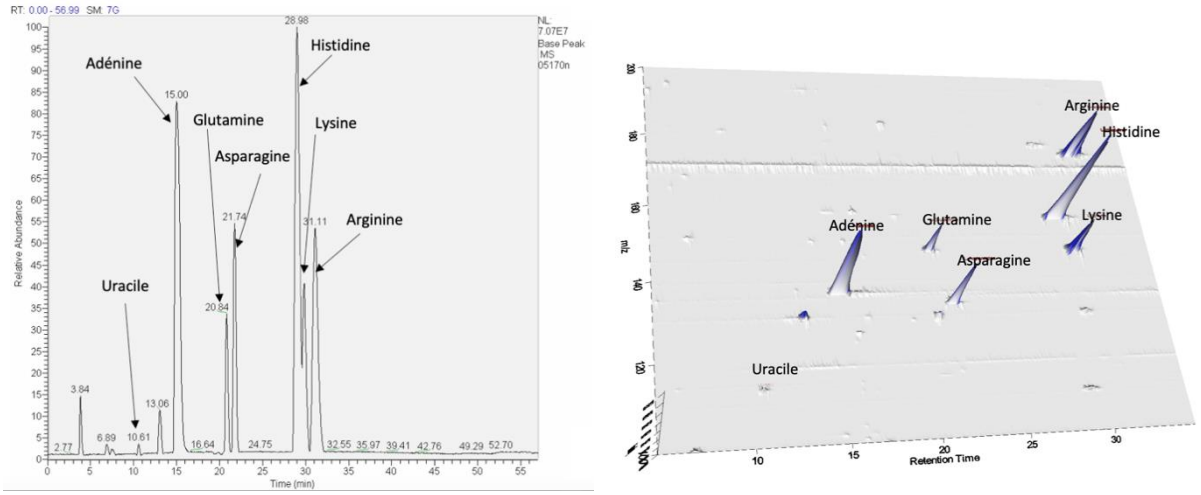


Figure 46 – (Gauche) Chromatogramme type « Base Peak », méthode acide. Zoom sur la zone 5-35 minutes et lissage Gaussien de 7 points. (Droite) Représentation en 3D du chromatogramme, méthode acide. Zoom entre 5-35 minutes et entre les masses 100 et 200.

2.2.6. Validation des méthodes

La méthode acide n'a pas subi de validation rigoureuse car les tests effectués indiquent une méthode stable, à savoir une variation des temps de rétention du mélange tests inférieurs à 2%, sur plusieurs semaines d'utilisation de la même solution tampon.

Cependant, la méthode basique présente de fortes variations des temps de rétention au cours du temps, allant jusqu'à plus de 12% de variation sur certains composés lors de tests effectués non présentés ici. Du fait de la connaissance de la non-stabilité des tampons carbonates montrée par ces résultats, un protocole de validation est créé pour évaluer cette non-stabilité ainsi qu'une estimation de la répétabilité et de la reproductibilité. Quatre séries sont effectuées selon le protocole suivant : (1) 6 injections successives, phase mobile 1 (Série 1), (2) 6 injections successives, phase mobile 1, consécutif à la Série 1 (Série 2), (3) 40 injections successives, phase mobile 2.1 (Série 3) et (4) 25 injections successives, phase mobile 2.2 (Série 4). Chacune des phases mobiles a été réalisée à 20mmol.L^{-1} (±0,1mmol.L^{-1}) et ajustée à un pH de 9,2. Les phases mobiles 2.1 et 2.2 ont été réalisées le même jour et la phase 2.1 est utilisée directement alors que la 2.2 est conservée à 3°C pendant trois jours en flacon fermé avant utilisation.

Chaque série est statistiquement évaluée, pour un risque de 5%, avec la série suivante sur les 6 premières injections de chaque série. La série 3 est statistiquement évaluée pour la variabilité inter-jours ainsi que sur les demi-journées tandis que la série 4 est uniquement évaluée sur la variabilité entre les deux demi-journées d'analyse. Quelques paramètres importants sont présentés dans le Tableau 12. Le nombre de points par série montre qu'au moins 6 injections sont utilisées pour les traitements statistiques, ce qui est un nombre minimal satisfaisant pour la significativité des statistiques effectuées. De plus, les écarts temporels entre les différentes séries évaluées sont réalisés de telle manière que l'évolution des solutions ainsi que leur variabilité puissent être évaluées sur des temps courts (1ère, 2ème et 3ème série), mais également sur des temps longs (3ème et 4ème série).

	Nombre de points par série	Intervalle de temps entre les deux débuts de série (heures)
1ère vs 2nd série	6/6	8h08min
2nd vs 3ème série	6/6	12h19min
3ème vs 4ème série	6/6	76h15min
3ème série – inter-jours J1-J2	8/18	10h50min
3ème série - inter-jours J2-J3	18/14	24h23min
3ème série - inter-jours J1-J3	8/14	35h13min
3ème série - Demi-journée 1 vs Demi-journée 2	9/9	12h12min
3ème série - Demi-journée 2 vs Demi-journée 3	9/9	12h11min
4ème série - Demi-journée 1 vs Demi-journée 2	9/9	12h12min

Tableau 12 - Nombres de points pour chaque série évaluée ainsi que l'intervalle temporel entre les deux premières injections.

Le calcul de la dérive des temps de rétention au cours du temps pour la 3ème et la 4ème série montre que cette dérive est inférieure à 5% pour l'ensemble des composés testés, déviation qui est acceptable. Une représentation graphique des déviations est disponible en Figure 47, comportant également une indication pour pouvoir comparer les résultats pour un nombre d'injections comparable entre les deux séries. Ainsi, à un nombre d'injections identique entre les deux séries, soit 25 injections consécutives, il apparait que la 4ème série présente une dispersion des temps de rétention qui ne présente pas de croissance ou décroissance des temps de rétention au cours du temps alors que l'on discerne clairement des croissances et décroissances des temps de rétention pour la 3ème série. Ces résultats semblent montrer que la conservation du tampon quelques jours à 3°C stabilise le tampon et permet des analyses qui sont alors statistiquement identiques.

L'étude statistique des variances et des moyennes révèle que même si la variance est globalement identique entre chaque série testée, les moyennes ne sont pas statistiquement comparables, mis à part en ce qui concerne les variances entre demi-journées. Cette constance globale de la variance associée à une dérive statistique des temps de rétention moyens semble indiquer une vitesse de dégradation de la phase mobile constante au cours du temps, et indépendante de la solution utilisée. Il apparait également que les moyennes des temps de rétention sont comparables lors de l'utilisation de la solution qui a été conservée pendant quelques jours à 3°C alors que pour toutes les autres conditions testées, les temps moyens ne sont pas comparables.

Ces tests indiquent qu'il semble important de laisser reposer la solution ainsi préparée avant utilisation. Ces tests indiquent également que dans des conditions où la solution est stable, on doit s'attendre à une répétabilité et reproductibilité de la méthode de l'ordre de 1 à 2%.

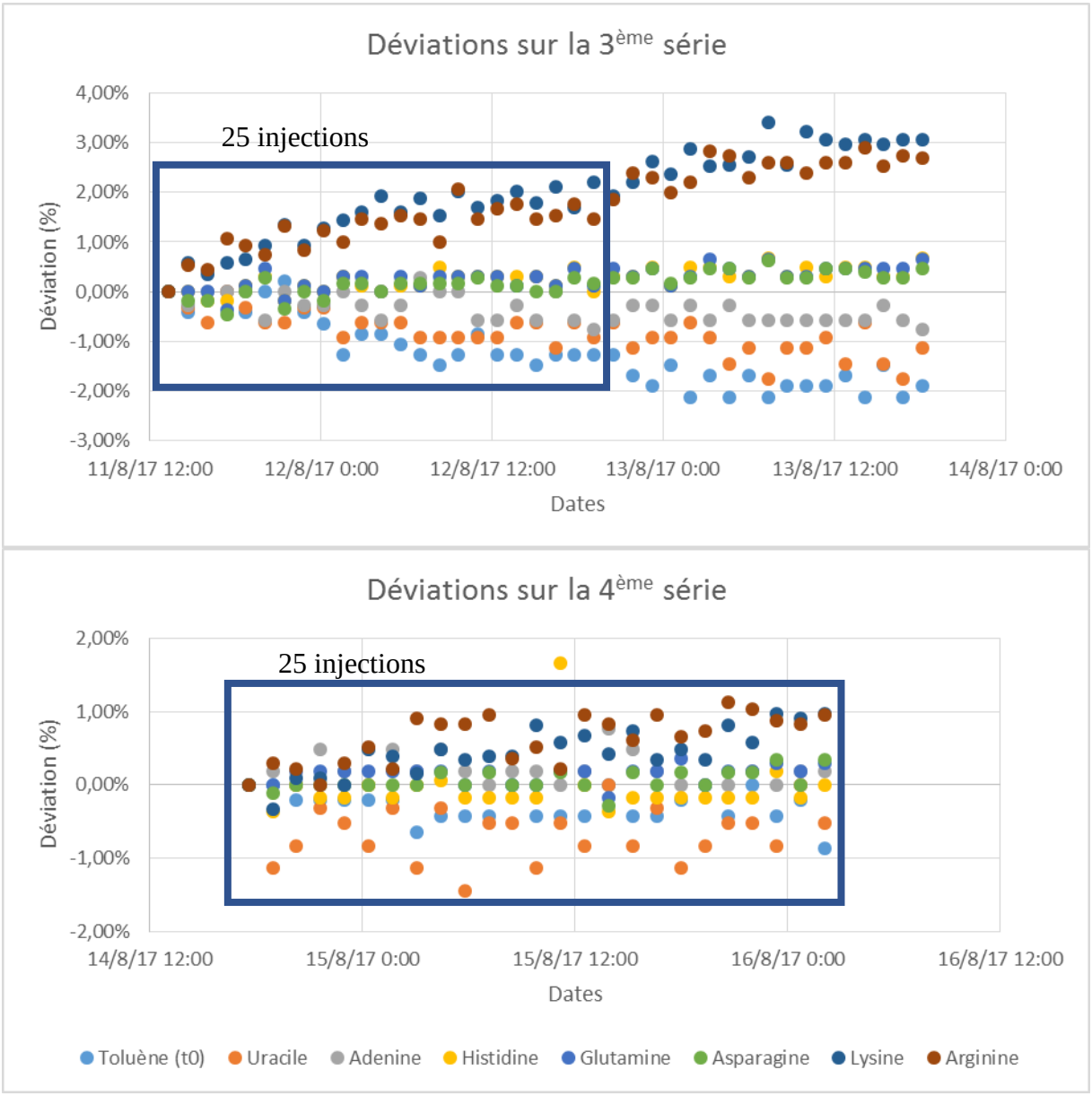


Figure 47 - Comparaison graphique des déviations entre les séries 3 et 4
Référence des déviations : temps de rétention de la première injection

2.2.7. Considérations de pH

Une limite éventuelle à prendre en compte est le fait de considérer le pH qui s'applique en tout temps sur la colonne comme étant constant et fixé par le pH de la phase aqueuse. En effet, le pH d'un mélange lorsqu'il est mélangé à un solvant organique n'est pas nécessairement égal au pH de la phase aqueuse, mais dépend du ratio de mélange entre les deux phases. Comme le pH a un effet important sur la rétention des composés chargés sur les méthodes HILIC, il est important de comprendre l'influence de la variation du pH dans le mélange des phases aqueuse et organique pour interpréter les rétentions observées.

Une façon d'exprimer les pH selon le mode de calibration peut être effectué selon l'Équation 7 où ${}^{s}_{s}pH$ correspond au pH dans la phase hydro-organique calibrée dans la même composition hydro-organique, ${}^{s}_{w}pH$ correspond au pH dans la phase hydro-organique calibrée dans l'eau et δ est le terme qui quantifie d'une part l'énergie de transfert de Gibbs d'une mole de protons en phase aqueuse à l'état standard vers la phase hydro-organique à une température donnée et d'autre part la différence entre la jonction liquide effectuée lors de la calibration du pH-mètre en phase aqueuse et la mesure effectuée dans la phase hydro-organique.

$$ {}^{s}_{s}pH = {}^{s}_{w}pH - \delta $$

Équation 7 - Représentation thermodynamique des pH selon la calibration, extrait de [43]

En effet, en chromatographie, le pH à prendre en compte est le pH thermodynamique, le ${}^{s}_{s}pH$, et non le pH apparent que l'on mesure de la même manière que pour les solutions aqueuses, le ${}^{s}_{w}pH$ [44]. Ainsi, il apparait nécessaire soit de mesurer directement le pH thermodynamique, soit de pouvoir convertir le pH mesuré classiquement. Cependant, la détermination de δ n'est pas une tâche aisée et n'est pas forcément disponible pour tous les tampons et toutes les phases organiques disponibles. De la même manière, il existe plusieurs façons de mesurer les pH :

- Mesure du pH dans la solution hydro-organique à partir du système d'électrodes calibré dans des solutions standards préparées dans des conditions de composition de solvant identiques, permettant d'obtenir directement ${}^{s}_{s}pH$;
- Mesure du pH dans la solution hydro-organique à partir du système d'électrodes calibré dans des solutions standards aqueuse classiques, permettant d'obtenir directement ${}^{s}_{w}pH$;
- La mesure du pH dans la fraction de phase mobile aqueuse, puis d'ajouter la fraction organique à partir du système d'électrodes calibrées dans les solutions standards aqueuse classiques.

Pour des raisons pratiques, il est bien plus facile de mesurer selon la seconde méthode car les solutions de calibration standard sont utilisées alors que les solutions de calibration qui seraient nécessaire dans la première méthode de mesure ne sont que peu ou pas disponibles. Il apparait alors nécessaire de trouver une façon simple et efficace de modéliser δ aux différentes compositions en phase organique pour pouvoir convertir les valeurs de pH mesurées en valeurs de pH thermodynamiques utiles pour la modélisation des rétentions en chromatographie. Pour calculer le pH thermodynamique, certaines hypothèses doivent être effectuées qui sont plus ou moins discutables selon les situations. La première est le caractère totalement dissocié de l'acide dans le mélange de solvants considéré, soit l'assimilation de l'activité à la concentration. La seconde est le fait de pouvoir assimiler la valeur de δ calculé pour un acide fort à n'importe quel autre acide. Les autres hypothèses sont également importantes mais ne seront pas considérées dans le cas de cette étude.

Une modélisation de δ est proposée, δ_m, pour des compositions en acétonitrile allant de 0 à 90% (v/v) et des températures de 15°C à 60°C selon l'Équation 8 où X est la fraction en acétonitrile, a, b et c les paramètres de modélisation et t la température en degrés Celsius.

$$ \delta_m = X \frac{a + bt}{1 + cX} $$

Équation 8 - Modélisation de δ_m, extrait de [44]

L'estimation des paramètres a, b et c, du fait des hypothèses, peut être reprise directement de l'étude et appliquée à nos conditions expérimentales selon le Tableau 13

	a	b	c	R^2	Écart-type
Valeur	$-2{,}323.10^{-3}$	$-1{,}544.10^{-5}$	$-9{,}48.10^{-3}$	0,992	0,023 (0-80% ACN) 0,05δ (>80%ACN)
Incertitude	$1{,}2.10^{-5}$	$2{,}6.10^{-7}$	$7{,}5.10^{-6}$		

Tableau 13 – Paramètres de modélisation de δ_m, adapté de [44]

L'estimation du pH thermodynamique peut ainsi être discuté selon l'évolution de la valeur de δ_m avec la composition en acétonitrile. Ainsi, comme présenté sur la Figure 48, l'impact est assez faible à faible proportion en acétonitrile, et augmente fortement aux concentrations élevées en acétonitrile. Ainsi, en première approximation, le pH thermodynamique peut être

assimilé au pH mesuré pour les concentrations en acétonitrile n'excédant pas 40% (v/v). Pour les concentrations supérieures, une correction doit systématiquement être apportée.

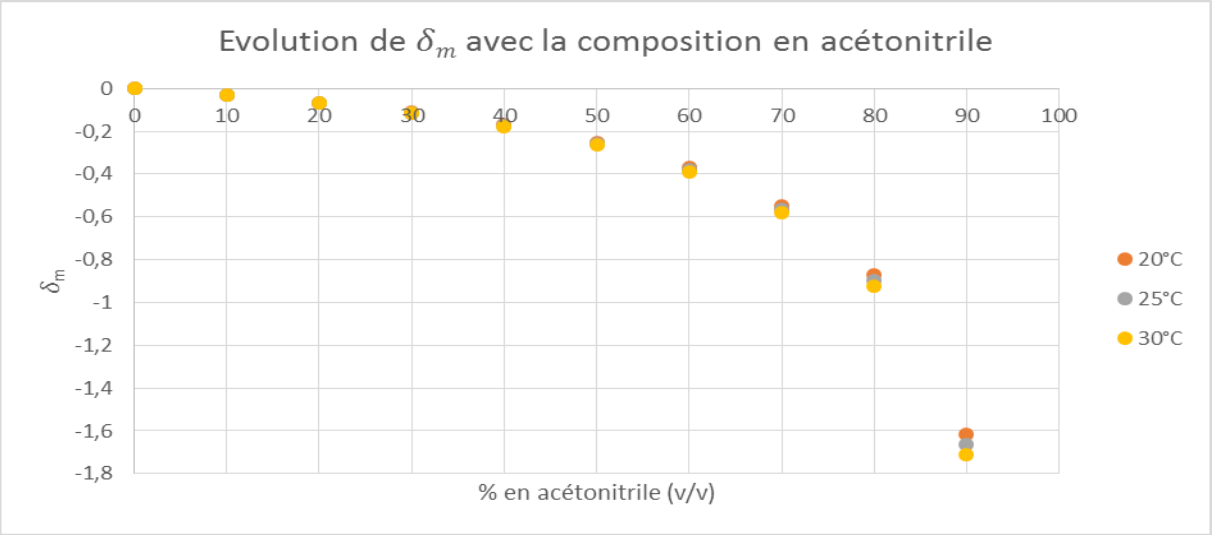


Figure 48 – Diagramme représentant l'évolution de δ_m avec la composition en acétonitrile

Ainsi, même si les méthodes sont appliquées avec une composition en acétonitrile qui dépasse les 40%, on considère en première approximation que le pH de la phase aqueuse est similaire au pH de la phase mobile. Cependant, cette approximation peut expliquer des déviations dans les temps de rétentions pour les composés dont le pKa est comparable au pH de la phase mobile puisque les paramètres physico-chimiques calculés ne sont alors pas équivalents entre le pH de la phase aqueuse et le pH réel de la phase mobile. Comme les outils de calculs des paramètres physico-chimiques n'ont pas encore été transféré sous IGOR PRO et utilisent encore les outils développés sous MS Excel[45], cette correction n'est pas effectuée dans les travaux proposés et devra éventuellement être considérée lors de développements futurs.

2.2.8. Systématique de développement

Basé sur cette expérience de développement, ainsi que sur l'encadrement d'une stagiaire de M1 axé sur le développement de méthodes chromatographiques, un protocole de développement de méthode systématique a été rédigé. Ce protocole a pour objectif d'être une ressource pour des utilisateurs futurs de la chromatographie au laboratoire, utilisateurs qui ne sont pas forcément experts de la méthode mais qui ont besoin d'utiliser l'instrument pour leurs analyses. Ainsi, ce document fournit divers niveaux de lectures en fonction de l'informations souhaitée, et permet la compréhension et la modification de méthode existantes, ainsi que l'adaptation ou le développement de nouvelles méthodes chromatographiques.

2.3. Prédiction des temps de rétention

2.3.1. Théorie

2.3.1.1. *Pourquoi prédire des temps de rétention ?*

Identifier une molécule en chimie n'est pas un acte anodin. En effet, par le jeu des isomères, isobares ou encore des énantiomères, une molécule a plusieurs dizaines voire centaines de variations qui rendent des molécules plus ou moins similaires les unes avec les autres. Cette variation engendre une complexité importante en ce qui concerne l'identification de molécules. Ainsi, Sumner et al[46] ont proposé une façon de normaliser l'identification de molécules pour la métabolomique, mais qui est applicable à n'importe quelle identification chimique, en faisant la différence entre annotation et identification. Ce travail de recherche de standardisation fait référence, de façon plus large, à l'ensemble des façons de rapporter des résultats en science analytique.

Annoter une molécule est quand une (ou plusieurs) de ses propriétés mesurées correspondent aux propriétés rapportées dans une base de données. Ces mesures peuvent être (ou non) acquises de la même façon que celles de la base de données. A l'inverse, une identification est quand au moins deux méthodes orthogonales (ex : chromatographie et spectrométrie de masse en tandem) donnent un résultat strictement équivalent aux mêmes analyses effectuées sur un standard chimique et effectué strictement dans les mêmes conditions.

Dès lors, identifier est un processus bien plus strict que l'annotation puisqu'elle requiert la possession d'un standard chimique au laboratoire et une analyse de ce standard dans les mêmes conditions que celles de l'échantillon. Cela demande donc de l'argent pour acheter les standards et beaucoup de temps machine disponible si les standards existent commercialement. Sinon, il faut envisager une synthèse organique pour obtenir le standard. A l'inverse, annoter une molécule ne requiert qu'une analyse et une base de données établie. La simplicité du processus d'annotation se doit alors de différencier différentes façons d'annoter une molécule :

- Composé inconnu : ces composés ne peuvent pas être classés, mais peuvent être différenciés étant donné leur signature spectrale ;
- Classe de composé caractérisé putativement : basé sur les propriétés physico-chimiques d'une classe de composés chimiques, ou par similarité spectrale avec d'autres composés connus de cette famille de composés ;
- Composé annoté putativement : basé sur les propriétés physico-chimiques et/ou des similarités spectrales issues de bases de données ou de bibliothèques.

Souvent, le processus d'identification d'une molécule remonte la chaine d'annotation depuis une molécule inconnue, puis une molécule dont la classe est connue, puis une annotation et enfin une identification. Cette chaine d'annotation avant d'arriver à une identification a l'avantage de réduire la liste des molécules possibles et donc de sélectionner seulement les meilleures possibilités pour effectuer l'identification finale, réduisant les coûts et le temps nécessaire pour effectuer une identification.

Cependant, la dernière étape nécessite une comparaison avec une base de données. En chromatographie couplé à la spectrométrie de masse, cela veut dire que pour chaque composé de la base de données, il faut lui associer une masse exacte et un temps de rétention. Associer une masse exacte est trivial, associer un temps de rétention l'est beaucoup moins. C'est pourquoi il est intéressant de pouvoir prédire les temps de rétention de composés en se basant sur leurs propriétés physico-chimiques et ainsi de pouvoir réduire la liste des composés potentiels avant identification, si cela est nécessaire.

2.3.1.2. *Objectifs de développement*

Les objectifs du développement d'une méthode de prédiction des temps de rétention sont multiples :

- Sélectionner des composés calibrant sur une large gamme de paramètres physico-chimiques et s'assurer de leur détection ;
- Être capable, à partir d'un minimum d'injections de mélanges de composés calibrant, de générer un modèle statistiquement acceptable permettant de prédire théoriquement plusieurs milliers de structures moléculaires ;
- Être capable, à partir d'une liste de temps de rétention extraite d'un échantillon inconnu, de proposer une liste de structures moléculaires probables, et par opposition, d'exclure des structures non-probables ;
- Obtenir un logiciel modulaire, utilisable couramment et avec la possibilité d'être intégré dans *Attributor*.

Le développement est fait sous IGOR PRO et non directement dans Attributor. Cela permet de séparer le développement des temps de prédiction du développement du module de prédiction des temps de rétention et ainsi de limiter les interférences. Cela nécessitera cependant un travail d'intégration dédié lorsqu'il sera décidé d'inclure ce module à Attributor.

2.3.2. Pratique

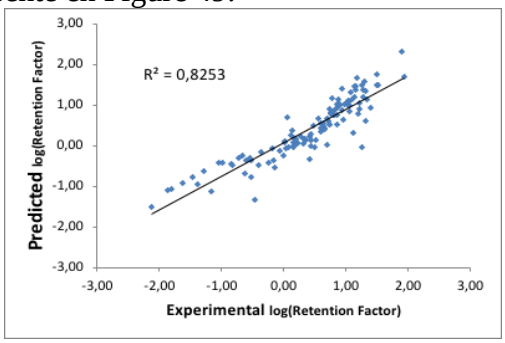
2.3.2.1. *Méthode de référence*

L'équipe ayant proposé les méthodes sélectionnées en chromatographie a également publié un modèle de prédiction des temps de rétention valide pour leur méthode à pH acide [45]. La base de données supporte une large gamme de composés utilisés en métabolomique, avec des atomes exotiques tels que Phosphore, Chlore, Iode ou Soufre de présents.

Le modèle, une régression linéaire multivariée, comporte six paramètres :

- Le ratio du nombre de donneurs de liaisons hydrogène par la masse exacte (HBD/MW) ;
- Le nombre de phosphates ;
- Le nombre de liaisons ayant une rotation possible ;
- Le nombre de charges positives à pH 3,5 (pos(3,5)) ;
- Le nombre de charges négatives à pH 3,5 (neg(3,5)) ;
- Le logD à pH 3,5.

Le modèle utilise 120 composés comme calibreurs et fournit un modèle acceptable, comme présenté en Figure 49.



MLR Equation	HBD/MW	Phosphates	ROTATABLE BONDS	pos (3.5)	neg (3.5)	logD (3.5)	constant
co-efficient	19,72	0,46	-0,05	0,20	0,24	-0,24	-0,87
std error	6,64	0,11	0,03	0,07	0,08	0,03	0,13
r-square	**0,83**	0,35					
F stat	89	113					
ss	66,7	14,12					

Accuracy windows		
within	<10%	48%
within	<20%	78%
within	<35%	94%
within	<50%	99%

Figure 49 – Estimation de la précision des prédictions (auteurs) : ±35% (94% des prédictions sont à ±35% d'erreur)

Du fait d'un grand nombre de composés ayant une composition atomique exotique, on ne conserve que les composés en CHNO. Cette base de données pourra être complétée par de nouvelles molécules suivant les besoins.

Également, du fait de la modification profonde de la structure de la base de données à cause de l'exclusion d'un grand nombre de composés, une nouvelle analyse statistique doit être effectuée pour obtenir un modèle de prédiction acceptable. Sur les 120 composés initiaux, seuls 81 sont en CHNO uniquement, ce qui pourrait nous permettre de calculer directement un nouveau modèle basé sur la régression de référence. Même si les résultats sont corrects (R^2 au-delà de 0,8 ; fenêtre de précision comparable à celle des auteurs), on peut se poser la question si le modèle est statistiquement acceptable du fait du nombre de paramètres engagés.

Une nouvelle étude statistique complète, utilisant l'ensemble des paramètres physico-chimiques de la base de données est donc menée. Cette analyse statistique est réalisée pour obtenir le meilleur modèle possible en partant d'un modèle contenant un nombre important de coefficients puis en le réduisant au fur et à mesure des itérations. Pour obtenir un modèle acceptable, plusieurs conditions d'acceptation définis à partir de paramètres statistiques sont définis *à priori* comme suit :

- Paramètres – La probabilité p, calculée à partir de la statistique de Student, permet d'indiquer la probabilité que le coefficient considéré soit non significatif, à l'erreur statistique près. Cette probabilité est calculée comme suit :

$$p = 2 * \left(1 - StatsStudentCDF\left(abs\left(\frac{coefficient}{erreur\ type}\right); DDL\right)\right)$$

Équation 9 – Formule de calcul de la probabilité p sous IGOR Pro. StatsStudentCDF représente la distribution de Student cumulative ; DDL représente de nombre de degrés de liberté de modèle, défini comme le nombre de composés moins le nombre de paramètres du modèle moins 1.

En pratique, on considère $p \leq 0{,}05$ comme représentant un coefficient significatif et à l'inverse, une probabilité élevée comme représentant un coefficient non-significatif. Si nécessaire, on retire du modèle le coefficient ayant une probabilité élevée pour réaliser la prochaine itération.

- Qualité globale – Coefficient de détermination R^2. Ce coefficient permet de décrire la qualité de la régression et est calculé en effectuant la régression linéaire suivante :

$$Predicted\ RT = f(Experimental\ RT)$$

En pratique, on considère $R^2 > 0{,}8$ comme étant suffisant pour qualifier le modèle d'acceptable.

- Qualité globale – La distribution finale des résidus présente une tendance gaussienne, i.e. une répartition visuellement distribuée et centrée en zéro. Des tests statistiques existent pour estimer ceci, mais ne sont pas considérés ici puisque les déviations sont facilement observables sur le graphique des résidus. En pratique, on veillera à ce que la distribution ne présente pas de courbure ou de tendance marquées indiquant une faille du modèle considéré.
- La fenêtre de précision – telle que définie par les auteurs de l'étude, on indique le pourcentage de prédictions qui sont égaux ou inférieurs à une erreur donnée. Les auteurs indiquent donc la précision de leur modèle à ±35% puisque 95% des prédictions ont une erreur inférieure ou égale à ±35%. Ce résultat est différent du calcul usuel de l'intervalle de confiance qui considère la moyenne et d'écart-type. La fenêtre de précision est représentée sur les modèles par une zone violette.

2.3.2.2. *Analyse statistique*

Pour effectuer une analyse statistique, on choisit un mélange de composés qui se doit d'être le plus représentatif possible de la base de données. Du fait de la présence d'un peu moins de 20 000 composés dans cette base, fabriquer un mélange représentatif ayant un nombre limité de composés est une tâche complexe nécessitant de nombreuses itérations. Dans un souci pratique, on cherchera plutôt à déterminer quel espace de la base de données est représenté par le mélange considéré, et ainsi se limiter à un mélange de composés plus simples, avec des molécules qui *a priori* se détectent en ESI-Orbitrap. L'analyse statistique plus détaillée est disponible en Annexe II.

Le modèle initial possède 6 paramètres purement linéaires. L'ensemble des paramètres disponible dans la base de données représente plusieurs dizaines de paramètres, réduits à 8 seulement du fait de la connaissance des interactions sur la colonne. En effet, une colonne HILIC possède 3 types distincts de rétention : partage, liaisons hydrogènes et liaisons ioniques, comme présenté en partie 2.2.2. Ainsi, toutes les grandeurs telles que le nombre d'atome d'un certain type ou la taille de la chaine carbonée sont exclus d'emblée des paramètres testés. Ainsi, le logP et le logD, la charge partielle nette, le nombre de groupements donneurs et accepteurs de liaisons hydrogènes, le nombre de liaisons libres (i.e. liaisons qui permettent au groupement une rotation libre dans l'espace) et la masse de la molécule sont les paramètres initiaux considérés. Après réduction statistique, le processus renvoie respectivement 3 et 6 composantes pour le modèle acide et pour le modèle basique. Cependant, pour les deux modèles, on observe une distribution des résidus qui indiquerait la présence de termes non linéaires. Ainsi, différents essais ont permis de déterminer que l'inverse de la masse des composés permettait d'obtenir un bon modèle. Basé sur ce nouveau paramètre, le processus statistique est effectué à nouveau et un modèle à 6 et 7 paramètres, différents des paramètres de la publication, est déterminé. Les régressions sont présentées en Figure 50 pour le résultat du modèle à pH acide et en Figure 51 pour le résultat du modèle à pH basiques.

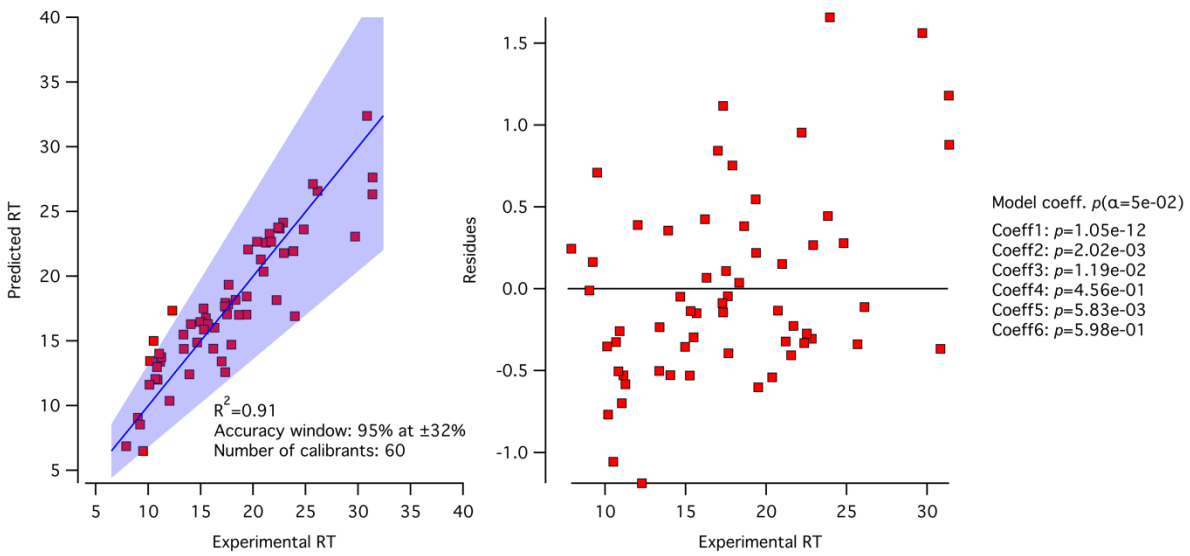


Figure 50 – Modèle pour le pH acide, 6 paramètres considérés dont un hyperbolique et une constante.

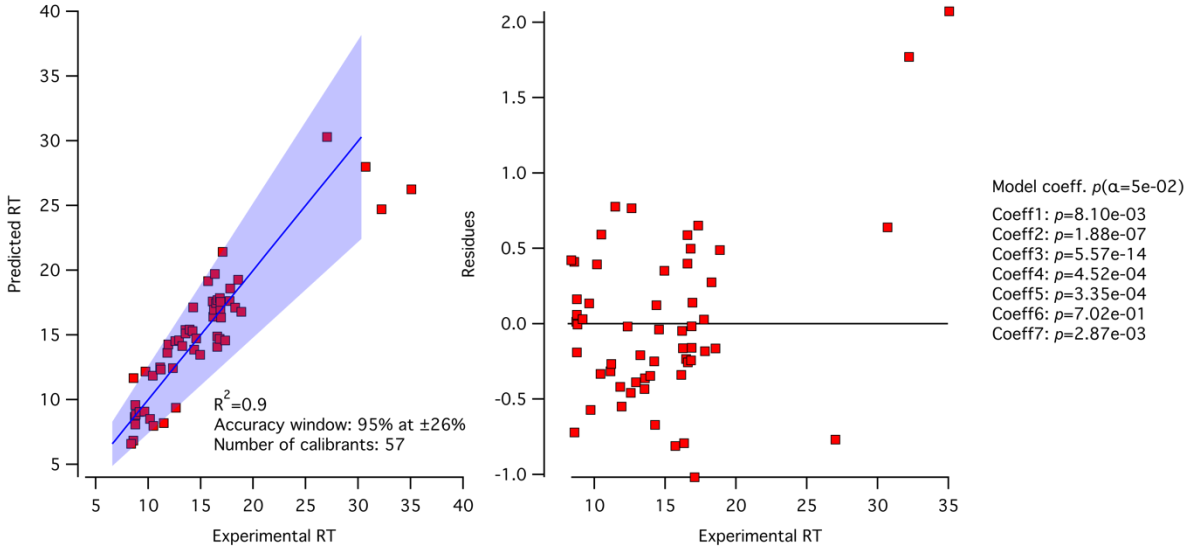


Figure 51 - Modèle pour le pH acide, 7 paramètres considérés dont un hyperbolique et une constante.

Ces modèles permettent de prédire avec une précision de respectivement ±32% et ±26% pour une méthode acide et une méthode basique. La seule condition est de s'assurer que le temps de rétentions de référence soient corrects, et nécessite alors une injection de mélange calibrant à chaque analyse. Cette injection peut également avoir la même utilité qu'une injection d'un mélange de type « contrôle qualité », et ainsi servira à valider la bonne fonction de la colonne avant son utilisation pour un échantillon à analyser. On s'assurera également, vu que la liste des composés calibrant est limitée, que les molécules prédites aient leurs propriétés physico-chimiques qui soient incluses dans l'espace des propriétés physico-chimiques définis par la calibration. Toute molécule prédite qui possède au moins une propriété hors du domaine de prédiction doit alors clairement être identifiée et la confiance de son temps de rétention prédis doit être ajusté.

2.4. Conclusion

Ce chapitre a présenté en trois temps distincts : (1) le développement de méthodes en spectrométrie de masse, (2) le développement de méthodes en chromatographie, et (3) l'identification moléculaire.

Le développement de méthodes en spectrométrie de masse s'est articulé autour de l'optimisation des méthodes d'acquisition en Orbitrap, suivit de la définition de systématiques d'attribution et de validation des données pour l'analyse d'échantillons solubles en ESI-Spectrométrie de masse, puis a été étendu et révisé pour l'analyse en LDI-Spectrométrie de masse. Par la suite, une méthode systématique et détaillée de développement de méthodes en chromatographie a été présentée, puis mise en application par la présentation du processus et résultats de développement de méthodes dédiées à l'analyse d'échantillons synthétiques complexes. Enfin, il a été rappelé la complexité d'effectuer une identification complète de chaque composé d'un échantillon complexe, et l'outil de prédiction des temps de rétention est alors introduit pour réduire l'espace des molécules possibles et ainsi pouvoir si nécessaire, effectuer des identifications dans un temps et un coût raisonnable.

3. Traitement des données : développement d'un algorithme de traitement des données pour la chromatographie en phase liquide

L'analyse non-ciblée d'échantillons organiques complexes par chromatographie couplée à la spectrométrie de masse haute résolution nécessite des traitements de données spécifiques, sans apriori sur la nature des composés présents. À notre connaissance, aucun logiciel commercial n'est disponible pour cette problématique, l'objectif de la plupart d'entre eux étant d'analyser des échantillons biologiques pour les sciences omiques. Plusieurs logiciels gratuits existent mais sont des boites noires en première approximation, comme MZMine par exemple.

Comme présenté en partie 1.5, les objectifs de développement du logiciel sont les suivants :

- Traitement du bruit et alignement des spectres
- Création d'une carte *m/z* en fonction du temps (carte ionique)
- Détection et modélisation des signaux

Les fonctions relatives à la génération des formules stœchiométriques et aux éventuels traitements spectre par spectre, présentées succinctement en 1.3.5, sont déjà incluses nativement dans Attributor et n'ont donc pas à être développées.

Ce chapitre va détailler pas à pas le processus de traitement des données issues du couplage entre la chromatographie et la spectrométrie de masse haute résolution. Tout d'abord, les données issues de l'instrument doivent être reconstruites. Du fait de la taille des données brutes, il est alors nécessaire d'effectuer des traitements initiaux comme traiter le bruit ou rééchantillonner les spectres par exemple. Après la reconstruction des données, une carte ionique est obtenue et sera la base de travail pour la suite des traitements effectués. En chromatographie, l'intérêt de l'analyse est de générer un couple (masse ; temps) pour chaque signal. Pour ce faire, il faut alors être capable de détecter les signaux dans la carte ionique : c'est le rôle des algorithmes de détection des signaux. Le problème de ce type de détection de signaux est qu'il ne permet pas de résoudre les composés qui sont partiellement séparés par le processus chromatographique : on ajoute alors une étape de modélisation des signaux détectés, et on génère alors l'ensemble des informations nécessaires pour pouvoir interpréter l'analyse effectuée.

3.1. Reconstruction des données

Cette partie a été développée et programmée en collaboration avec F.-R. Orthous-Daunay.

3.1.1. Structure des données brutes, problématique

Les données de chromatographie couplée à la spectrométrie de masse sont extraites des fichiers RAW générés à la fin de l'acquisition. Trois tableaux sont suffisants pour représenter l'ensemble des données issues de la chromatographie couplée à la spectrométrie de masse :

- Un tableau index temps=f(masse) : l'information des masses détectées en fonction du temps. Chaque coordonnées (ligne ; colonne) représente une masse et un temps unique ;
- Un tableau index temps=f(intensité) : l'information des intensités détectées en fonction du temps. Chaque coordonnées (ligne ; colonne) représente une masse et un temps unique ;
- Un tableau index temps=f(temps, Total Ion Count) : tableau pour effectuer la correspondance entre le numéro de colonne et le temps réel d'acquisition. On associe également par convenance la valeur du TIC, pour référence.

Ces données sont donc chargées en l'état dans le module de traitement, générant le premier problème de traitement de données : l'alignement en masse. En effet, la génération des données empiles les lignes et les colonnes sans aucun regard pour aligner les mêmes masses sur une

même ligne, comme illustré en Figure 52. Cela engendre une complexité de traitement puisqu'il n'est pas possible d'assigner une ligne à une masse, rendant alors le traitement résultant lent. Il est donc nécessaire d'aligner en masse les tableaux de données.

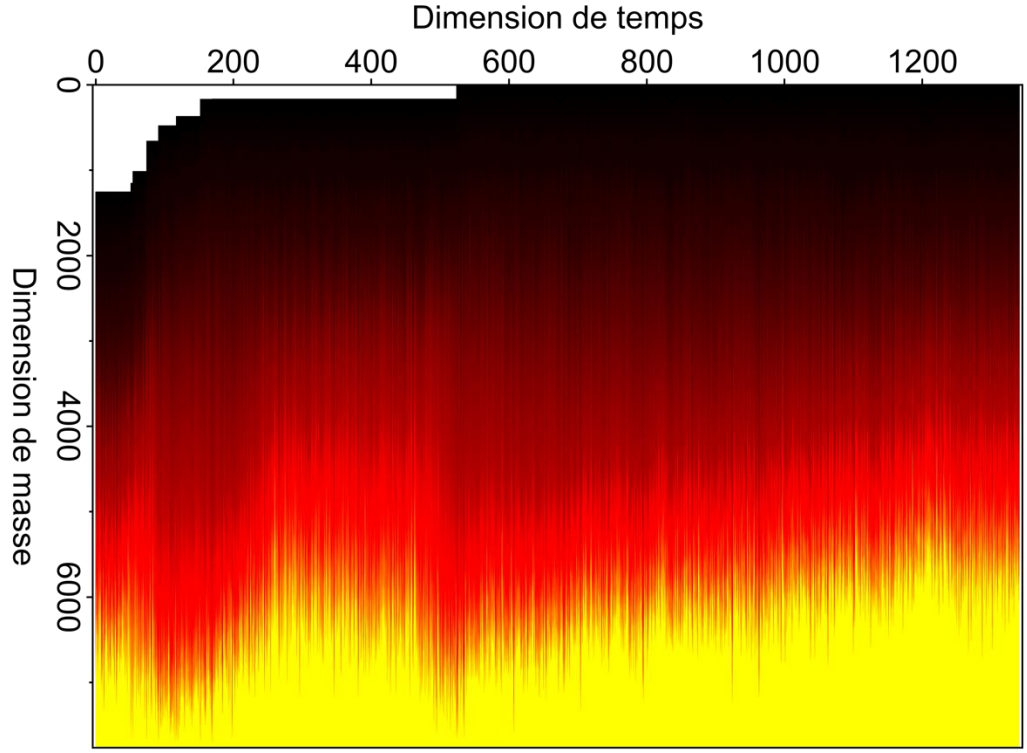


Figure 52 – Représentation d'un tableau de données brutes. Les zones jaunes et blanches représentent des zones où il n'y a pas de données, et sont ainsi remplies par défaut (NaN pour le blanc et dernière masse mesurée pour le jaune). On notera également que la taille des spectres en masse (i.e. longueur d'une colonne) est variable au cours du temps. Seuil de bruit fixé à 200 000.

Cet alignement est simple à effectuer d'un point de vue algorithmique, mais génère une quantité de données qu'il est nécessaire de pouvoir stocker dans la RAM de l'ordinateur. Pour représenter le problème concrètement, on considère l'analyse de l'échantillon présenté en partie 4.3, avec un traitement d'alignement en masse, acquis entre *m/z*=[50-550] Da et sur 40 minutes (purge et rééquilibrage de la colonne exclus du traitement) génère ~1500 colonnes et ~1 300 000 lignes, soit près de 2 000 000 000 de valeurs à stocker. En considérant un système 64bits et un stockage en « *double précision* », la taille du tableau de données résultant est de ~15Go de mémoire pour un unique tableau, sachant qu'il y en a deux (information de masse et d'intensité) et que le logiciel a également des données chargées. Cette utilisation de mémoire n'est pas compatible, voire même impossible avec un grand nombre de systèmes. Ainsi, il est nécessaire de réduire le nombre de données à considérer avant d'effectuer l'alignement des spectres.

Un autre problème, directement lié à l'instrumentation, est la non-constance de l'échantillonnage en temps. En effet, les spectromètres de masse à transformée de Fourier utilisent des « *bins* », qui séparent l'espace des masses, continu, en espace échantillonné et donc discontinu. Ainsi, une même masse sera échantillonné par des « *bins* » ayant une masse moyenne différente d'un temps à un autre, ce qui augmente également artificiellement le nombre de masses à considérer lors de l'alignement des spectres.

Ce décalage systématique de l'échantillonnage doit également être corrigé par l'algorithme d'alignement des spectres.

3.1.2. Réduction des données et échantillonnage en masse

Pour réduire les données, plusieurs voies peuvent être envisagées :

- Traitement du bruit : on annule et retire toutes les masses qui sont inférieures à un niveau d'intensité donné ;
- Traitement en temps : on demande à l'utilisateur d'indiquer la plage temporelle qui l'intéresse. En pratique, on retirera par exemple la purge et le reconditionnement de la colonne du traitement de données ;
- S'assurer d'un échantillonnage uniforme : après déduction du bruit, la duplication des masses représentant la même information est problématique et augmente de manière considérable le nombre de données à stocker. S'assurer ou construire d'un échantillonnage uniforme est donc primordial ;
- Stockage des données : il n'est peut-être pas nécessaire de stocker les données chromatographiques sous forme de nombre à « *double précision* ». On peut alors utiliser un stockage de la donnée sous 32bits, réduisant d'un facteur 2 la taille du tableau généré. Le problème d'utiliser un stockage 32bits est que l'ensemble du processus de masse déjà présent fonctionne en 64bits. Il est alors convenu de rester en 64bits pour tout ce qui concerne les données finales, et de passer en 32bits pour toute fonction intermédiaire ne modifiant pas les données ;
- Limiter la gamme de masse : réduire la quantité de masses uniques à considérer va grandement réduire la quantité d'information à stocker. Pour être optimale, cette action est une décision à prendre lors de l'acquisition des données, et a également une justification instrumentale du fait de l'AGC. En effet, réduire la gamme de masse en effectuant des analyses en SIM va augmenter la sensibilité sur cette gamme de masse restreinte, et ainsi obtenir plus d'informations sur l'échantillon. Pour une analyse complète, faire des fenêtres de masse de largeur 50 Da semble être efficace[47], mais nécessite une quantité d'échantillon suffisante pour effectuer de multiples analyses sur l'ensemble des gammes de masses par pas de 50Da.

Le traitement du bruit et le traitement en temps sont deux étapes distinctes qui doivent être effectuées dans le bon ordre : on coupe en temps et ensuite on traite le bruit. L'interface utilisateur n'a ainsi que 3 paramètres à indiquer lorsque les données sont chargées : les deux bornes temporelles et le niveau de bruit à considérer. C'est tout ce qui est nécessaire à priori pour reconstruire les données et effectuer la détection et la modélisation des signaux chromatographiques.

Le traitement en temps recherche les bornes temporelles indiquées par l'utilisateur et ne conserve que les données qui sont strictement incluses dans l'intervalle demandé. Les données hors de cet intervalle sont simplement supprimées des données prises en comptes par la suite. Le traitement du bruit repose sur l'algorithme « FAT Noise » développé sous Attributor pour le traitement des spectres de masses en infusion directe. Cet algorithme renvoie un seuil de bruit pour chaque spectre, déterminé en considérant le premier maxima de la dérivée de l'histogramme des intensités. Ce point représente l'intensité à laquelle on suppose qu'apparait les signaux de type artefacts induits par l'instrument et la transformée de Fourier. Déterminer ce seuil pour chaque spectre relatif au pas temporel permet de reconstruire une distribution des seuils de bruits. On peut également remarquer, comme présenté en Figure 53, que la distribution des seuils de bruits est différente lorsqu'un échantillon complexe est analysé comparé à une analyse de quelques composés uniquement. Si l'on étudie les valeurs non rangées par ordre croissant (non représenté ici), alors on peut remarquer que le seuil de bruit est dépendant de si l'on observe du signal ou non, ainsi que du nombre et de l'intensité des signaux détectés.

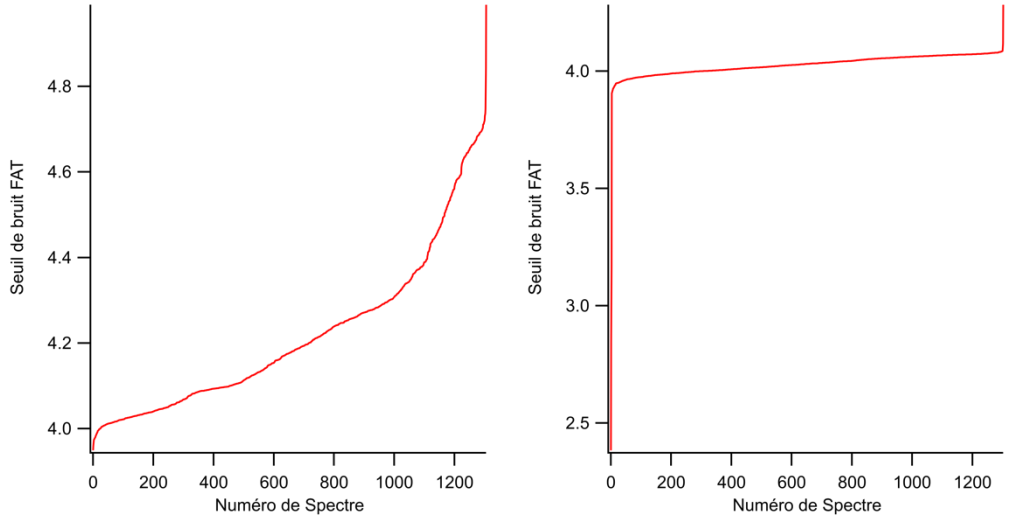


Figure 53 – Distribution des seuils de bruits FAT pour une analyse d'un échantillon complexe (gauche) comparé à une analyse de 12 standards (droite). Les valeurs sont rangées dans l'ordre croissant et les valeurs du seuil de bruit sont représentées en échelle log.

Cette différence notable indique pourquoi il est important, pour l'analyse des échantillons complexes, de considérer « un bon » niveau de bruit pour son analyse, puisque le placement du curseur de bruit a des conséquences importantes sur les informations disponibles après le traitement. Par expérience, on peut considérer qu'un bon seuil de bruit se situe autour du seuil moyen + trois fois l'écart type du seuil de bruit. Cela permet souvent d'obtenir un traitement chromatographique initial, et par la suite, de descendre manuellement le niveau de bruit si nécessaire. Dans certains cas, si la variation n'est pas très importante, il peut être nécessaire de monter le niveau de bruit pour obtenir un traitement car trop d'information est alors prise en compte.

Après traitement du bruit (i.e. détermination du seuil et retrait des informations situées en dessous de ce seuil), on effectue un « rebinage », c'est-à-dire que l'on va déterminer mathématiquement le pas moyen de chaque bin utilisé pour l'acquisition des données, et échantillonner de nouveau chaque spectre en utilisant ce pas moyen. Par construction de l'instrument, le pas d'échantillonnage semble être une variable discrète et quantifiée, comme montré en Figure 54. Cela pose un problème quant à la détermination mathématique de ce pas ; il est donc choisi de considérer l'enveloppe externe inférieure de cette distribution comme étant le pas moyen d'échantillonnage. Pour permettre une détermination suffisamment précise, il est nécessaire ici d'avoir suffisamment de données, et donc de ne pas choisir un niveau de bruit trop restrictif ou une gamme temporelle trop courte.

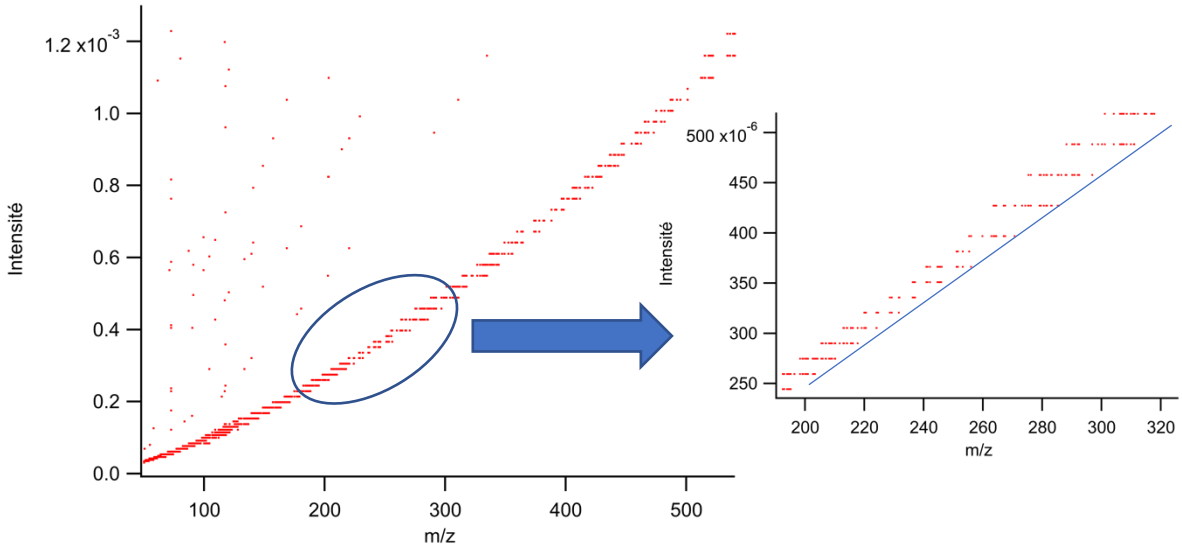


Figure 54 – Représentation de la différence de masse en fonction de la masse. Le pas d'échantillonnage semble ainsi être une variable discrète et quantifiée. La ligne bleue représente approximativement l'enveloppe inférieure de la distribution. Seuil de bruit fixé à 200 000.

L'erreur effectuée de la sorte est relative à un sur-échantillonnage des données par l'application d'un pas plus petit que la majorité des pas moyens, mais à l'avantage de ne pas faire perdre en résolution l'analyse en masse. L'évaluation de l'enveloppe inférieure est effectuée en considérant pour chaque masse, la plus petite différence de masse présente, et un polynôme de degré 3 est utilisé pour modéliser le nouveau pas d'échantillonnage. L'échelle de masse qui sera utilisée par la suite est ainsi recalculée à partir des coefficients issus de la modélisation et de toutes les masses mesurées.

Pour illustrer l'alignement des spectres, on présente en Figure 55 un signal en masse avant et après rebinage. Même si après rebinage, le spectre est en dents de scies, on note que les masses sont désormais alignées pour les différents temps, créant 10 masses uniques pour ce signal en particulier au lieu des 40+ masses uniques avant traitement. Un traitement spécifique pour traiter le spectre en dents de scies est effectué plus tard dans le processus, et est donc ignoré à ce point.

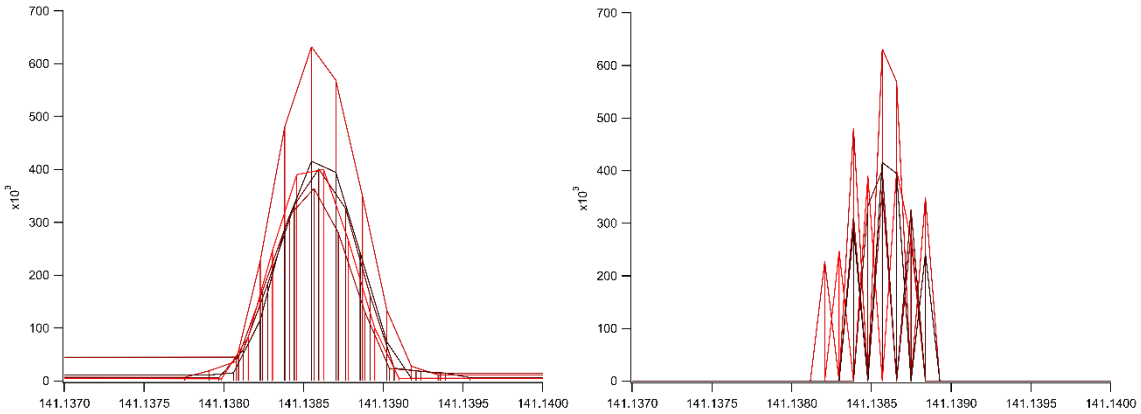


Figure 55 – Illustration du processus de rebinage sur cinq spectres consécutifs.

3.1.3. Reconstruction des données, création de cartes

Une fois les données réduites et la nouvelle échelle de masse déterminé, les données doivent être reconstruites. En s'appuyant sur la nouvelle échelle de masse, l'ensemble des données en masse sont recalculées en se basant sur cette échelle, et une carte alignée en masse est ainsi

générée. L'alignement en masse nécessite d'effectuer un remplissage pas à pas de la nouvelle carte en extrayant une par une chaque information mesurée, calculer la nouvelle masse et insérer l'intensité dans l'index de la nouvelle échelle de masse. À la différence d'une carte comme présenté en Figure 52 où chaque ligne ne représentait pas une masse, on construit ici une réelle carte où l'intensité est représentée en fonction de l'index temporel et de l'index en masse. Un exemple de carte reconstruite comparée aux données brutes est fourni à titre d'illustration en Figure 56. Il est à noter que seuls un traitement de bruit et un rééchantillonnage des masses a été effectué entre ces deux cartes. La dimension des masses et l'index des masses sont ici de manière fortuite similaires ; cette similitude n'est pas vérifiée pour l'ensemble des exemples disponibles et est grandement dépendante du niveau de bruit choisi.

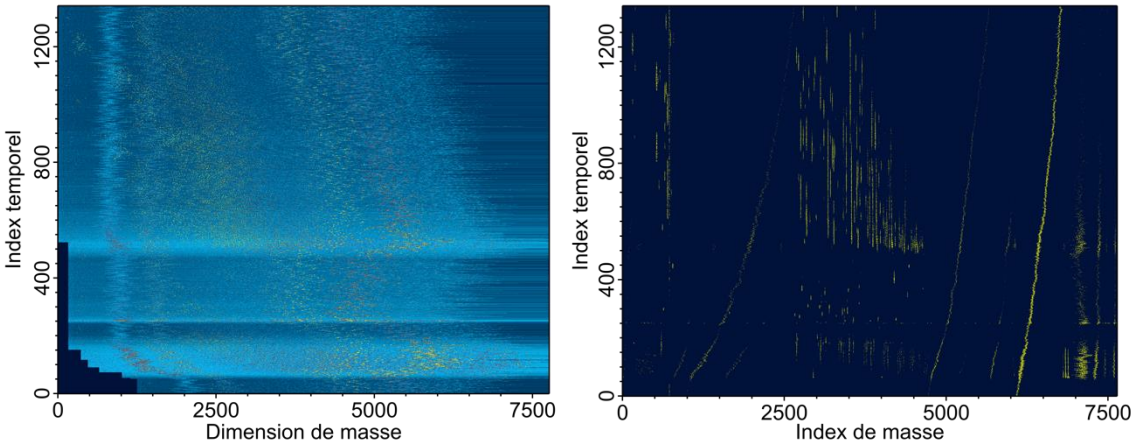


Figure 56 – Comparaison de la carte en intensité avant (gauche) et après (droite) traitement du bruit et rééchantillonnage en masse. Seuil de bruit fixé à 200 000.

Du fait du sur-échantillonnage introduit précédemment, il faut néanmoins s'assurer que les données soient correctes et ne présentent pas de déformation importante. Il est présenté en Figure 57 Gauche, un zoom sur une région contenant plusieurs signaux chromatographiques et l'on peut remarquer qu'en fonction du temps, des pixels manquent alternativement sur l'ensemble des signaux. Deux solutions sont possibles : changer le rééchantillonnage des masses ou effectuer un traitement de l'image pour combler les trous générés. Du fait que les manques sont seulement d'un pixel et rarement de plusieurs pixels à la suite pour un même temps, il est possible de combler le trou de données générées en faisant de l'interpolation linéaire sur les données. Le résultat est présenté en Figure 57 Droite.

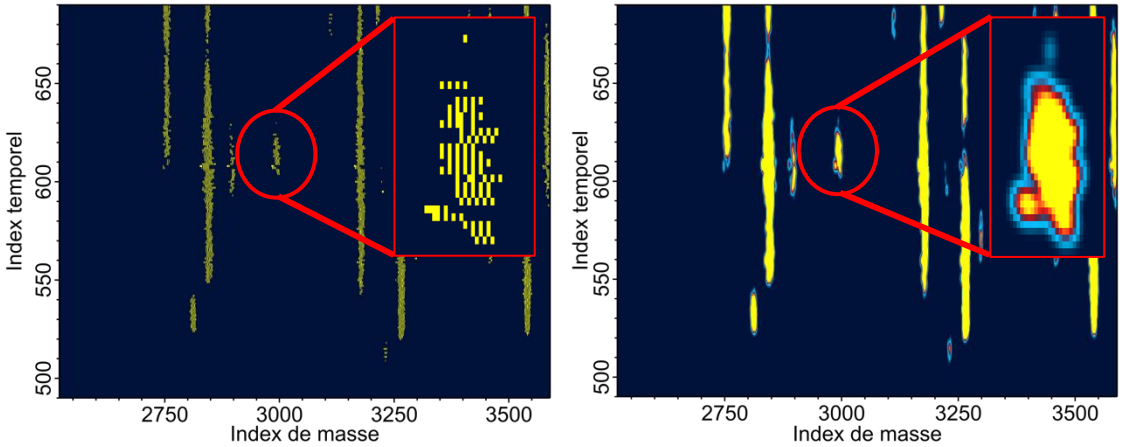


Figure 57 – Illustration du sur-échantillonnage (gauche) et du traitement correctif (droite). Seuil de bruit fixé à 200 000.

Cette carte est le produit final, et n'est aucunement modifié par la suite. Avant même d'effectuer du traitement d'image (détection et modélisation des signaux), il est possible de

générer avec ce traitement des données de spectrométrie de masse, avec ou sans information temporelle. Ce genre de vérification peut permettre de vérifier s'il y a des données (DMvM) ou encore d'effectuer une attribution des signaux en masse et comparer à l'infusion directe. En effet, on peut générer un spectre de masse moyen pour une plage temporelle, ou pour un temps donné. Ces informations permettent à priori de pouvoir se faire une idée sur a séparation chromatographique, avant même d'effectuer des traitements de données et de pouvoir évaluer la qualité du traitement de données effectuée. En Figure 58 est illustré un spectre moyen (profil) et son diagramme de défaut de masse associé pour l'échantillon utilisé dans les illustrations précédentes, zoomé sur la partie *m/z*=[50-300] Da qui présente le plus de données du fait du niveau de bruit sélectionné (200 000). Ces données reconstruites sont comparées aux données issues de l'infusion directe pour estimer s'il existe ou non des biais du fait de l'analyse. On peut observer des valeurs de défaut de masse plus basses dans les données reconstruites que dans les données d'infusion directe, ainsi qu'un chevauchement complet des masses les plus intenses entre les deux jeux de données. La présence de points supplémentaires est dû à la différence de méthode d'acquisition entre l'infusion directe et les données reconstruites à partir de la chromatographie, le bruit et la sensibilité étant différent en utilisant des micro-scans ou en effectuant la moyenne de scans avec un signal variable au cours du temps.

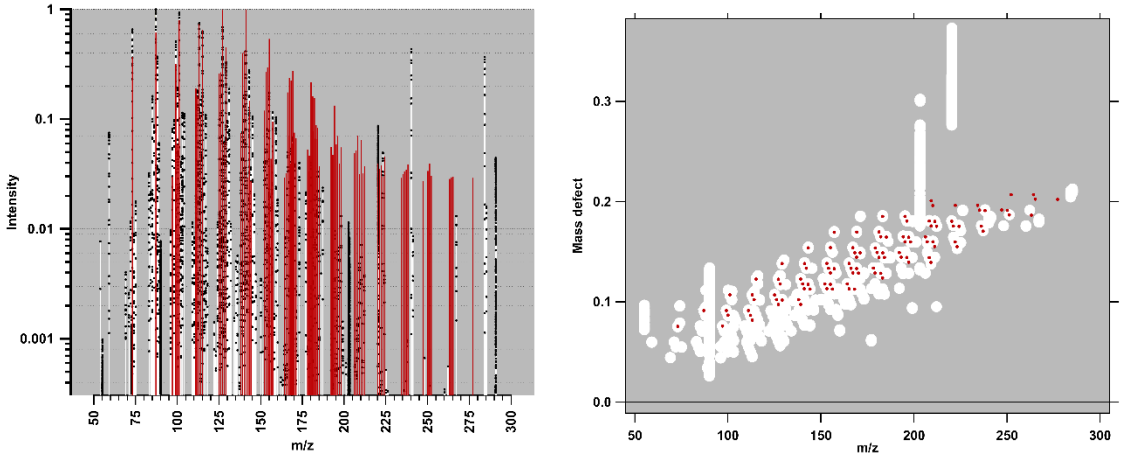


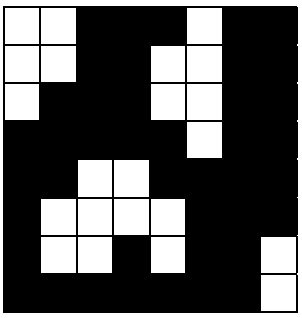
Figure 58 – Données reconstruite à la suite du traitement de données (en blanc) comparé aux données obtenues en infusion directe (en rouge). Seuil de bruit fixé à 200 000.

3.2. Outils algorithmiques pour la détection des signaux chromatographiques

3.2.1. Détection des ilots en masse/temps – Hoshen-Kopelman

Cette partie du code a été développée par F.-R. Orthous-Daunay, ma contribution est sur la définition des besoins et sur le filtrage des résultats.

Une fois la carte reconstruite, que l'on appellera par la suite *Carte Ionique*, on doit trouver et caractériser les signaux chromatographiques. Il faut alors construire une liste des signaux, et définir leur étendue temporelle et en masse : c'est le rôle de l'algorithme d'Hoshen-Kopelman. Cet algorithme a été publié en 1976 [48] et permet de partitionner le réseau considéré en dénombrant le nombre d'amas d'un même type dans une matrice finie. Autrement dit, cet algorithme nous permet d'assigner à un signal chromatographique l'ensemble de ces pixels constitutifs et de les indexer sous un identifiant unique. Pour simplifier le problème de chromatographie, estimons que la carte ionique que nous avons est une matrice de valeurs, représenté en Tableau 14, qui indiquent « j'ai du signal » ou « je n'ai pas de signal », représentés respectivement comme étant une case blanche et une case noire. L'algorithme va alors commencer sur la ligne du haut, case de gauche, et vérifie de gauche à droite, case par case, la valeur de gauche et celle du dessus. Une fois la ligne terminée, on recommence le processus sur la seconde ligne, case de gauche, etc jusqu'à ce que l'algorithme soit passé sur l'ensemble des cases.



Index en masse

Index en temps

1	1	0	0	0	2	0	0
1	1	0	0	2	2	0	0
1	0	0	0	2	2	0	0
0	0	0	0	0	2	0	0
0	0	3	3	0	0	0	0
0	3	3	3	3	0	0	0
0	3	3	0	3	0	0	4
0	0	0	0	0	0	0	4

Tableau 14 – Illustration du fonctionnement de l'algorithme de détection des signaux chromatographiques. Le code couleur sur la figure de droite est pour simplifier la lecture de la matrice. Les flèches rouges indiquent le sens de lecture de l'algorithme.

Plusieurs cas sont alors à prendre en compte, I[case] étant la valeur de l'intensité sur le pixel considéré, I[gauche] l'intensité du pixel situé à gauche et I[dessus] l'intensité du pixel situé au-dessus :

- Si I[case] est nulle : l'index de la case est égal à lui-même ;
- Si I[case] est non nul et I[gauche] et I[dessus] sont nuls : l'index de la case est égal à lui-même ;
- Si I[case] et I[gauche] sont non nuls, et que I[dessus] est nulle : l'index de la case est identique à l'index de la case de gauche ;
- Si I[case] et I[dessus] sont non nuls, et que I[gauche] est nulle : l'index de la case est identique à l'index de la case du dessus ;
- Si les trois intensités sont non nulles :
    - Si l'index de la case du dessus est identique à l'index de la case de gauche : l'index de la case est identique à l'index des deux cases ;
    - Si l'index de la case du dessus est différent de l'index de la case de gauche : cela indique que l'ilot du haut et celui de gauche doivent être fusionnés. Pour ce faire, on change l'ensemble des index de gauche et du dessus en un nouvel index et l'index de la case est alors identique à ce nouvel index.
- S'il n'y a pas de case à gauche et/ou au-dessus : on ne considère que la présence de la case présente et on applique les deux premières règles.

Une fois cette itération terminée, on note que, du fait de la première règle, l'ensemble des cases ayant une intensité nulle ont toutes un index différent. On réindexe alors les cases sans signal sous un index unique et on obtient alors la carte ionique des index, comme représenté en Figure 59. En pratique, est considéré comme étant un signal chromatographique tout signal qui est au moins entouré d'une ligne continue de pixels sans intensité. Ce signal est alors appelé « ilot ».

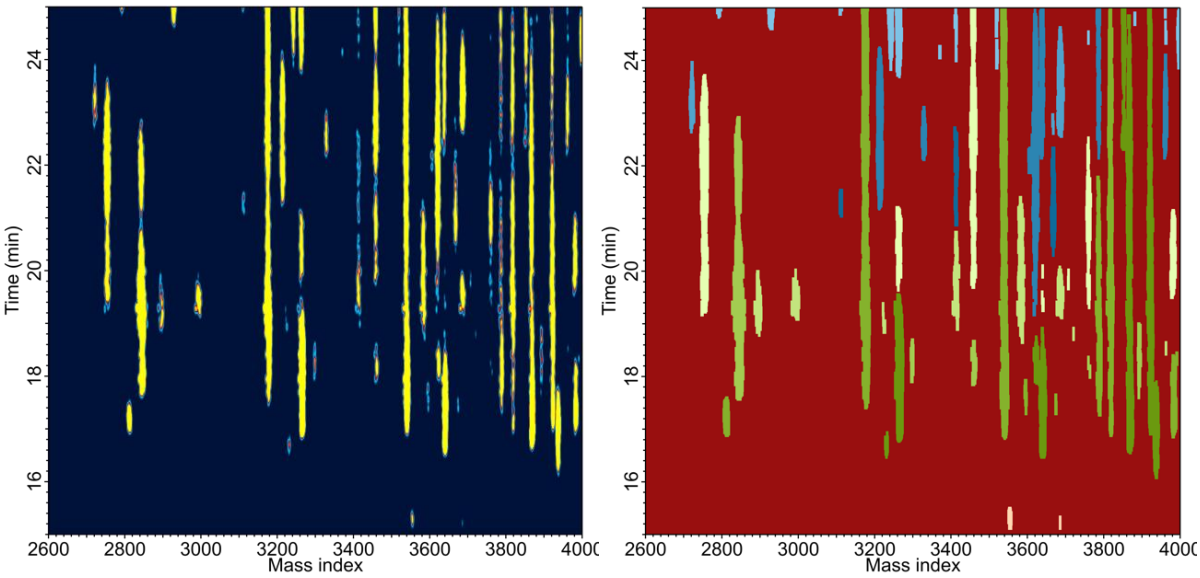


Figure 59 – Illustration du traitement de Hoshen-Kopelman sur une carte ionique. (Gauche) est la carte ionique avant traitement, (Droite) est la carte ionique après traitement. Chaque signal chromatographique qui est séparé par au moins une suite de pixels continus sans intensité est codée d'une couleur différente.

En plus d'une carte, cet algorithme permet d'établir une liste de l'ensemble des ilots détectés (i.e. dont l'index est différent de l'index du fond) ainsi qu'un ensemble de paramètres chromatographiques d'intérêts, tels que l'étendue en temps et en masse du signal, sa masse moyenne ou encore le temps associé au maximum d'intensité. Cette liste va permettre d'effectuer un premier tri des signaux chromatographique selon les critères arbitraires suivants :

- On retire l'ensemble des ilots qui ont en étendue en masse nulle, i.e. dont l'étendue en masse ne représente qu'un unique pixel sur l'échelle des masses. Conserver ses ilots reviendrait à dire que l'on a une résolution infinie avec un Orbitrap, ce qui n'a pas de sens ;
- On retire l'ensemble des ilots qui ont moins de 10 pixels d'étendue chromatographique. Puisque les ilots doivent être modélisés plus tard, on s'assure que l'on a suffisamment de points dans la dimension temporelle pour effectuer le traitement ;
- On retire l'ensemble des ilots qui ont moins de 30 pixels au total. Ces ilots sont beaucoup trop faibles en intensité pour être interprétés par la suite.

À la suite de cette détection, on utilise le résultat de ce traitement pour sauvegarder une carte qui ne contient que les ilots détectés, sauvegardant ainsi de l'espace en mémoire. Cette opération est réalisée en linéarisant la carte en un vecteur unique d'intensités non nulles, et est adossé à l'échelle en masse réalisée lors du rééchantillonnage. C'est sur ce vecteur, convertible en carte, qu'est effectué la suite des traitements et où est stocké l'information relative aux signaux chromatographiques.

L'avantage pratique de ce traitement est qu'il est rapide (6 secondes de traitement pour une carte de plus de 1 000 000 pixels) et qu'il permet sans ambiguïté de détourer chaque ilot de la carte ionique. D'autres algorithmes de traitement du signal ont été testés par ailleurs et n'ont pas donné les performances atteintes par cet algorithme.

3.2.2. Déconvolution des signaux chromatographiques – Expecta-Maxima

Cette partie du code a été développée par F.-R. Orthous-Daunay, ma contribution est sur la définition du problème et des besoins. Cette partie du code ne fonctionne pas encore correctement et nécessite donc des ajustements et des optimisations pour accélérer les traitements. Elle est néanmoins mentionnée ici car la définition du problème et la solution

générale sont là, il suffit de régler l'application. Des pistes d'amélioration sont détaillées à la suite.

Bien que l'Orbitrap ait une résolution importante, la complexité de la matière analysée fait qu'il peut arriver que des ilots soient convolués en masse, mais séparés en temps, ou alors que la résolution en masse soit juste insuffisante pour un retour propre à la ligne de base. Cela crée des problèmes avec la détection des signaux Hoshen-Kopelman, puisque ces ilots sont alors un seul et unique ilot détecté, au lieu des multiples ilots que l'on peut voir à l'œil nu lorsque l'on vérifie les cartes. Un exemple est fourni en Figure 60 où l'on distingue clairement les différents ilots en intensité, mais du fait de l'algorithme, on ne peut alors détecter qu'un unique ilot contenant plusieurs signaux.

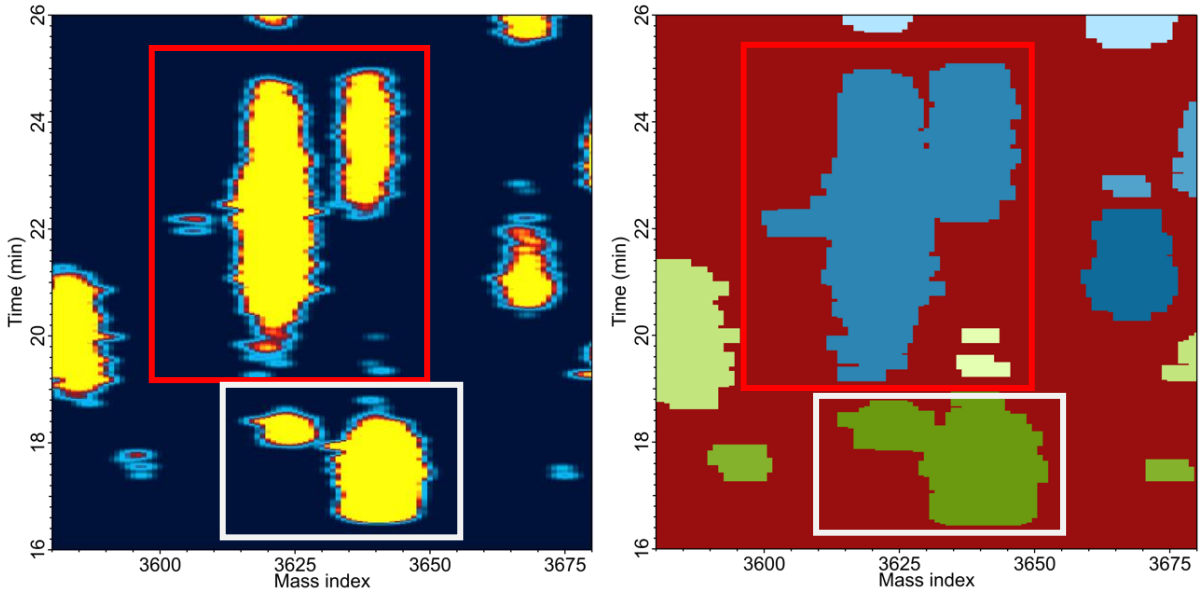


Figure 60 – Illustration de la perte de résolution du fait de l'algorithme de détection.

Ainsi, pour solutionner ce problème, il est nécessaire de prendre en compte l'intensité des ilots et d'effectuer une modélisation en 3D de l'espace [masse ; temps ; intensité]. Une approche simple est d'approcher les signaux chromatographiques par des gaussiennes en 3D. Si plusieurs gaussiennes sont ajustées de cette manière, on peut alors réattribuer les pixels dans l'espace [masse ; temps] à chacune des gaussiennes projetées, et ainsi résoudre les ilots convolués de cette manière.

L'approche actuelle est une approche de classification non-supervisée, qui cherche à déterminer automatiquement le nombre de gaussiennes à considérer et d'ajuster ce nombre pour rendre l'erreur d'ajustement la plus faible possible. Un problème récurrent dans nos données est que la convolution est souvent la combinaison d'un signal intense et d'un signal plus faible, et l'ajustement des gaussiennes retire la gaussienne de plus faible intensité dans la plupart des cas. Il faudrait donc passer à un algorithme supervisé qui initie le nombre de gaussiennes en fonction du nombre de pics en masse, et qui ne permet pas, ou difficilement, à l'ajustement de retirer les gaussiennes peu significatives. On pourrait également considérer la même supervision dans l'espace temporel, et ainsi effectuer les étapes suivantes (i.e. détection et modélisation des signatures temporelles) en une unique étape. Cela nécessite cependant des modifications importantes du code, qui n'ont pas pu être effectuées avant l'écriture de ce manuscrit.

De récentes investigations ont également montré que ce problème de convolution des signaux pourrait être générés par le traitement que nous effectuons, et particulièrement du fait de l'algorithme de reconstruction de la carte à partir d'un vecteur unique. En effet, à partir de la nouvelle échelle de masse issue du processus de rééchantillonnage, seules les masses qui ont une intensité non nulle sont considérées, et ainsi vectorisées. Cela veut dire que des ilots séparés

en masse se retrouvent concaténés directement l'un à côté de l'autre sans aucun moyen de les distinguer mis à part en vérifiant si l'on observe un saut en masse à l'intérieur du signal, comme présenté en Figure 61. On remarque également que les bords des pics présentent également un saut en masse, qui est très surement dû au même problème ainsi qu'à l'algorithme de comblement qui est utilisé.

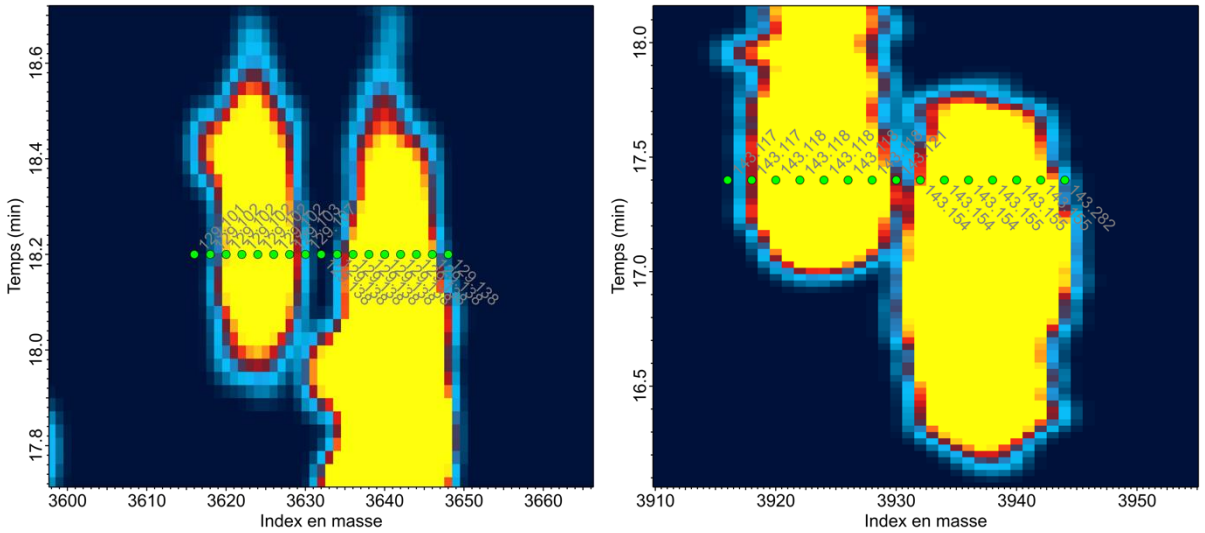


Figure 61 – Illustration d'un saut en masse à l'intérieur de deux différents ilots. On représente par des points verts les différentes masses associées à chaque pixel, et on note la masse de ce pixel. Le saut de masse entre chaque pic est symbolisé par le changement de sens de l'affichage des masses : passage de 129.103±0.002 à 129.138±0.001 pour le premier ilot, et de 143.118±0.002 à 143.155±0.004.

La solution dans ce cas est d'ajouter dans la dimension de masse des masses d'intensité nulle autour de chaque cluster en masse, permettant ainsi d'encapsuler les signaux de la carte dans un rectangle de masse d'intensité nulles, et ainsi de séparer effectivement les ilots lors des traitements par la suite. Cette correction ne peut pas être effectuée dans les temps pour ce manuscrit, et la résolution de ce problème ne rend pas forcément caduque le traitement par Expecta-Maxima tel qu'envisagé. En effet, on s'attend néanmoins à ce que la résolution instrumentale à plus haute masse ne soit pas forcément suffisante pour séparer correctement les ilots. Ainsi, il convient de corriger la façon de vectoriser et de reconstruire la carte mais également de s'assurer à plus haute masse que l'on puisse déconvoluer proprement les ilots si nécessaire.

3.2.3. Détection et modélisation des signatures temporelles

Cette partie représente deux développements distincts du code : la détection des signaux dans la dimension temporelle et leur modélisation. La détection a fait l'objet d'un travail en collaboration avec F.-R. Orthous-Daunay tandis que la modélisation a été définie et réalisée seul.

Une fois les signaux détectés, et déconvolués lorsque l'algorithme sera au point, on estime alors que chaque signal détecté ne représente qu'un unique signal en masse. L'objectif est alors de déterminer le nombre de signaux dans la dimension temporelle, *i.e.* le nombre d'isomères qui sont suffisamment séparés pour que l'on puisse détecter plusieurs pics en temps. En chromatographie, ajuster des signaux temporels est fait depuis plusieurs années par l'utilisation de Gaussiennes convoluées avec une exponentielle (*Exponentially Modified Gaussian*, EMG)[49,50]. L'avantage de ce type de fonction est que la composante exponentielle permet d'approcher les asymétries de pics (*fronting*, *tailing*) générées par le système chromatographique. En effet, la saturation de la colonne va par exemple générer un relargage à décroissance exponentielle, justifiant le *tailing* du signal chromatographique associé. La forme gaussienne est, quant à elle, parfaitement justifiée par la distribution normale des parcours moyens dans une colonne remplie de particules ayant une taille moyenne constante. Ainsi,

chaque composant de la modélisation utilisant des EMG est explicable par la physique à l'origine de la séparation en chromatographie.

Cependant, comme n'importe quel ajustement, déterminer à priori le nombre d'EMG à considérer pour ajuster un signal n'est pas chose aisé. Sur-ajuster avec trop d'EMG donnera des résultats satisfaisant d'un point de vue mathématique, mais ne permettra pas d'interpréter les données par la suite. Il est dont critique de pouvoir classifier les signaux temporels, et ainsi pouvoir générer des conditions initiales d'ajustement qui permettent une interprétation des résultats par la suite. Des algorithmes relatifs aux calculs topologiques permettent de fournir les informations dont on a besoin, tel que l'algorithme *Persistent Homology*. Ce genre d'algorithme a été décrit dans les années 1990[51] et fait l'objet de multiples extensions de code disponibles en libre accès.

L'idée de cet algorithme est de détecter les maximas et minimas locaux. Pour ce faire, on compare trois à trois chaque valeur en progressant dans le graphe. Plusieurs cas se présentent alors :

- Maximum local : lorsque le point a une valeur supérieure aux deux points proches ;
- Minimum local : lorsque le point a une valeur inférieure aux deux points proches ;
- Régime transitoire : lorsque l'ensemble des trois valeurs est dans une progression croissante ou décroissante.

Ce genre de logique est très sensible au bruit, générant de nombreux maximas et minimas locaux. Pour régler ce problème, la notion de « persistance » et de « mort » est introduite. Chaque maximum local se voit attribuer une valeur de « naissance » égale à son intensité tandis que chaque minimum local a une valeur de « naissance » nulle. Pour les valeurs de « mort », deux cas se présentent, basé sur la valeur des maximas adjacents :

- Si la « naissance » du maximum de droite est supérieure à celle de gauche, alors on définit la valeur de « mort » du point de gauche comme étant égale à l'intensité du point considéré ;
- Si la « naissance » du maximum de gauche est supérieure ou égale à celle de droite, alors on définit la valeur de « mort » du point de droite comme étant égale à l'intensité du point considéré ;

La persistance est alors définie comme la différence entre les valeurs de « naissance » et de « mort » : « un maximum avec une persistance élevée est un sommet ayant subi peu de mort ». Une illustration du processus de définition de la persistance est présentée en Figure 62.

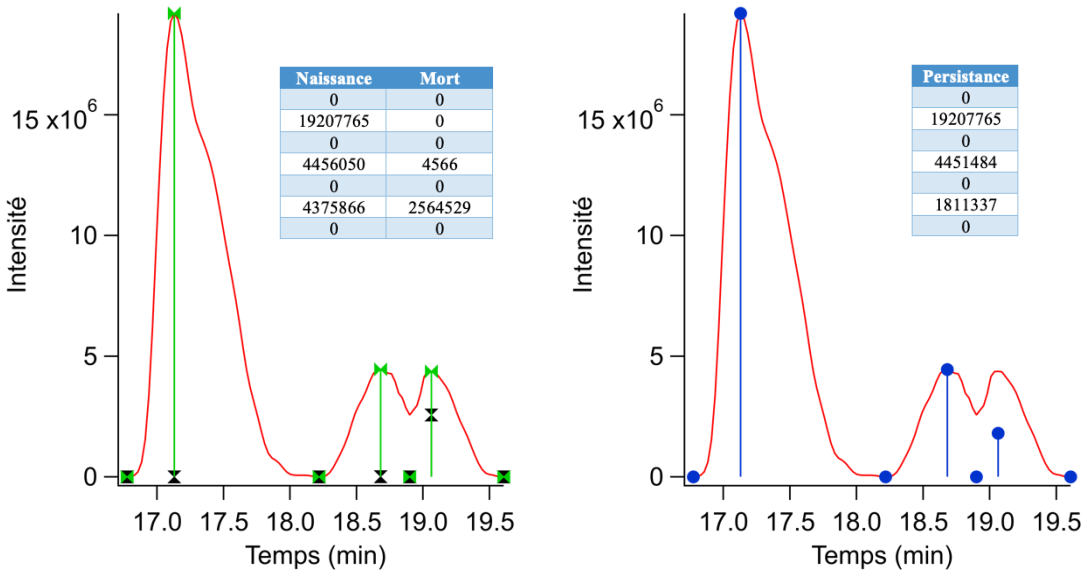


Figure 62 – Simulation d'un signal chromatographique et de la détection temporelle effectuée. (Gauche) représentation des valeurs de « naissance » (vert) et de « mort » (noir). (Droite) représentation de la persistance.

Une fois les maximas détectés et leur valeur de persistance connue, il est dès lors facile d'effectuer un filtrage arbitraire sur les valeurs de persistance pour ne conserver que les sommets suffisamment définis. Dans notre cas, on définit que tout maximum ayant une valeur de « naissance » supérieure à deux fois sa valeur de « mort » est conservée. Cela génère des problèmes, notamment dans l'exemple considéré puisque le maximum le plus à droite est alors détruit. C'est pour cela que l'on donne une interface permettant de changer ses valeurs pour tous les signaux et ainsi récupérer ses ratés. Néanmoins, ce filtrage agressif permet de retirer très efficacement le bruit et les épaulements non significatifs des détections, et ainsi d'avoir des conditions initiales plus robustes pour la modélisation.

Une EMG est définie selon Équation 10, mais cette formule n'est pas applicable directement pour la chromatographie car le coefficient λ est confondu avec l'asymétrie du pic et sa hauteur, deux paramètres distincts nécessaires à la chromatographie :

$$EMG(x;\mu;\sigma;\lambda) = \frac{\lambda}{2} e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2x)} erfc\left(\frac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right)$$

Équation 10 – Définition mathématique de l'EMG. La fonction erfc est la fonction « erreur complémentaire »

Pour effectuer la modélisation, on utilisera plutôt la forme présentée en Équation 11, qui permet de fournir la hauteur H, la largeur l, l'asymétrie a et le temps c :

$$EMG(x;H;c;l;exp) = He^{\frac{a}{2}(2c+al^2-2x)} erfc\left(\frac{c+al^2-x}{\sqrt{2}l}\right)$$

Équation 11 – Définition de l'EMG utilisée pour les conditions initiales. Le facteur d'intensité H, lié à la hauteur du pic remplace ici le facteur placé devant l'exponentielle et la largeur l remplace la valeur de l'écart-type σ. Le temps c est équivalent à la moyenne μ.

Bien que cette équation soit fausse, elle permet d'approximer très précisément les conditions initiales de l'ajustement et de converger en quelques cycles vers une solution. Par expérience, les valeurs de a, de H et de l sont les valeurs critiques qui permettent ou non la convergence de l'ajustement. Les meilleurs résultats, illustrés en Figure 63, sont obtenus avec :

- a est fixé à 1 ;
- H étant la moitié de la persistance ;
- l étant le dixième de la distance entre les deux minimas.

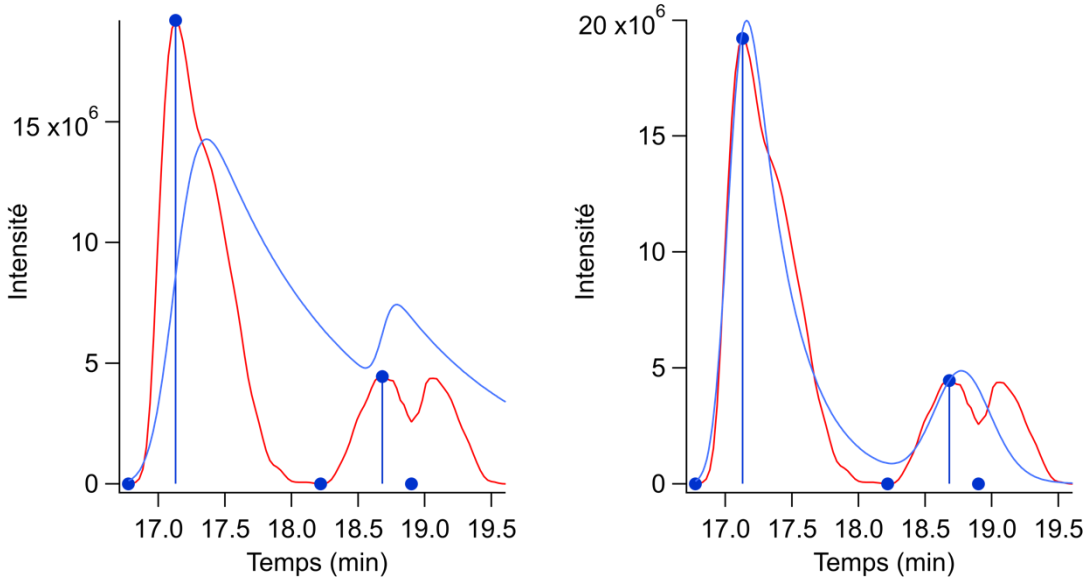


Figure 63 – Illustration des fit EMG initiaux (gauche) et finaux (droite).

L'ajout manuel du pic de droite permet d'effectuer une modélisation acceptable du signal, tel que présenté en Figure 64. L'expérience montre que cet algorithme fonctionne automatiquement pour 80% des signaux, les 20% restant étant du bruit et des artefacts instrumentaux qui ne sont pas détruits par les différents traitements et des cas de signaux tels qu'illustrés par l'exemple utilisé ici.

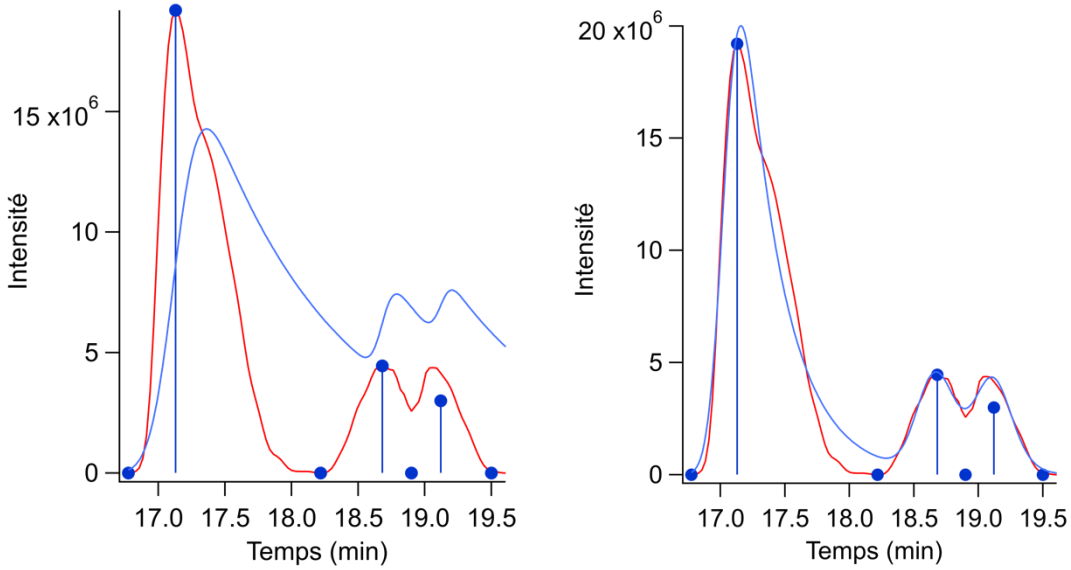


Figure 64 – Illustration des fit EMG avec l'ajout manuel d'un pic non conservé par la classification automatique.

À la suite de ce traitement, il faut extraire des modélisations les grandeurs chromatographiques nécessaires à l'interprétation des données : la hauteur du pic, son aire, son temps de rétention et son facteur de forme. La masse moyenne du pic est déjà connue étant donné les traitements de détection des signaux. Plusieurs solutions sont disponibles pour retrouver ses paramètres : (1) utiliser l'expression mathématique des EMG pour retrouver le temps au sommet, l'aire et la hauteur, (2) réutiliser Persistent-Homology pour retrouver le temps, la hauteur et faire calculer l'aire en utilisant le résultat de la modélisation ou (3) toute combinaison intéressante. En pratique, on utilise les coefficients pour retrouver les paramètres de hauteur et de temps au sommet, et l'aire sous la courbe modélisée pour retrouver l'aire du signal chromatographique. L'asymétrie du pic chromatographique n'est pas convertie et on utilise directement le coefficient issu de la modélisation qui donne déjà un ensemble d'informations suffisantes (*i.e.* quand a=1, le pic est gaussien).

Pour finir, chaque signal chromatographique est alors listé et est associé un couple [masse ; temps de rétention] qui permet d'identifier de façon unique chaque signal. Cette liste peut être exportée pour d'autres analyses, les masses extraites pour être attribuées ou encore le couple [masse ; temps de rétention] peut être comparé à une base de données et ainsi annoter voire identifier les signaux détectés.

3.3. Conclusion

Ce chapitre a présenté en détail le développement d'un logiciel de traitement des données issues du couplage entre chromatographie et spectrométrie de masse. Le traitement initial des données a été abordé, à partir du défi présenté par la quantité d'information jusqu'au traitement de bruit et de rééchantillonnage jusqu'à la génération de la carte ionique. Ce traitement initial est alors la base de l'ensemble des traitements suivants. Même si ce traitement est imparfait, puisqu'il génère la fusion de signaux en masses, il permet néanmoins un traitement fidèle des données qui ne sont pas impactées par ce problème.

Comme l'information d'intérêt pour pouvoir interpréter l'analyse est de générer une liste de couples (masse ; temps), il est alors nécessaire de détecter les signaux temporels : c'est le rôle de l'algorithme d'Hoshen-Kopelman. La détection des ilots ainsi effectuée ne prend pas en compte le problème de résolution des signaux, que ce soit en masse ou en temps. Ce problème fait que la détection d'isomères non résolus n'est pas possible directement : c'est alors le rôle de la modélisation sur l'ensemble des signaux détectés précédemment d'extraire l'information isomérique non résolue. On obtient alors la liste des couple (masse ; temps de rétention) et prête à être utilisée pour l'interprétation des données.

4. Applications aux échantillons organiques complexes : analyse d'analogues d'aérosols d'atmosphères d'exoplanètes

L'ensemble des développements proposés aboutissent à leur utilisation pour des échantillons organiques complexes d'intérêt pour la planétologie. Nous avons ainsi décrit précédemment le fonctionnement des spectromètres de masse et de la chromatographie, ainsi que du développement des méthodes de traitement des données associées. Ces méthodes sont complétées par un outil de prédiction des temps de rétention nécessaire à la réduction de l'espace des composés possibles pour les identifications. Nous avons également présenté le développement d'un logiciel de traitement des données chromatographiques dédiée.

Dans ce chapitre, les trois échantillons qui donnent des résultats en spectrométrie de masse sont analysés par différentes méthodes en infusion directe et par chromatographie. L'idée est d'avoir une image globale de la diversité moléculaire présente dans ce type d'échantillon, puis d'aller voir plus en détails les différentes fractions d'un échantillon en particulier. Enfin, l'analyse chromatographique permettra de commencer à déterminer quels isomères de molécules d'intérêt biochimique peuvent être annotées, voire identifiées et ainsi apporter des éléments supplémentaires à l'analyse effectuée par Moran et al. [26].

4.1. Choix des échantillons analysés

Des échantillons fournis par Sarah Hörst et son équipe de l'université de Johns Hopkins (Maryland, USA) [25–27,52] sont utilisés comme analogues de matière organique complexe. Cette expérience est présentée en partie 1.1.3.3 et vise à simuler des exoplanètes de type super-Terres et mini-Neptunes. De multiples compositions, listées en Tableau 15, ont été utilisées dans la chambre de réaction et les résidus récupérés pour analyse. La métallicité indiquée ici est le facteur multiplicatif pris en compte pour chaque élément (hydrogène et hélium exclus) comparé à la composition élémentaire du soleil [26]. Des calculs thermodynamiques estiment ensuite quelles molécules sont stables à cette composition et température et l'expérience est ainsi menée suivant ces ratios.

	Métallicité : 100×		Métallicité : 1 000×		Métallicité : 10 000×	
600K	***Insoluble***	72% H2 6,3% H_2O 3,4% CH_4 18,3% He	***Soluble***	42% H_2 20% CO_2 16% H_2O 5,1% N_2 1,9% CO 1,7% CH4 13,3% He	***Insoluble***	66% CO_2 12% N_2 8,6% H_2 5,9% H_2O 3,4% CO 4,1% H_2
400K	***Insoluble***	70% H_2 8,3% H_2O 4,5% CH_4 17,2% He	***Soluble***	56% H_2O 11% CH_4 10% CO_2 6,4% N_2 1,9% H2 14,7% He	***Soluble***	67% CO_2 15% H_2O 13% N_2 5% He
300K	***Soluble***	68,6% H_2 8,42% H_2O 4,51% CH_4 1,23% NH_3 17,24% He	***Soluble***	66% H_2O 6,6% CH_4 6,5% N_2 4,9% CO_2 16% He	***Soluble***	67,3% CO_2 15,6% H2O 13% N_2 4,1% He

Tableau 15 – Tableau des compositions et de leur solubilité dans le méthanol.

Ces échantillons d'analogues d'aérosols d'atmosphère d'exoplanètes servent à nous donner une idée de la diversité qu'il est possible d'obtenir en utilisant des mélanges plus complexes. En effet, les premières réactions de ce type ont servi à modéliser Titan, avec un mélange simple de N_2/CH_4 à $95/5\%_{vol}$ et ce genre de réactions en phase gaz est novatrice dans son approche multivariée du problème avec pas moins de neuf compositions différentes effectuées dans cette première étude [25–27,52] qui utilise de multiples gaz tels que CO, CO_2, H_2O, N_2, CH_4 ou encore H_2. Une série d'échantillon plus récente [Vuitton et al, 2020, accepté] a également incorporé du H_2S dans le mélange réactif, générant également une diversité intéressante. L'ensemble de ces données ne modélisent pas une atmosphère d'exoplanète et sa diversité réelle, cependant, elles permettent d'avoir des idées de la diversité potentielle dans le but de limiter les exoplanètes à observer aux seules présentant potentiellement une diversité *a priori* conséquente.

Seuls les échantillons avec une métallicité de mille sont solubles dans le méthanol et fournissent des résultats en infusion directe. L'échantillon « 300K – 100x » ainsi que les deux échantillon « 400K – 10 000x » et « 300K – 10 000x » semblent soluble dans le méthanol mais fournissent peu d'informations en spectrométrie de masse. Ils ne sont donc pas discutés par la suite.

L'avantage de tels échantillons, comparés par exemple à des échantillons issus de modélisation de résidus de glaces cométaires, est premièrement leur disponibilité en quantités importantes. En effet, alors que les synthèses de glaces ne permettent d'obtenir que quelques centaines de nanogrammes de matières, les expériences en phase gazeuse, comme PHAZER, permettent d'obtenir jusqu'à plusieurs centaines de milligrammes en fonction des mélanges réactifs utilisés. Cette quantité importante de matériel disponible fait que ces échantillons peuvent alors être utilisés à des fins de développement et d'essais divers, sans avoir à se limiter drastiquement du fait de la disponibilité de l'échantillon.

En complément de la disponibilité physique de l'échantillon, la diversité atomique et fonctionnelle est également à prendre en compte. En effet, toute expérience de synthèse ne permet pas d'obtenir un mélange complexe composé de molécules en CHNO(S) seulement, au moins partiellement soluble, et présentant des signaux complexes en spectrométrie de masse. Ce sont ces différents critères qui ont indiqué, parmi l'ensemble des possibilités, les échantillons d'analogue d'aérosols d'atmosphère d'exoplanètes pour développer et tester les méthodes proposées dans ce travail.

4.2. Analyses par spectrométrie de masse

4.2.1. Analyses par ESI-Orbitrap

Les paramètres utilisés pour les analyses ESI sont disponibles en 0. Les analyses ESI ont été acquises sur trois gammes de masses : [50-300] Da, [150-450] Da et [400-1000] Da en utilisant 4 scans de 128 micro-scans. Chaque spectre est acquis en polarité positive et négative et les données sont traitées selon la systématique présentée en partie 2.1.2.

4.2.1.1. *Exploration de l'échantillon*

Il est important de comparer visuellement les données avant leur interprétation et leur attribution complète. On présente les six spectres de masses normalisés en Figure 65. Par soucis de concision, seule la gamme [150-450] Da est discutée pour l'ensemble des traitements en ESI-Orbitrap. Les spectres de masses des autres gammes de masses sont disponibles en Annexe IV. Ces spectres présentent des motifs périodiques classiques pour les échantillons d'analogues d'aérosols atmosphériques, que ce soit en polarité positive ou négative. L'échelle de couleur est appliquée aux signaux représentant au total plus de 99% de l'intensité totale du spectre. Quelques signaux intenses sont présents et sortent de la tendance des échantillons, et ne sont à priori pas des signaux appartenant aux échantillons. L'échantillon 400K est celui qui présente

les signaux les plus intenses dans les deux polarités, indiquant possiblement une plus grande sensibilité, *i.e.* une plus grande diversité moléculaire possiblement détectée, bien qu'il présente la même gamme dynamique que les autres analyses. Le concept de gamme dynamique est alors à prendre avec précautions, et sera discuté par la suite.

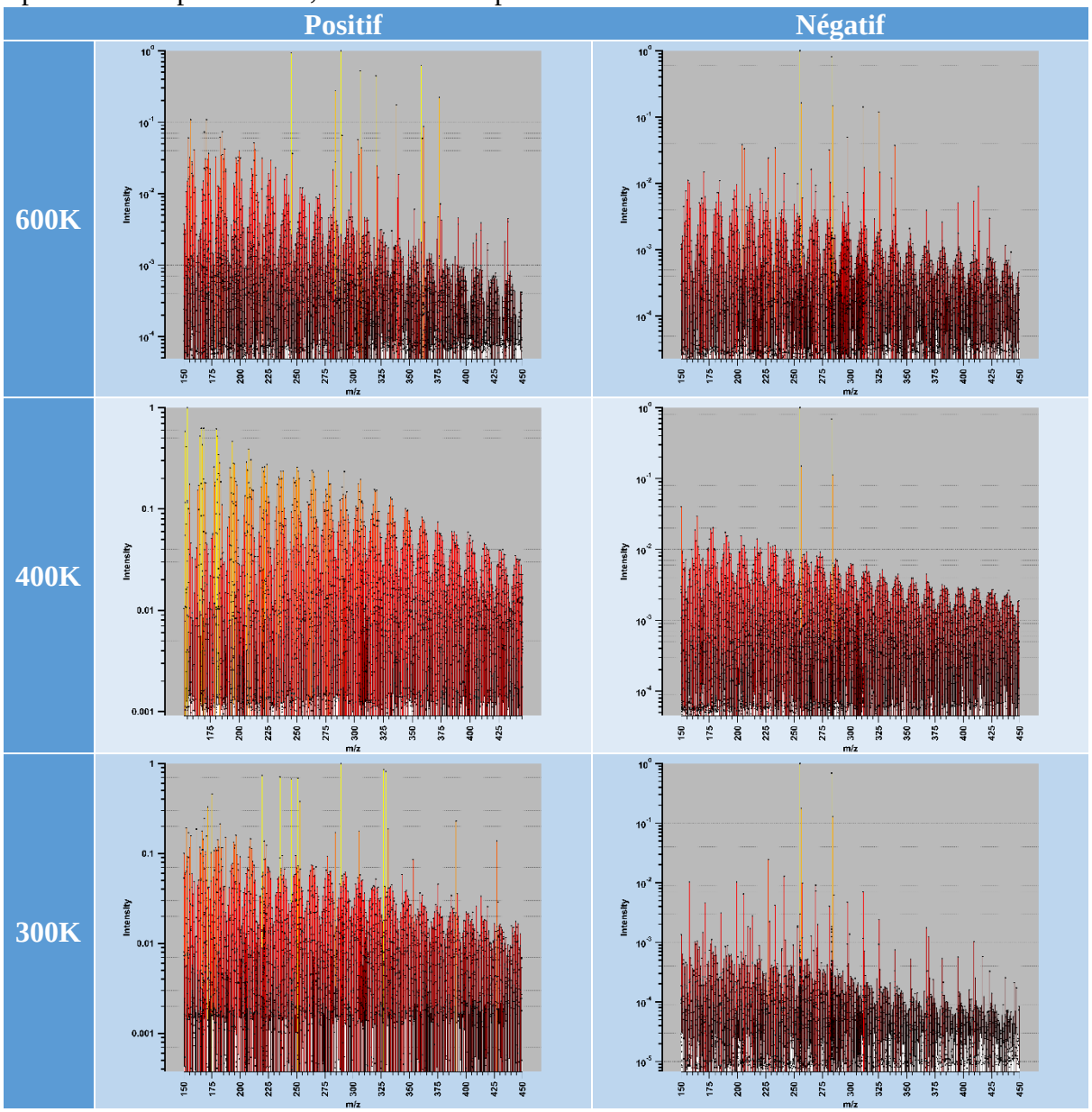


Figure 65 – Spectres de masses des échantillons, sur la gamme de masse [150-450]Da.

Les spectres de masse ne donnent que peu d'information par eux-mêmes, l'information d'intérêt étant à quelques fractions de Daltons alors que le spectre de masse s'étend sur plusieurs centaines de Da. Ainsi, on préfère les représentations de type Défaut de masse en fonction de la masse (DMvM), introduites en partie 1.3.5. Ces représentations DMvM permettent non seulement d'explorer le contenu de l'échantillon, mais permettent également la comparaison entre échantillons. Ces représentations sont disponibles en Figure 66 pour les trois échantillons. Bien que chacun des échantillons présentent des similarités – fuseau dense de molécules carbonées – on observe cependant des variations importantes en ce qui concerne la position du fuseau ainsi que de son étendue, révélant les différences notables de composition chimique *a priori*. On note également que l'échantillon à 400K semble plus intense et de distribution plus

dense que les deux autres échantillons, comme supposé par l'observation du spectre de masse réalisée précédemment.

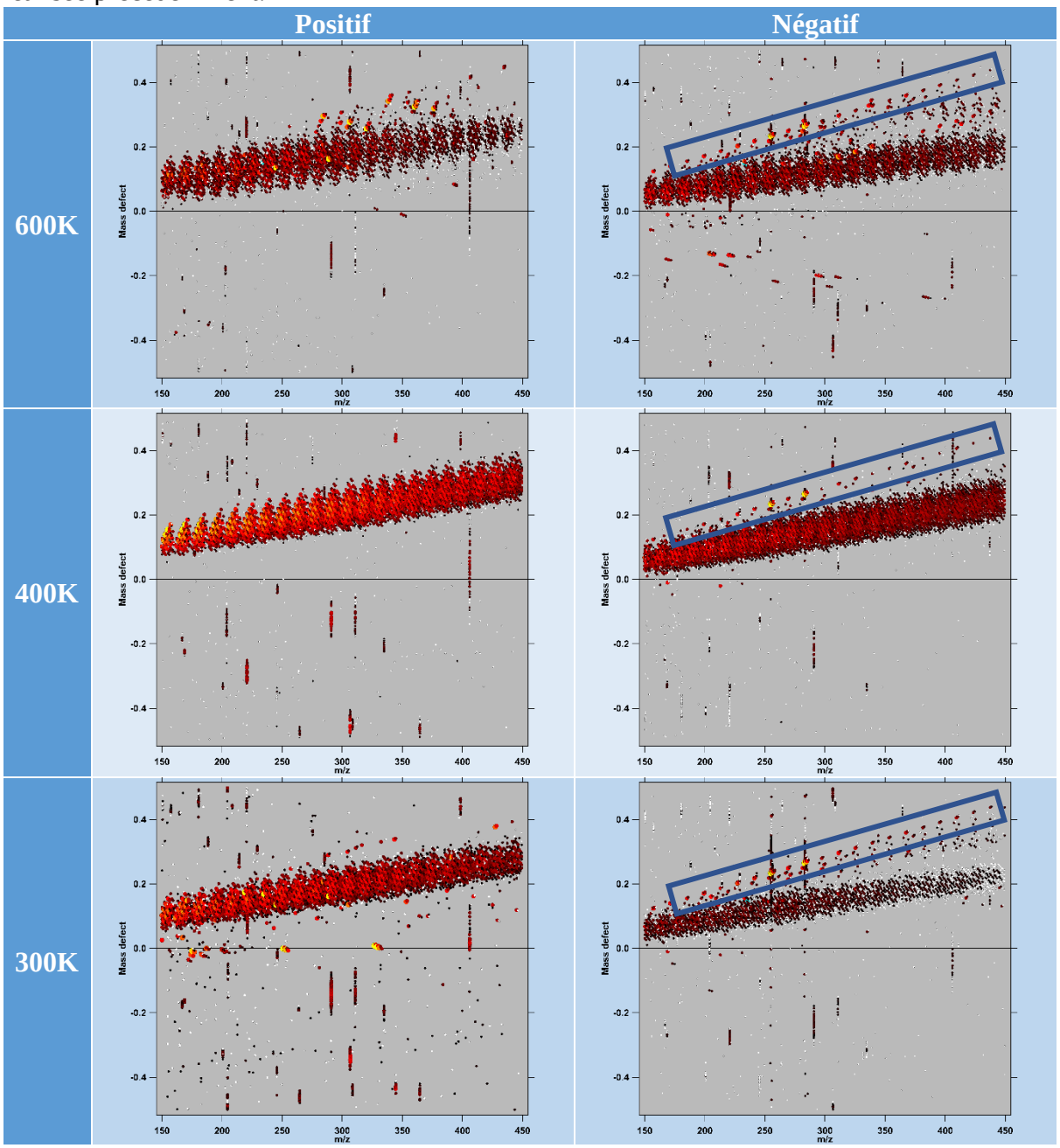


Figure 66 – Représentations graphiques des défauts de masse en fonction de la masse pour chacun des échantillons pour la gamme de masse [150-450]. Le code couleur représente l'intensité des signaux associés (du jaune au noir pour du plus intense au moins intense). En polarité négative, les rectangles bleus représentent des acides gras, et sont de la contamination et non une réponse issue de l'échantillon.

Pour pouvoir comparer plus précisément ces trois échantillons, une attribution complète est nécessaire. La systématique d'attribution présentée en partie 2.1.2 est appliquée à l'attribution des polarités positives et négatives. Ainsi, c'est six spectres qui sont attribués et leurs attributions nettoyées pour obtenir une liste d'attributions sur lesquelles baser un travail de comparaison. L'hypothèse qui permet de nettoyer et valider l'attribution est la nature polymérique de l'échantillon que l'on étudie, et donc qu'une molécule est forcément incluse dans une chaine d'autres molécules liées par une variation unique de ce que l'on peut appeler « monomère ». Cependant, à la différence de la science traditionnelle des polymères, non seulement le monomère ici n'est pas connu et est possiblement non-unique, mais la graine à

l'origine de la polymérisation est également inconnue et non unique. Par chance, le spectre de masse des échantillons révèle des motifs périodiques, que l'on peut lier et représenter. Par exemple, sur les spectres de masses représentées en Figure 67, où quelques familles polymériques sont représentées en violet. Le lien entre chaque point consécutif est une variation unique en CH_2. La courbe en intensité qui est représenté par ces quelques familles polymériques est visible pour l'ensemble des familles en CH_2 que l'on peut déterminer dans l'échantillon, et cette forme particulière est caractéristique des processus de formation de la matière organique par voie synthétique.

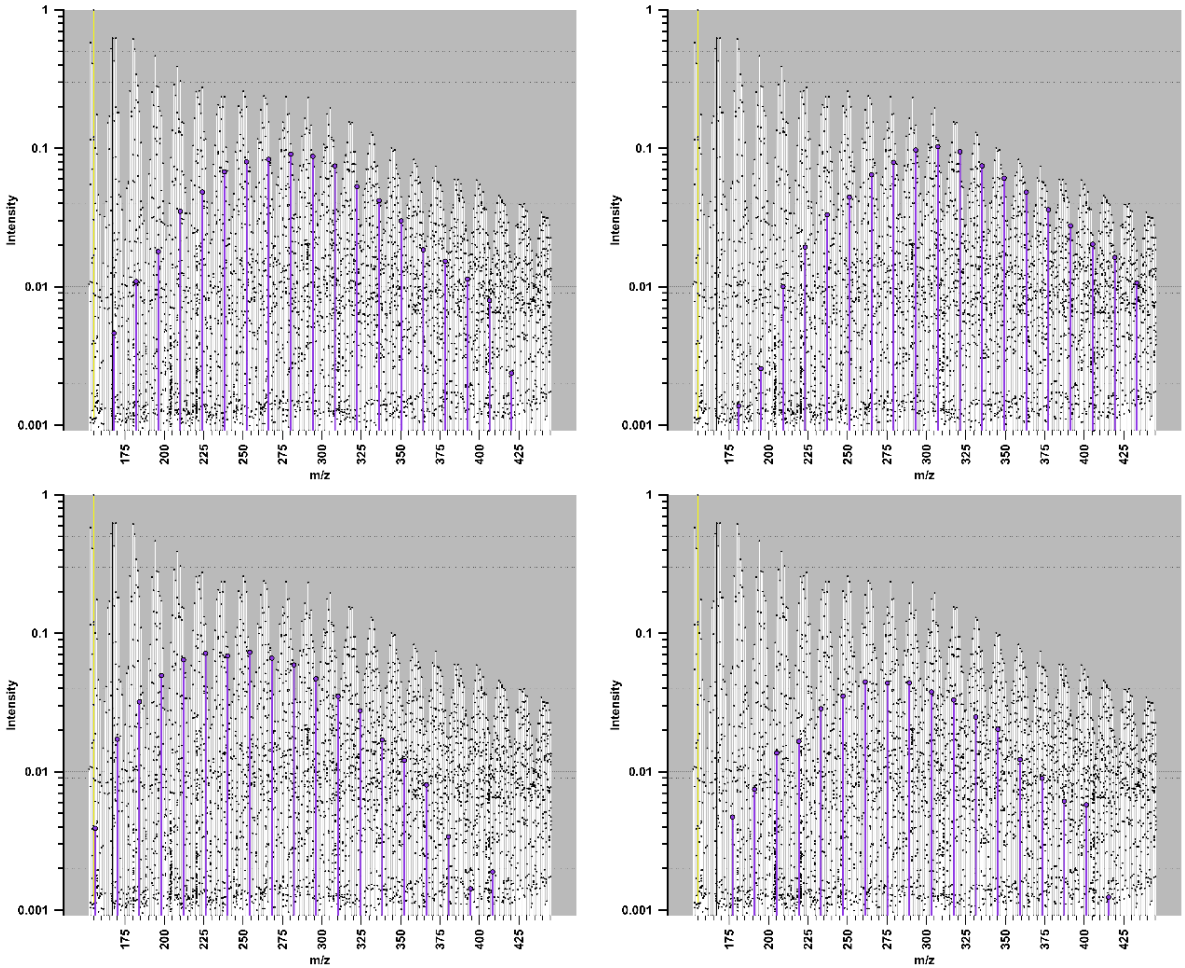


Figure 67 –Représentation en intensité de quelques familles moléculaire pour l'échantillon 400K en polarité positive. La différence entre deux points violets consécutifs est d'un unique groupement CH2.

4.2.1.2. *Attributions et représentation de la diversité*

Toute la complexité du traitement de donnée repose sur comment représenter ces milliers de points de façon visuelle. Une première façon de faire est de comparer l'ensemble des formules attribuées entre elles, et de déterminer la similarité ou non des échantillons entre eux. Pour ce faire, on génère un diagramme de Venn des attributions effectuées, présenté en Figure 68. On remarque tout d'abord que les trois échantillons possèdent une quantité importante de formules stœchiométriques qui leur sont propres. Puis, on note un recouvrement majeur entre les échantillons 300K et 400K (plus de 30% des attributions), alors que l'on observe un faible taux de recouvrement (mois de 10% des attributions) avec l'échantillon à 600K. Enfin, on note qu'environ 10% des attributions sont communes à l'ensemble des trois échantillons. Cette classification à priori simpliste révèle rapidement une similarité importante entre les échantillons 300K et 400K, avec un échantillon 600K différent des deux autres. Un

recouvrement partiel de 10% entre les trois échantillons indique également qu'un ensemble de formules stœchiométriques sont invariantes au mélange initial.

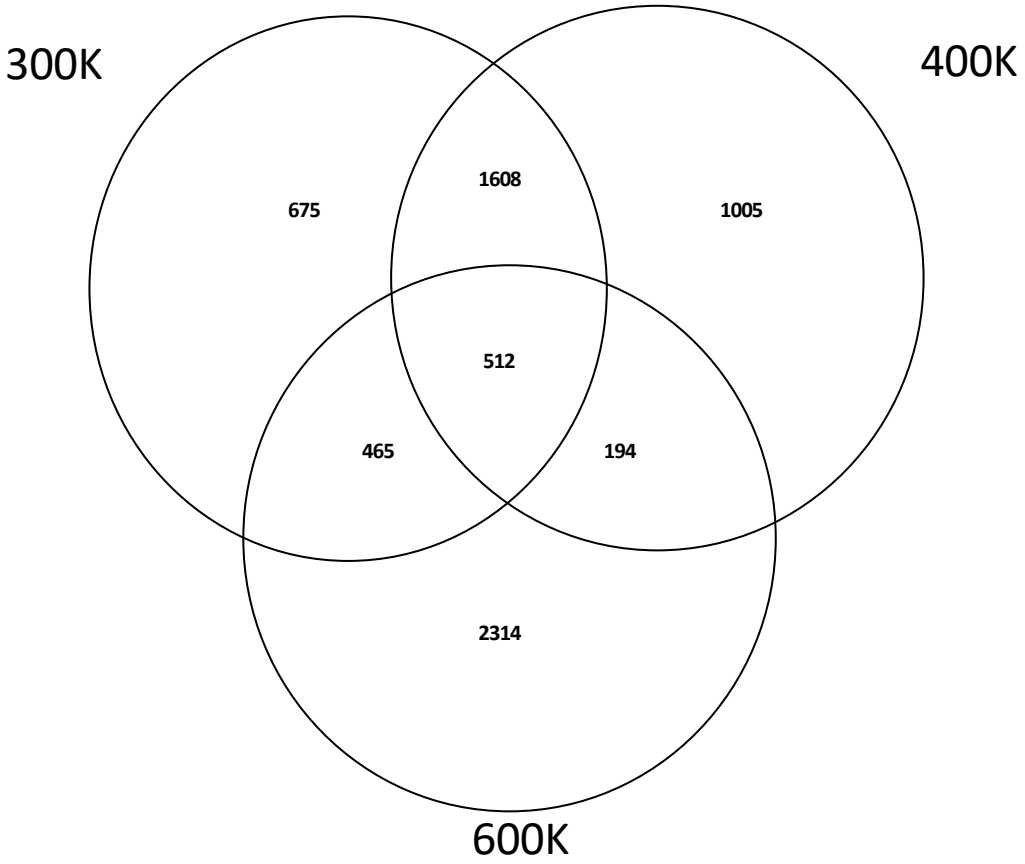


Figure 68 – Diagramme de Venn des attributions pour les échantillons 600K, 400K et 300K. Les tailles de cercles et recouvrement sont volontairement non proportionnels et arbitrairement choisi pour la lisibilité. Exemples de lectures de cette représentation : 675 représente les attributions uniquement présentes dans 300K ; 512 représente les attributions qui sont présentes dans les trois échantillons ; 194 représente les attributions qui sont à la fois dans 400K et dans 600K, mais pas dans 300K.

Du fait de la diversité de composés présents, discuter plusieurs milliers de molécules uniques n'est pas possible : il faut trouver un moyen de simplifier le problème. Considérons les familles moléculaires de ces composés, tels qu'acide aminé, base nucléique ou lipide, et comparons les composés attribués à une base de données. Ainsi, en utilisant la base de données utilisée pour les travaux de prédiction des temps de rétention [45], on peut se focaliser uniquement sur les acides aminés, peptides, base nucléiques et leurs dérivés. L'avantage est que ce sont des molécules azotées et oxygénées, groupes majoritairement présents dans les données à comparer. Comparer les molécules attribuées à la base de données permet de dénombrer les composés qui appartiennent à chaque famille moléculaire. L'histogramme de ce dénombrement, présenté en Figure 69, montre que les échantillons 400K et 300K sont similaires avec l'échantillon à 600K étant enrichi en composés d'intérêts biochimiques. L'étude de l'intersection des domaines (∩) semble également indiquer qu'une majorité des formules stœchiométriques sont identiques entre 300K et 400K, et que 600K semble contenir une majorité des molécules présentes dans les deux autres, comme indiqué par la classification en diagramme de Venn précédemment.

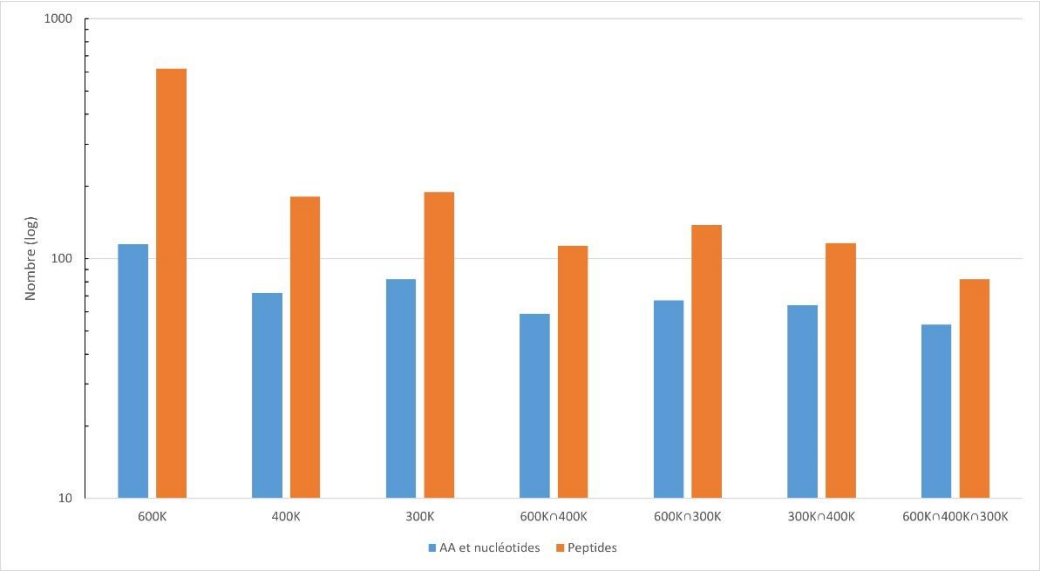


Figure 69 – Dénombrement des formules stœchiométriques correspondant pour quelques familles ciblées dans une base de données (Base de données : 330 Acides Aminés, base nucléiques et dérivés ; 1721 Peptides et dérivés). Les attributions en positif et négatif sont fusionnées et les doublons supprimés.

Par cette comparaison à la base de données, on sait qu'*a priori* une majorité de molécules d'intérêt biochimique sont potentiellement présentes dans les trois échantillons. On peut augmenter le focus et s'intéresser non pas à des familles, mais à quelques molécules, comme par exemple celles listées dans le travail de Sarah Moran qui reprend ainsi les trois échantillons présentés ici. L'intérêt initial de ce type de liste est d'éventuellement définir des objectifs de mesure pour l'analyse en chromatographie. Dans ce cas, l'intérêt est plutôt que la méthode chromatographique disponible et développée au laboratoire soit compatible avec l'analyse de ces échantillons puisque les composés détectés de cette manière sont ceux qui sont compatibles avec la colonne choisie.

Une autre façon de représenter la diversité et de classifier les échantillons est d'utiliser des transformations mathématiques telles que le « *Double Bound Equivalent* » (DBE) ou encore des représentations de Van Krevelen, comme discuté en 1.3.5. Ici, il est préféré des représentations plus simples de ratio d'hétéroatomes par le carbone en fonction de la masse. Ces représentations ont l'intérêt de pouvoir visuellement séparer les différentes familles en hétéroatomes, mais également de pouvoir voir clairement les différences majeures entre échantillons. Ainsi, comme présentés dans la Figure 70 et dans la Figure 71, des différences importantes peuvent être observées dans les distributions en oxygène et en azote entre les différents échantillons étudiés, révélant potentiellement une différence de mécanisme d'implantation des hétéroatomes dans les chaines carbonées. On note également un comportement différent de l'échantillon à 600K comparé aux deux autres échantillons. En effet, la forme globale de la distribution en oxygène et en azote est différente et présente une importante densité de composés avec une composition enrichie en hétéroatomes.

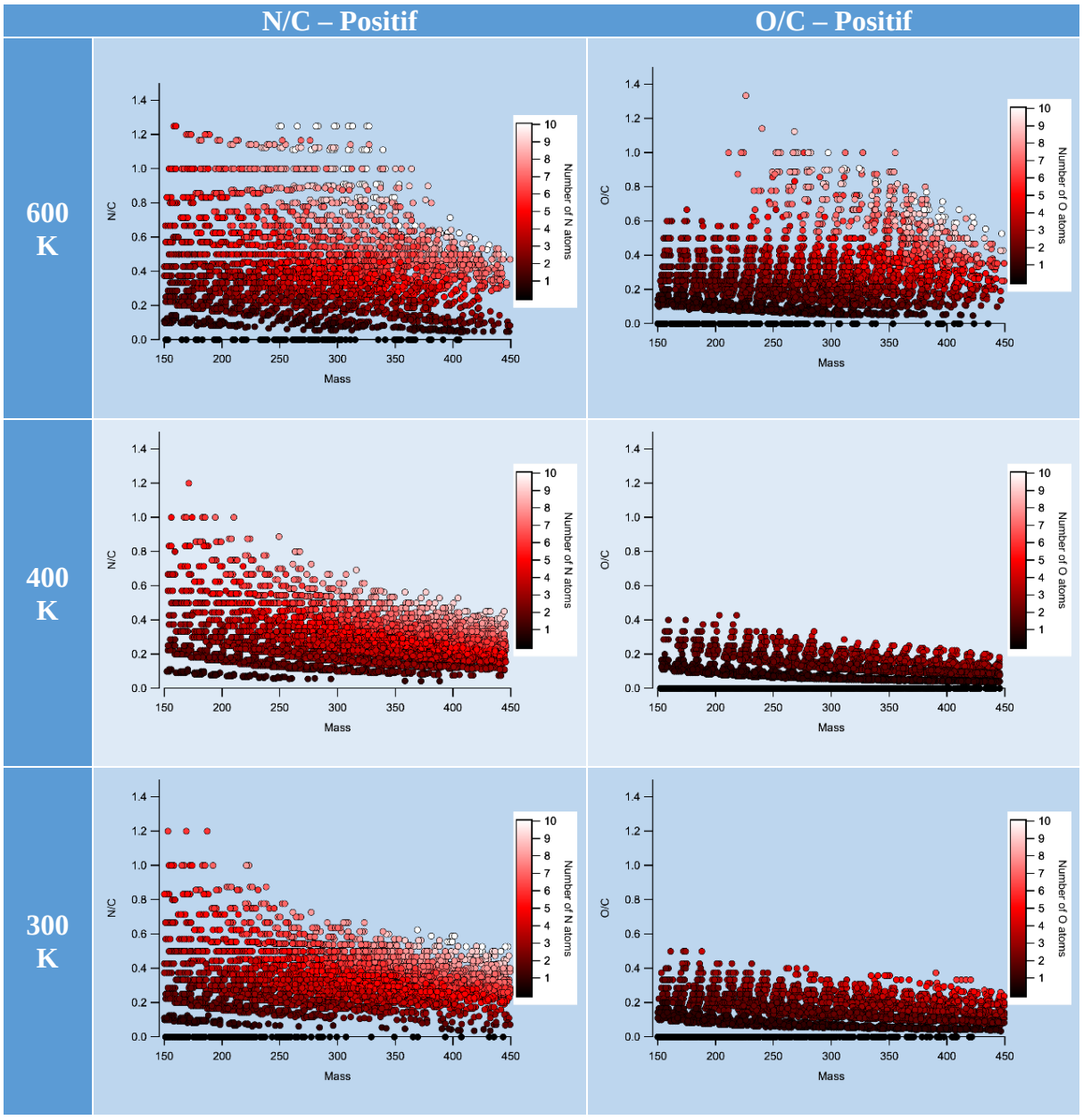


Figure 70 – Représentation des attributions en polarité positive des trois échantillons étudiés.

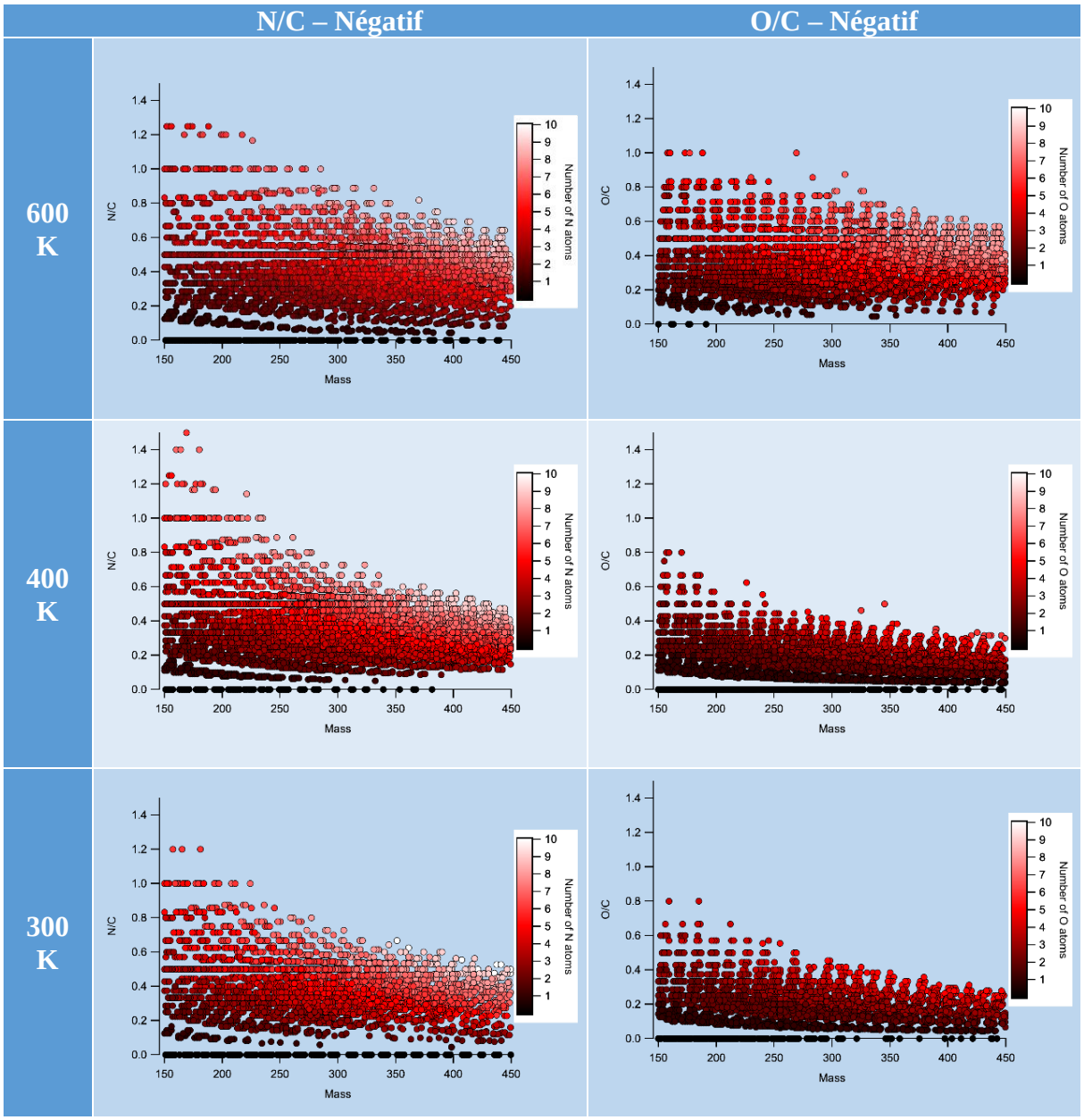


Figure 71 – Représentation des attributions en polarité négative des trois échantillons étudiés.

Cette différence n'est pas due à un problème éventuel d'attribution puisque l'on peut observer également un comportement du fuseau en DMvM qui est différent entre l'échantillon à 600K et les deux autres, comme présenté en Figure 72. En effet, on observe pour l'échantillon à 600K un fuseau situé à des valeurs plus bases de défaut de masse comparé aux valeurs observées pour les échantillons à 300K et à 400K, signifiant soit une insaturation plus importante, soit une présence plus importante d'hétéroatomes possédant un défaut de masse négatif.

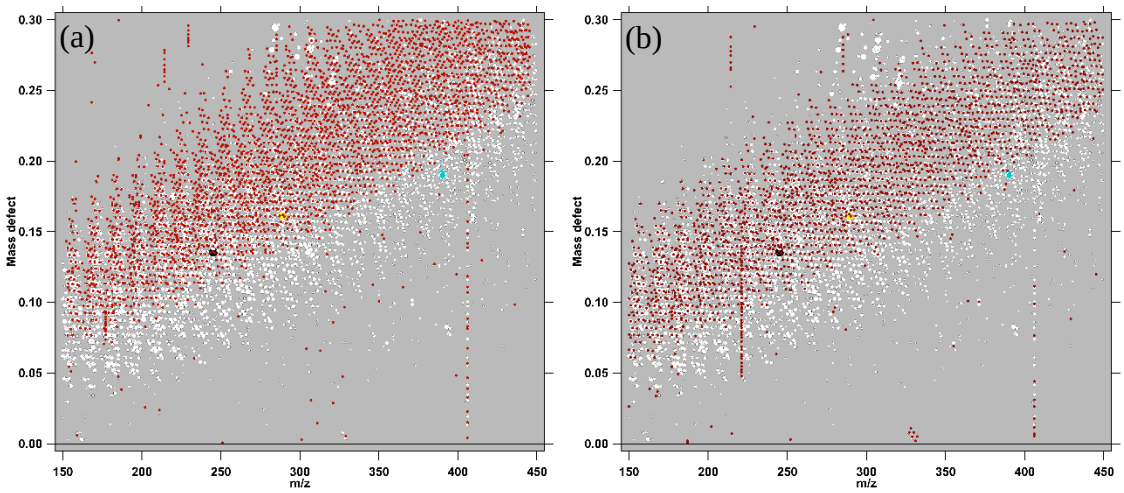


Figure 72 – Représentation en blanc du défaut de masse en fonction de la masse pour l'échantillon à 600K, polarité positive, avec en rouge en (a) l'échantillon 300K et en (b) l'échantillon 400K.

D'autres représentations permettent de visualiser cette différence entre échantillons, comme par exemple la représentation du nombre d'azotes en fonction du nombre d'oxygènes comme présenté en Figure 73. Ici, on ajoute également l'information stœchiométrique d'un Tholins de Titan (95% N_2, 5%CH_4) qui est décrit depuis plusieurs années dans de nombreuses publications [23,53,54] comme étant très azoté et peu oxygéné (i.e. contamination par l'oxygène de l'air). Cette figure nous indique plusieurs informations : (1) l'échantillon à 300K et 400K sont similaires en termes d'incorporation relative de l'azote et de l'oxygène, (2) que les échantillons 300K et 400K sont différents de l'échantillon à 600K par son implantation en oxygène mais pas pour son implantation en azote et (3) que l'implantation en azote des trois échantillons n'est pas comparable à l'implantation observée pour un échantillon réputé très azoté. Cette différence d'implantation d'hétéroatomes peut éventuellement s'expliquer par une différence dans la composition initiale du gaz réactif, comme démontré par la suite.

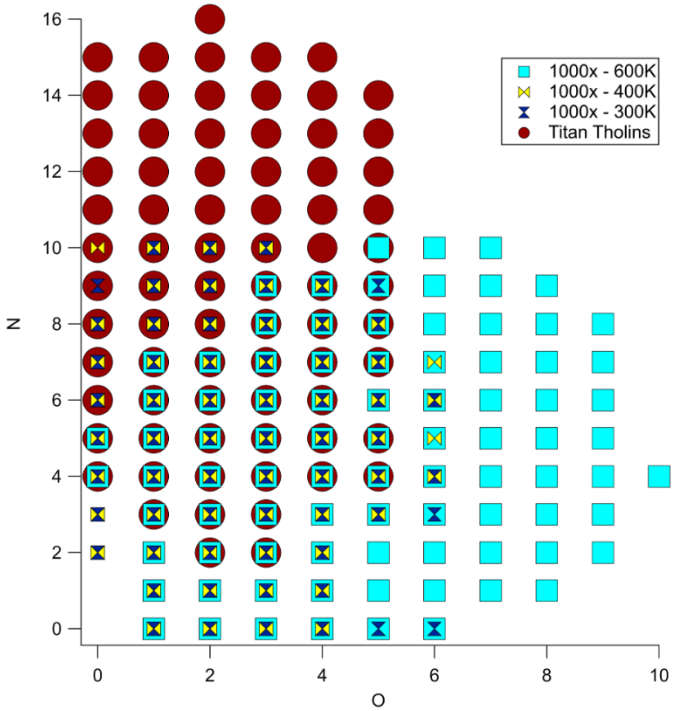


Figure 73 – Représentation du nombre d'azote en fonction du nombre d'oxygène. L'oxygène dans le Tholin de Titan provient d'une oxydation au contact de l'atmosphère.

Ainsi, les trois échantillons étudiés ont une composition initiale du gaz réactif en hydrogène et en azote similaires, mais une quantité de carbone et d'oxygène respectivement croissante et décroissante avec l'augmentation de la température, comme présenté en Tableau 16. L'échantillon à 600K, qui possède une incorporation importante d'oxygène, présente initialement une composition initiale en carbone plus importante pour une composition initiale en oxygène plus faible comparé aux deux autres échantillons étudiés.

Échantillon	Composition du mélange	Composition élémentaire
1000x – 300K	66%H_2O, 6.6%CH_4, 6.4%N_2, 4.9%CO_2, 16%He	8.2%C, 9.4%H, 10.6%N, 71.8%O
1000x – 400K	56%H_2O, 11%CH_4, 10%CO_2, 6.4%N_2, 1.9%H_2, 14.7%He	13.9%C, 8.8%H, 9.9%N, 67.3%O
1000x – 600K	42%H_2, 20%CO_2, 16%H_2O, 5.1%N_2, 1.9%CO, 1.7%CH_4, 13.3%He	19.2%C, 8.3%H, 9.7%N, 62.8%O
Titan	95%N_2, 5%CH_4	2.2%C, 0.7%H ; 97.1%N, 0%O

Tableau 16 – Composition élémentaire du mélange réactif ayant servi à la synthèse des échantillons discutés

Le niveau stable d'azote dans la composition initiale peut également expliquer l'incorporation similaire d'azote dans les trois échantillons. Néanmoins, on souhaite vérifier comment l'incorporation est effectuée, et cela est fait en étudiant le ratio *(H-N)/C* en fonction du DBE présenté en Figure 74, représentation qui permet grossièrement de déterminer par quel atome est majoritairement porté l'insaturation.

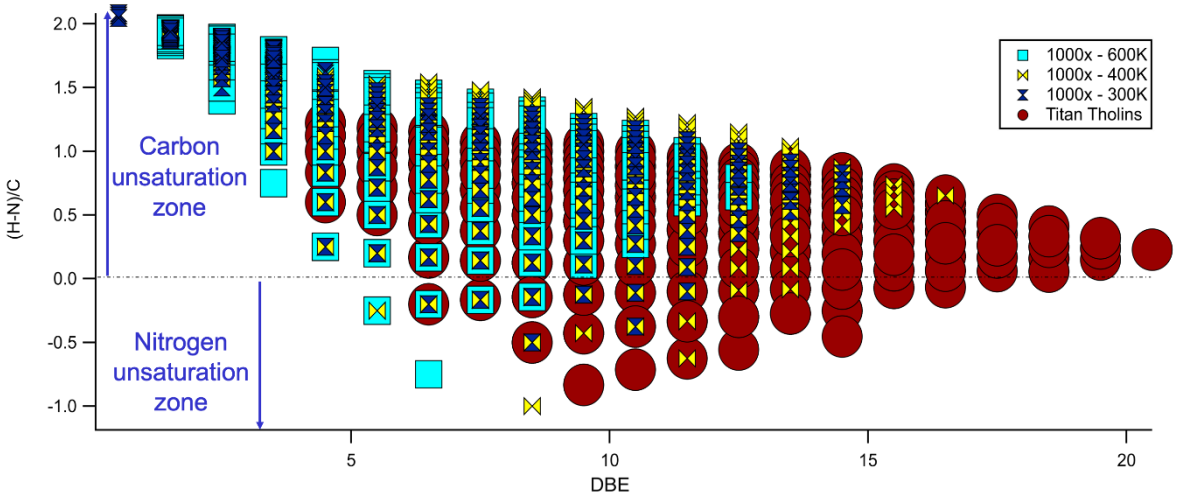


Figure 74 – Comparaison des DBE en fonction de l'insaturation portée par le carbone ou l'azote.

Ainsi, comparé à la Figure 73 qui indique que les échantillons 300K et 400K sont similaires, on a ici une indication concernant la différence en insaturation entre les deux échantillons. Ces deux échantillons sont également différents de l'échantillon à 600K qui ne présente que peu d'insaturation due à l'azote avec également globalement un DBE plus faible que les deux autres échantillons. Ces variations sont probablement également explicables par la variation de la composition élémentaire du gaz réactif utilisé. Cependant, cette variation de la composition initiale est effectuée sur plusieurs paramètres en même temps, il est alors impossible de faire ressortir la source à l'origine de la variation observée sur les formules stœchiométrique et l'incorporation différenciée des hétéroatomes.

4.2.2. Impact de la résolution – analyses Orbitrap et ICR

L'Orbitrap n'ayant qu'une résolution et une précision limitée, on peut se demander si les attributions effectuées sont fiables. Pour ce faire, on utilise un FT-ICR 12T, installé en 2016,

qui présente des résolutions supérieures au million à la masse 150 Da alors que l'Orbitrap est limité à quelques centaines de milliers. L'ICR possède de nombreuses améliorations, principalement en ce qui concerne le design de la cellule. L'Orbitrap utilisé a été conçu en 2007 et ne bénéficie d'aucune des améliorations apportées depuis (cellule améliorée, source HESI). Cette différence technologique initiale est à prendre en compte *a priori* pour la comparaison des résultats.

Par la suite, seul l'échantillon 400K est analysé. Ce choix a été fait étant donné l'intensité observée en spectrométrie de masse et la quantité de matière disponible avec cet échantillon en particulier, quantité nécessaire aux différentes analyses effectuées par la suite, depuis leurs essais jusqu'aux données finales. Les analyses en ESI-ICR ont été acquises sur la gamme de masse [100-800]Da, et sont ici réduites à [150-450] Da pour les besoins de la comparaison avec les données Orbitrap.

Comme présenté en Tableau 17, les analyses ICR sont plus denses que les analyses Orbitrap, avec une différence en nombre de signaux de plus de 250%. Cela se traduit également par un nombre de formules attribuées plus importantes (+200%) en ICR qu'en Orbitrap (^{13}C exclus des attributions).

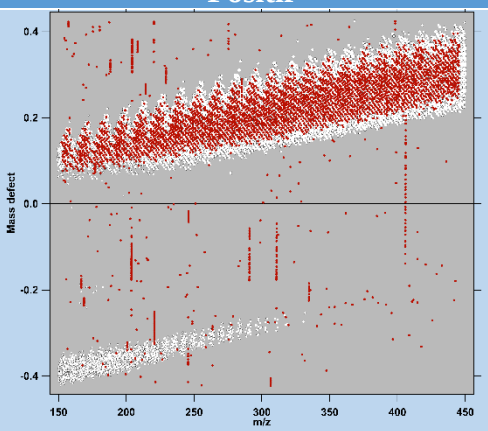
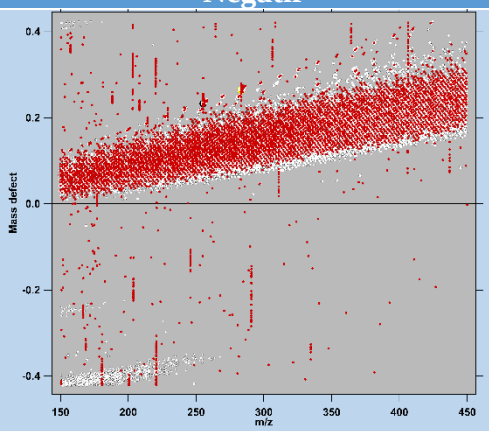
	Positif	Négatif
400K	 Nombre de points : • ICR : 14850 • Orbitrap : 4229 Nombre d'attributions (sans ^{13}C) : • ICR : 5601 • Orbitrap : 2503	 Nombre de points : • ICR : 12116 • Orbitrap : 5247 Nombre d'attributions (sans ^{13}C) : • ICR : 6221 • Orbitrap : 3523

Tableau 17 – Défaut de masse en fonction de la masse des analyses ICR (blanc), avec les analyses Orbitrap équivalentes en superposition (rouge).

Cette différence est expliquée par la gamme dynamique effective de l'ICR qui est plus importante que celle de l'Orbitrap utilisé, comme discuté en partie 2.1.3. On présente également en Figure 75 la superposition des attributions en Orbitrap et en ICR par-dessus le spectre de masse ICR. Les attributions ICR sont représentées directement sur le spectre, sans autre traitement. Les données Orbitrap doivent cependant être ajustées : on part de la masse exacte de chaque attribution, et on cherche dans le spectre ICR quelles masses correspondent. L'intensité Orbitrap est alors remplacée par l'intensité ICR, et l'erreur recalculée. C'est ainsi que les attributions Orbitrap, et plus particulièrement leur profil en intensité, sont recalculées pour le spectre ICR, et peuvent alors être comparées avec les attributions ICR directement. Cette figure confirme que la gamme dynamique effective de l'Orbitrap est inférieure à la gamme de l'ICR du fait de l'absence de molécules pour des intensités inférieures à 2.10^8 en Orbitrap alors que cette gamme d'intensité est riche en informations en ICR. Ce n'est pas la seule source de la différence entre Orbitrap et ICR, puisque si l'on ne considère que les signaux

supérieurs à 3.10^8, l'attribution ICR présente encore deux fois plus d'attributions que l'attribution de l'Orbitrap, avec respectivement 4697 et 2249 attributions.

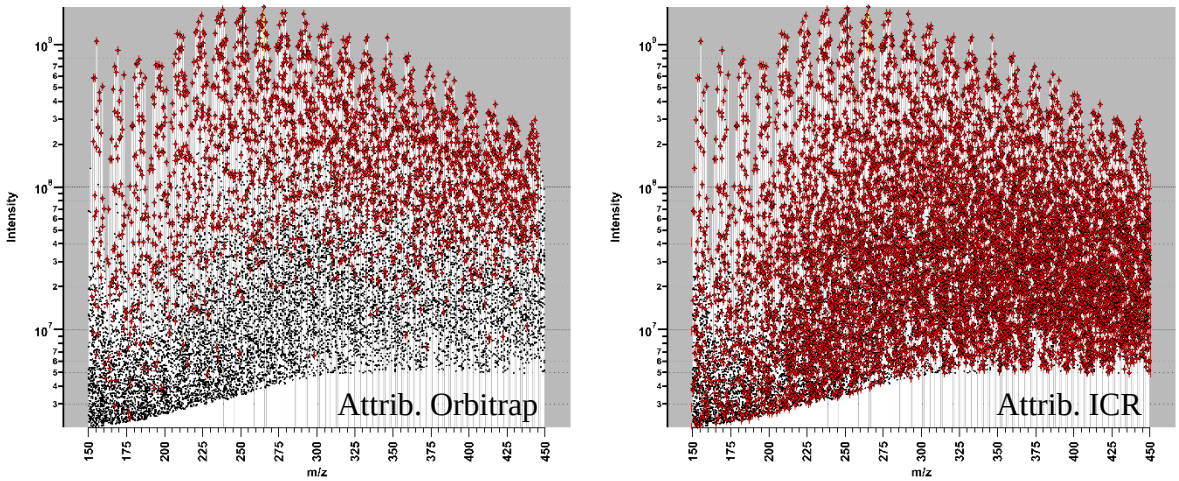


Figure 75 – Superposition des attributions Orbitrap et ICR, polarité positive seulement, par-dessus le spectre ICR. Les attributions Orbitrap sont recalculées (intensité, erreur) en se basant sur le spectre ICR pour obtenir ces résultats.

Malgré cette différence de gamme dynamique effective et de résolution entre Orbitrap et ICR, la représentativité globale de l'analyse en termes d'éléments est globalement conservée par les analyses en ESI. Ces données sont également comparables à l'analyse élémentaire réalisée par IRMS, comme présenté en Figure 76. La différence systématique observée avec les analyses de la phase soluble (ICR et Orbitrap) comparé à l'analyse globale (IRMS) est expliquée par : (1) l'analyse uniquement de la partie soluble par HRMS et (2) les analyses ESI sont biaisées vers les molécules acido-basiques. Les analyses Orbitrap, comparées aux analyses ICR sont effectivement moins sensibles, mais les données ne présentent pas de biais systématique de nature à changer les conclusions si l'on utilise un Orbitrap ou un ICR. Ainsi, les conclusions effectuées avec des données Orbitrap sont toujours valides.

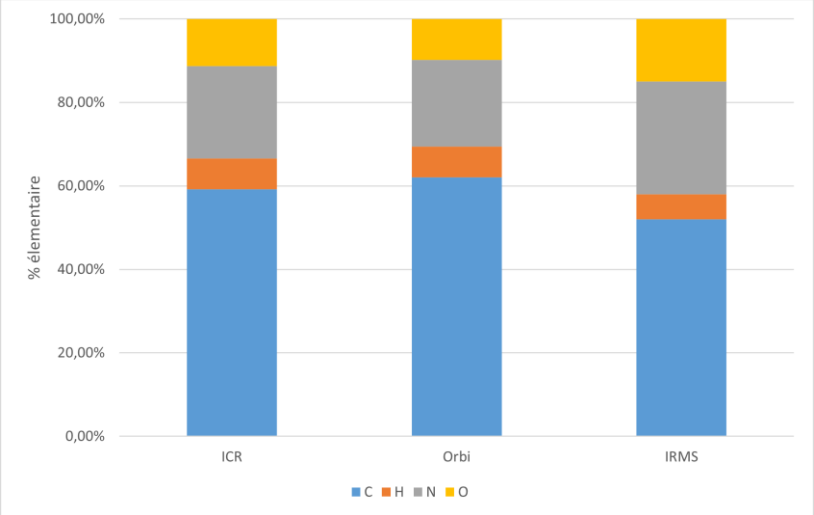


Figure 76 – Comparaison des analyses élémentaire en Orbitrap, ICR et IRMS, pour l'échantillon 400K.

Ainsi, en Figure 77 et Figure 78 qui représentent les attributions entre l'ICR et l'Orbitrap, on peut discerner une différence de densité de molécules détectées. Cette différence est possiblement à mettre en relief avec le fait que les acquisitions ont été effectuées avant les travaux d'optimisation présentés en 2.1.1, mais n'explique néanmoins pas totalement cette différence puisque passer de 4x128 à 1x1024 micro-scans n'ajoute qu'environ 40% de formules stœchiométriques supplémentaires selon les données disponibles. La différence est alors sur la puissance résolutive et la sensibilité de l'ICR comparé à l'Orbitrap.

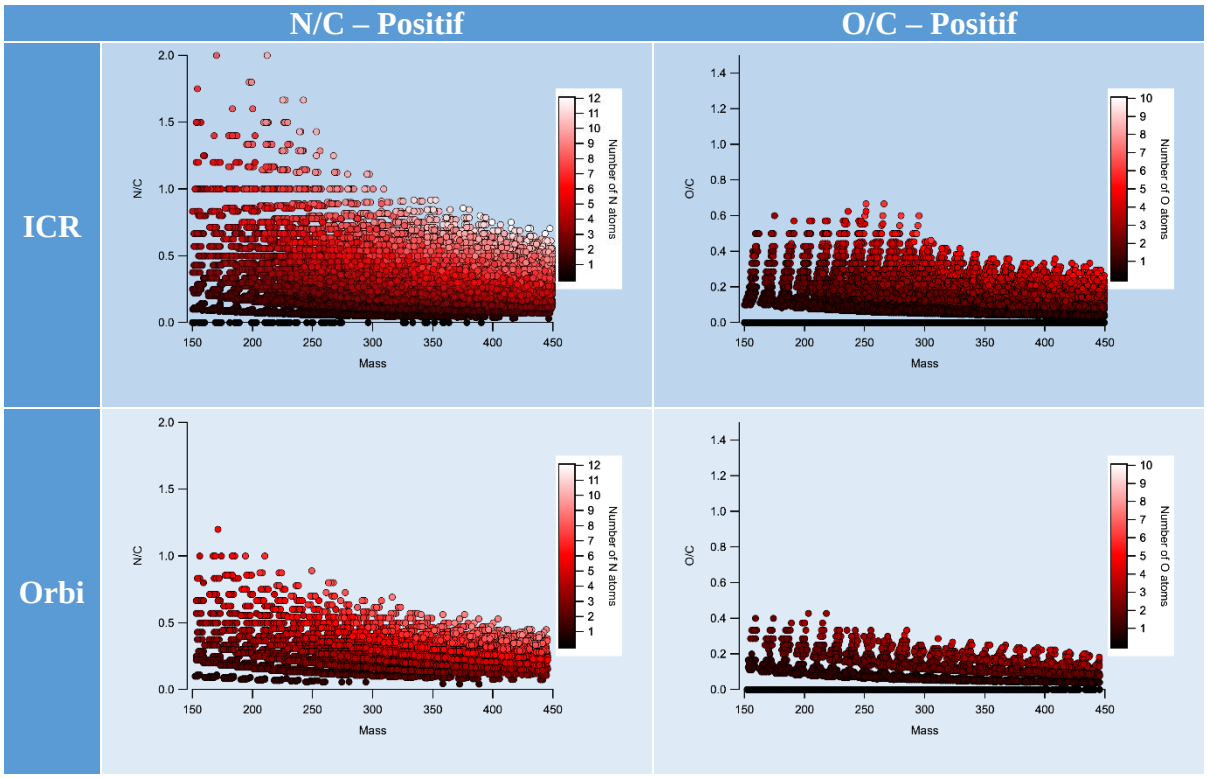


Figure 77 – Représentation des attributions en polarité positive des attributions ICR et Orbitrap pour la même gamme de masse.

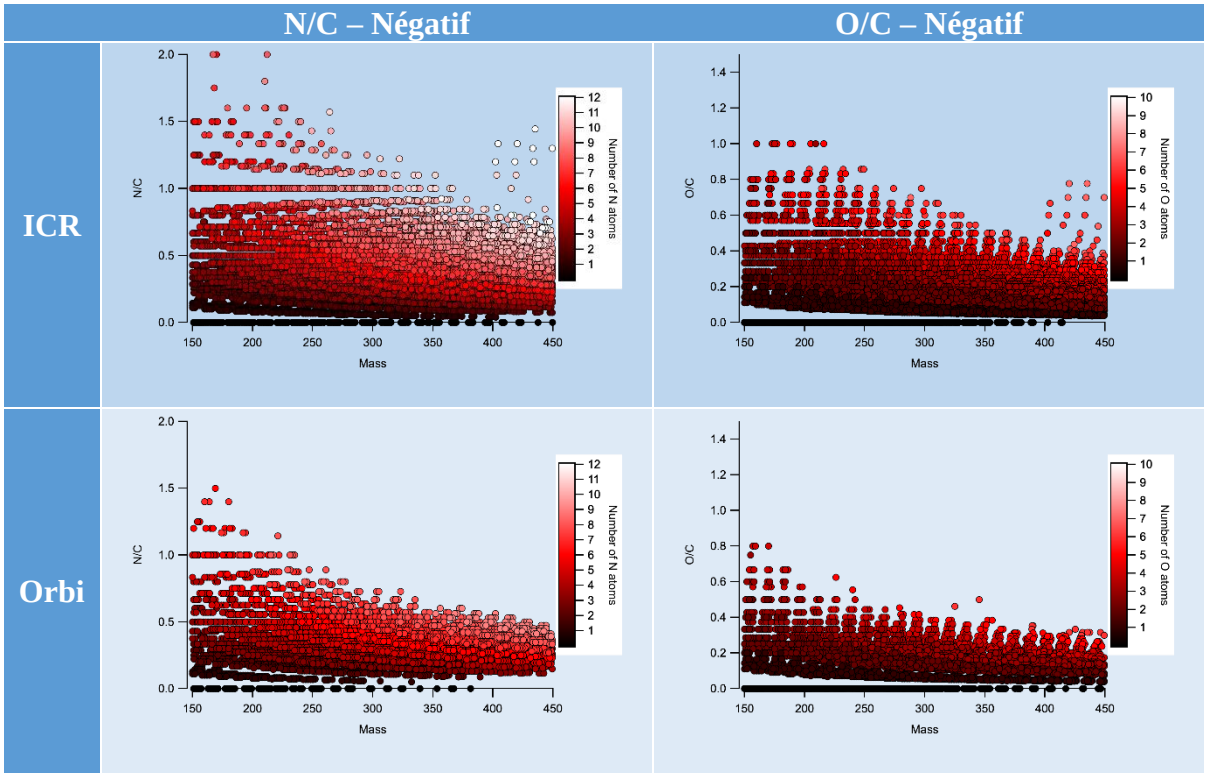


Figure 78 – Représentation des attributions en polarité négative des attributions ICR et Orbitrap pour la même gamme de masse.

On peut ainsi raisonnablement penser que la différence observée est relative à la différence d'instrumentation. En effet, les données Orbitrap acquises sur la gamme [400-1000] Da ne peuvent pas être utilisées du fait de la perte de résolution et de précision de l'Orbitrap. La puissance de l'ICR en termes de pouvoir résolutif est plutôt pour analyser les gammes de masses

supérieures à 450 Da puisque c'est à partir de là que la différence de résolution commence à avoir de l'importance pour l'attribution des formules stœchiométriques. Dès lors, sur la gamme de masse considérée, à savoir [150-450] Da, la puissance résolutive de l'ICR n'apporte aucun avantage comparé à l'Orbitrap. Ainsi, même si les analyses Orbitrap sont moins sensibles que celles de l'ICR, elles ne sont néanmoins pas dénuées de sens et représentent tout de même correctement la diversité observée puisque plus de 95% des attributions Orbitrap sont incluses dans les attributions ICR.

4.2.3. Phase totale, insoluble et soluble : analyses LDI

Les analyses LDI ont été acquises sur une gamme de masse unique : [98-1000] Da en utilisant 500 scans et un spectre de 8 millions de points. Chaque spectre est acquis en polarité positive et négative et les données sont traitées selon la systématique présentée en partie 2.1.3. L'ensemble des paramètres utilisés pour les analyses LDI sont disponibles en 0.

4.2.3.1. *Exploration des échantillons*

Il est également intéressant de regarder la différence entre analyses de la phase soluble, insoluble et de la phase totale. Une telle comparaison nécessite une source qui est capable d'ioniser de multiples phases, et la source LDI permet par exemple d'effectuer ces analyses. Chaque spectre LDI présenté par la suite est constitué de 500 spectres acquis à la suite, chaque spectre étant effectué sur une zone physique de l'échantillon différente des spectres précédents. L'analyse spectre par spectre lors de l'acquisition ne présentant pas de variations majeures, les échantillons sont supposés homogènes. De ce fait, la moyenne spatiale n'aura pas d'influence sur l'interprétation des résultats, mais cela ne serait par exemple pas possible si l'on effectuait l'analyse de la phase totale d'une météorite qui présente un mélange inhomogène de matrice, chondrules et autres grains [55].

Les spectres, acquis sur la gamme de masse [100-900]Da, sont également très denses et il faut trouver des représentations qui permettent de visualiser les différences entre les échantillons. Pour donner un ordre de grandeur de la taille des données, chaque acquisition a généré des spectres de plus de 15 000 points et permet l'attribution de 12 000 à plus de 20 000 formules stœchiométriques pour certaines analyses. Les analyses LDI-ICR sont comparées à l'analyse ESI-ICR qui est ici prise comme référence.

On peut remarquer une différence notable entre l'analyse de la fraction soluble et l'analyse des fractions insolubles et totales dans la Figure 79, du fait de l'élargissement significatif du fuseau sur les défauts de masses plus faibles pour les analyses de la fraction insoluble et totale. Cet élargissement est a priori révélateur de la présence d'une diversité en molécules plus insaturées que celles détectées pour la fraction soluble. On note également l'apparition d'un deuxième fuseau de molécules dans l'analyse LDI de la fraction soluble comparé à l'analyse ESI. Ce second fuseau est également présent dans les analyses des autres fractions en LDI négatif. Comme la source LDI créé un plasma à la surface pour ioniser les molécules, il est possible que ce second fuseau soit créé par des réactions chimiques non-souhaitées, dénaturant alors les molécules analysées. Les échos à défaut de masse négatif sont également anormaux pour des molécules composées uniquement de CHNO qui ne devraient pas pouvoir présenter des signaux à défaut de masse aussi négatifs. Enfin, les signaux à défaut de masse nuls après la masse 500 Da en positif sont des fullerènes produits par la combustion des molécules carbonées, et servent de molécules calibrantes en LDI positif.

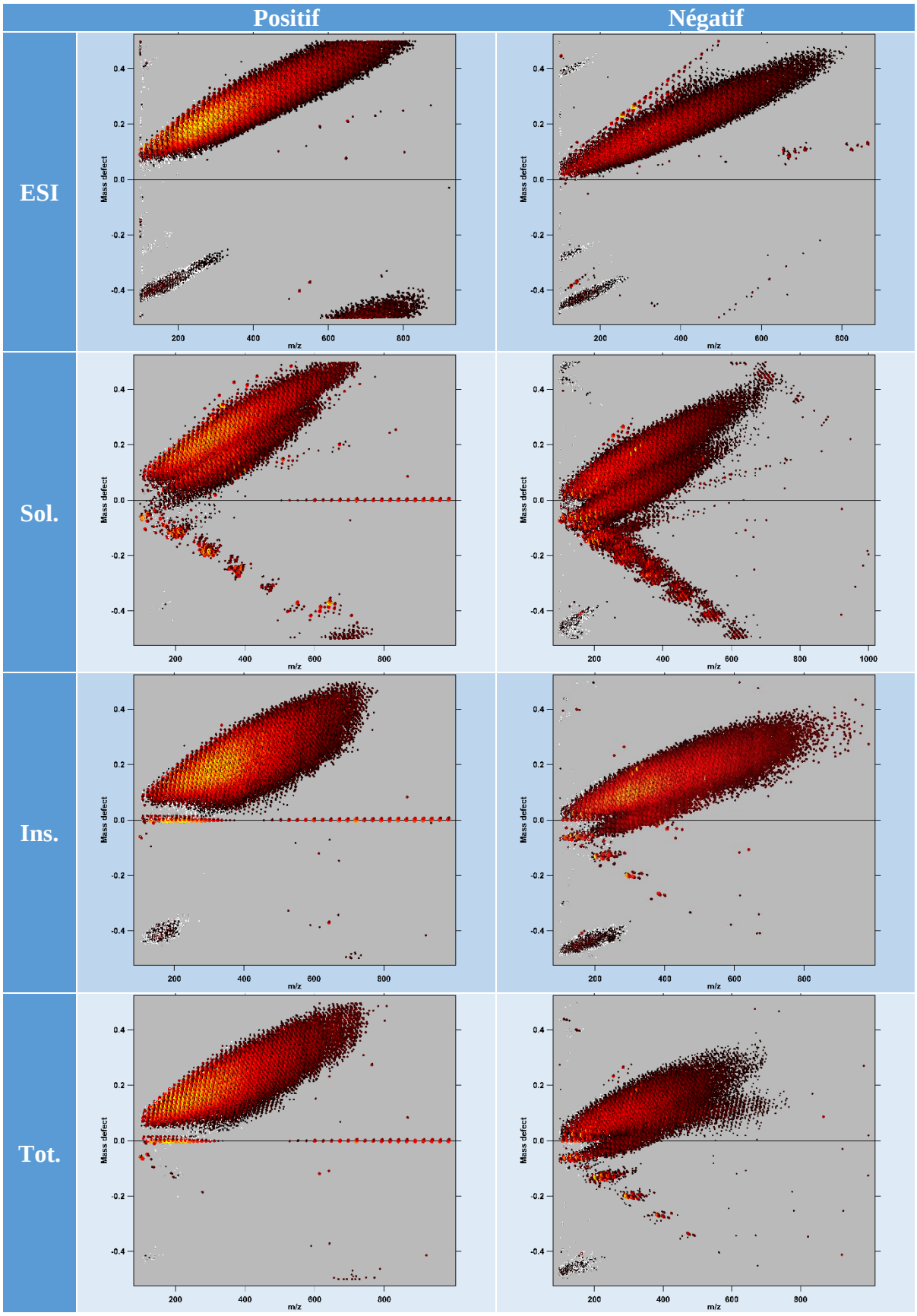


Figure 79 – Représentation des défauts de masse en fonction de la masse pour les différentes fractions analysées.

De façon générale, on s'attend également à ce que l'analyse de la fraction totale soit la somme de l'analyse de la fraction soluble et de la fraction insoluble. On présente en Figure 80

un zoom sur un massif à faible masse et un massif à une masse plus élevée. On peut alors remarquer qu'à faible masse, les trois fractions sont comparables et qu'à plus haute masse, la fraction soluble (en bleu) est différente des deux autres. La comparaison des attributions faite plus tard permettra de déterminer le taux de recouvrement entre les différentes fractions. Néanmoins, du fait de la différence visuelle des DMvM et le fait d'une densité moindre de signal après 625 Da pour la phase totale, cela engendrera par défaut des différences importantes. Il faudra alors peut être comparer les attributions sur une gamme de masse restreintes pour éviter ce biais et pouvoir ainsi conclure sur les données.

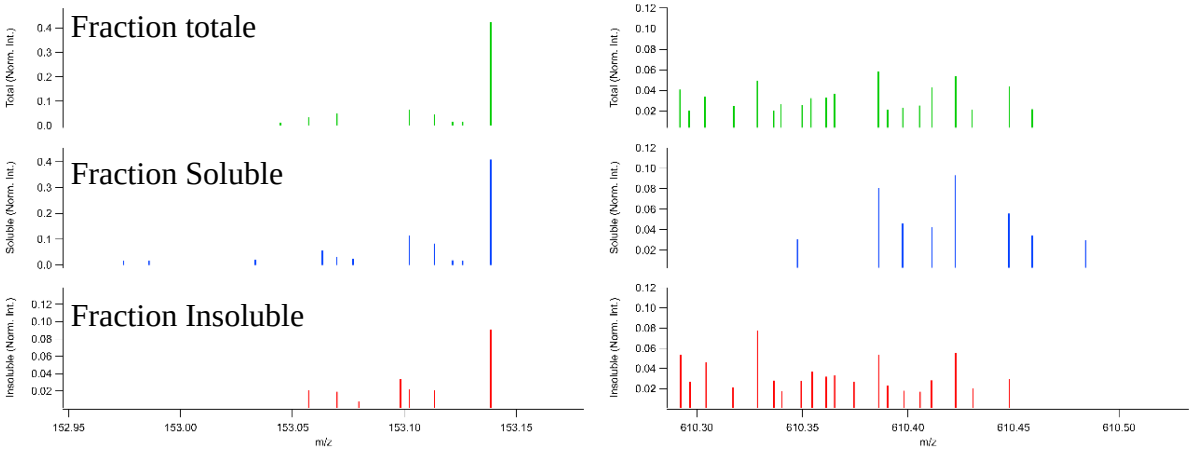


Figure 80 – Comparaison des trois fractions sur un massif à faible masse (153Da) et un massif à masse plus élevée (610Da).

Maillard et al [17] ont également effectué ce type de travail avec des Tholins de Titan. On note le même motif de répétition entre notre échantillon et le leur en comparant la fraction soluble (HCN) et la fraction insoluble (C_2H_2), comme montré en Figure 81. Cette variation de motif principal explique le décalage observé en intensité lorsque, par exemple autour de la masse 260Da, la fraction soluble présente un maxima d'intensité alors que la fraction insoluble présente un minima d'intensité. Il existe également de multiples autres motifs de polymérisations possibles[40], mais ceux mentionnés ici sont ceux qui expliquent le mieux les variations d'intensité observées.

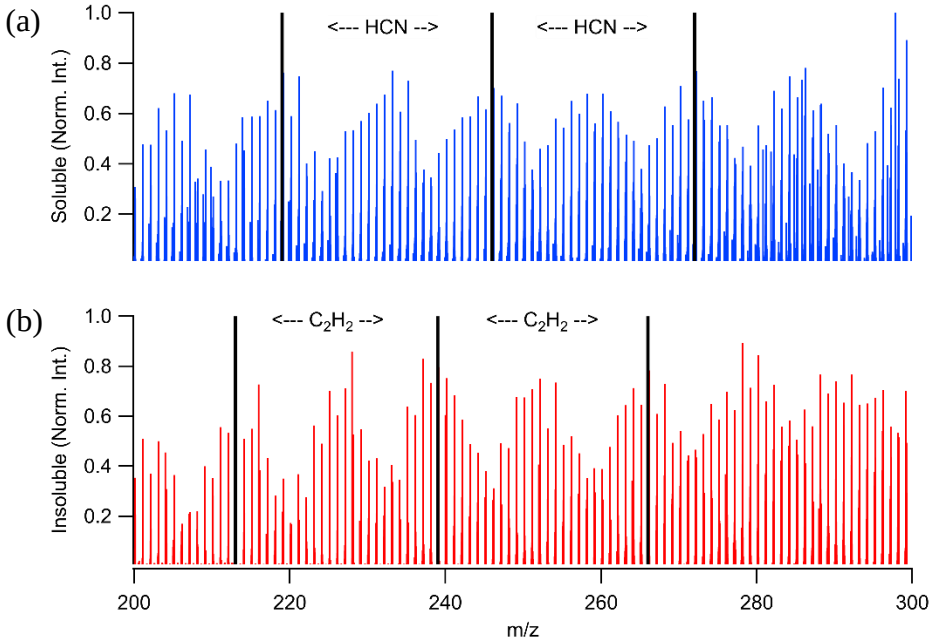


Figure 81 – Zoom sur le spectre de masse de (a) la fraction soluble et (b) la fraction insoluble pour mettre en évidence le motif périodique principal des deux échantillons.

4.2.4. Attribution des données LDI

Les attributions de la phase totale doivent être comparées aux attributions de la phase soluble et insoluble pour déterminer l'inclusion de la phase soluble et insoluble dans la phase totale. Comme les données présentent des différences importantes de sensibilité pour la phase totale, cette comparaison n'est effectuée que sur la gamme de masse restreinte [120-620]Da. Le résultat de cette comparaison est représenté sous forme de diagramme de Venn en Figure 82. Ce diagramme indique que la majorité des molécules détectées dans la phase totale sont retrouvées soient dans les deux autres, soit seulement dans la phase insoluble. On note également un nombre important de molécules uniques à la phase soluble et à la phase insoluble, en accord avec les conclusions de Maillard et al [17].

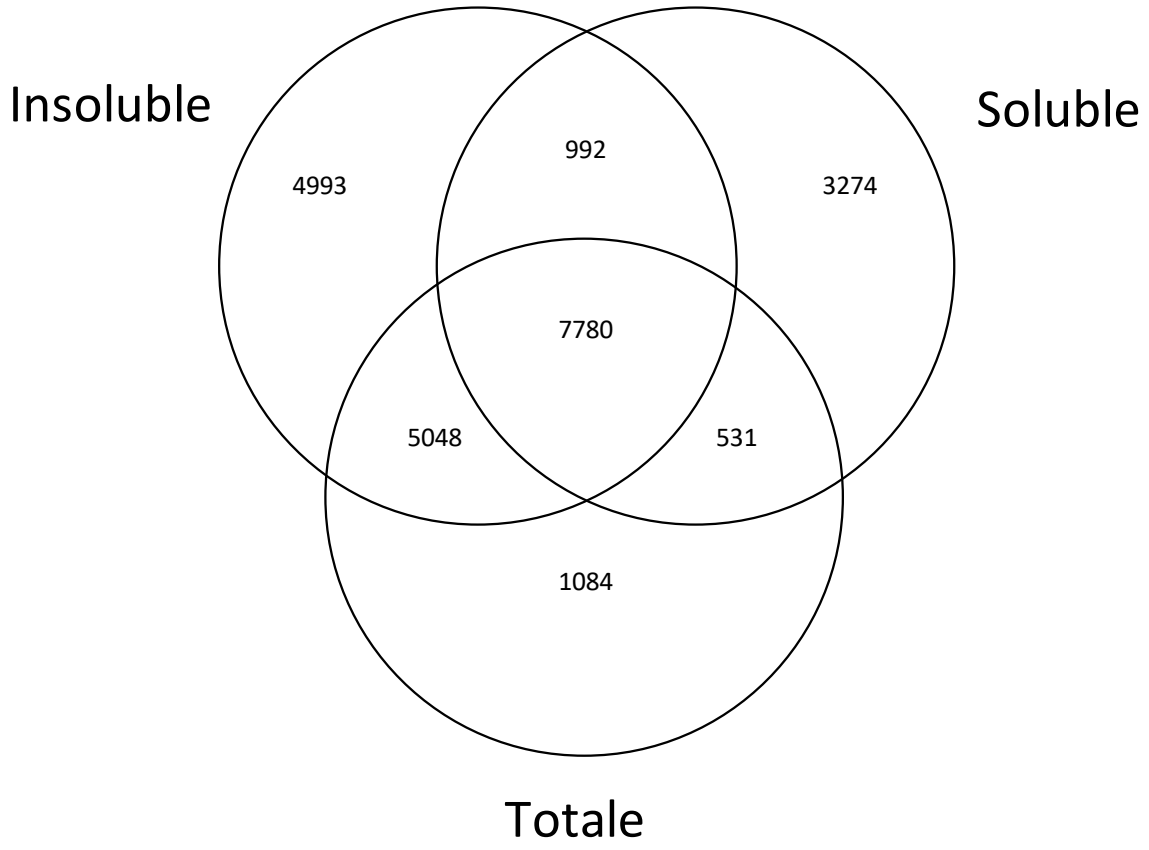


Figure 82 – Diagramme de Venn des attributions de la phase soluble, insoluble et totale. Les formules communes entre chaque échantillon sont présentes dans l'intersection des espaces.

Parmi ces différentes classes, il est intéressant de déterminer ses caractéristiques globales, et l'une d'elle est sa diversité en termes de DBE, présenté en Figure 83. On note ainsi une différence importante entre la fraction soluble et les deux autres, avec une valeur de DBE globalement plus faible pour la fraction soluble. Cela est attendu puisque le DBE mesure le degré d'insaturation et de cycles des molécules. Or, plus une molécule est insaturée, moins elle sera soluble dans le méthanol, expliquant alors la différence entre la phase soluble et les deux autres.

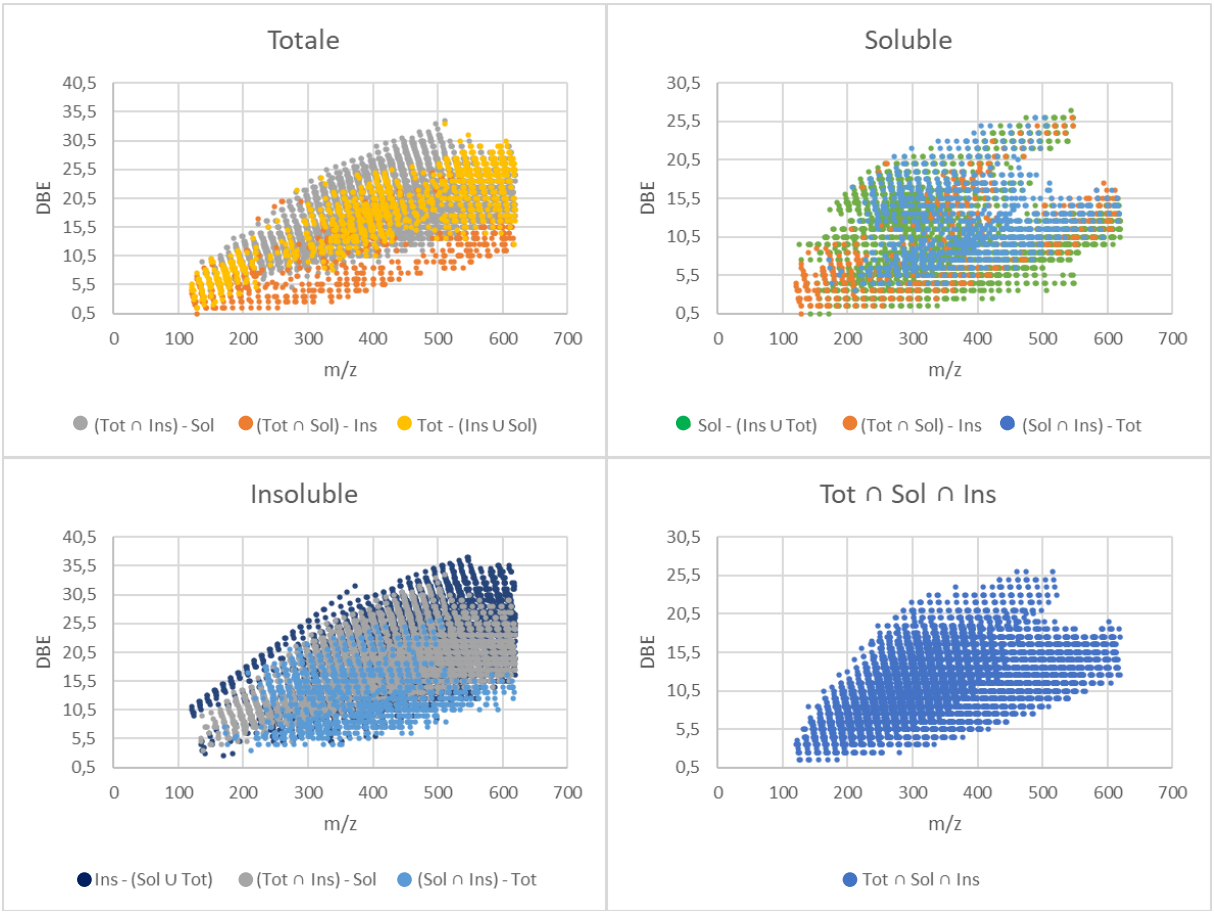


Figure 83 – Représentation des DBE en fonction de la masse pour les trois fractions, avec en évidence les différents domaines révélés par le diagramme de Venn. Les légendes indiquent les opérations d'espace relative à chaque domaine. « ∪ » est l'opérateur d'union : Ins ∪ Tot représente alors l'espace de toutes les molécules composant Ins et Tot. « ∩ » est l'opérateur d'intersection : Ins ∩ Tot représente alors l'espace des molécules uniquement présentes dans Ins et dans Tot. L'opération « – » est la soustraction d'un ensemble à un autre : Ins – (Sol ∪ Tot) représente alors l'ensemble des molécules qui ne sont présentes que dans Ins ; (Tot ∩ Ins) – Sol représente alors l'ensemble des molécules communes entre Tot et Ins, auquel on retire les molécules également présentes dans Sol.

On note également dans la phase soluble et dans la fraction commune à tous les échantillons la présence d'un deuxième lobe de molécules, qui semble distinct du lobe principal observé sur la fraction totale et insoluble. Une telle séparation est soit (1) la preuve d'un mécanisme secondaire lors de la synthèse, (2) la preuve d'une altération lors de l'ionisation, (3) un problème de sensibilité qui ne permet pas de détecter les molécules apparemment manquantes pour ne former qu'un unique fuseau. Du fait que l'on utilise une source LDI qui crée un plasma, on s'attend à des recombinaisons chimiques de ce fait, et donc à observer une partie de l'échantillon qui est altéré par l'ionisation. La présence de ce double fuseau semble en indiquer la possibilité dans les analyses effectuées, quand bien même il a été pris soin d'ajuster la puissance du laser pour en limiter les effets par différents essais effectués avant les acquisitions.

Une autre façon d'explorer la diversité moléculaire, en comparant les représentations de ratios d'hétéroatomes en fonction de la masse est présenté en Figure 84 et Figure 85. Les représentations des O/C sont relativement similaires en polarité positive en LDI, alors qu'en polarité négative, des différences significatives sont observables. Des différences intéressantes sont également visibles dans les représentations en azote, avec des distributions uniques pour la fraction insoluble en polarité négative par exemple. Ces représentations montrent également sans surprise la nature polymérique de l'ensemble, que ce soit la matière soluble, insoluble ou totale. De façon générale, on remarque tout de même que la matière insoluble est plus azotée que la matière soluble, sans forcément observer une variation significative du nombre

d’oxygène incorporé concernant les analyses en polarité positive. On remarque également que l’analyse en polarité positive présente une diversité en hétéroatomes moindre comparées aux analyses en polarité négative. Cette différence de rendement d’ionisation est peut-être explicable par des structures qui s’ionisent préférentiellement en perdant un proton ou un électron, générant alors des ions ou radicaux négatifs. Ces structures peuvent a priori être composées de fonctions acides, connues pour perdre facilement un proton, ou alors présenter des squelettes carbonés insaturés conjugés, permettant de stabiliser la charge dans l’espace.

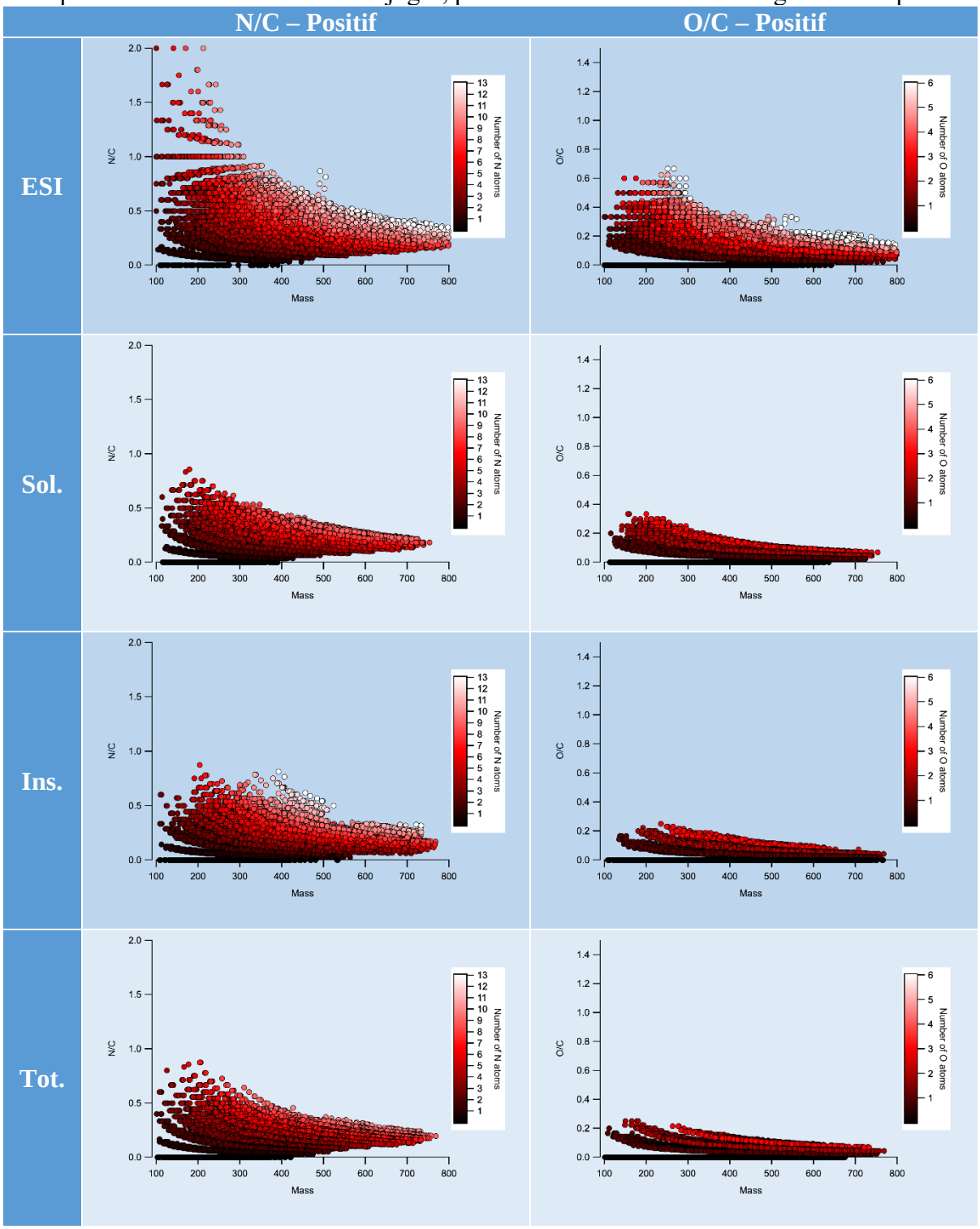


Figure 84 – Représentation des attributions en polarité positive des attributions ICR pour différentes fractions d’un même échantillon.

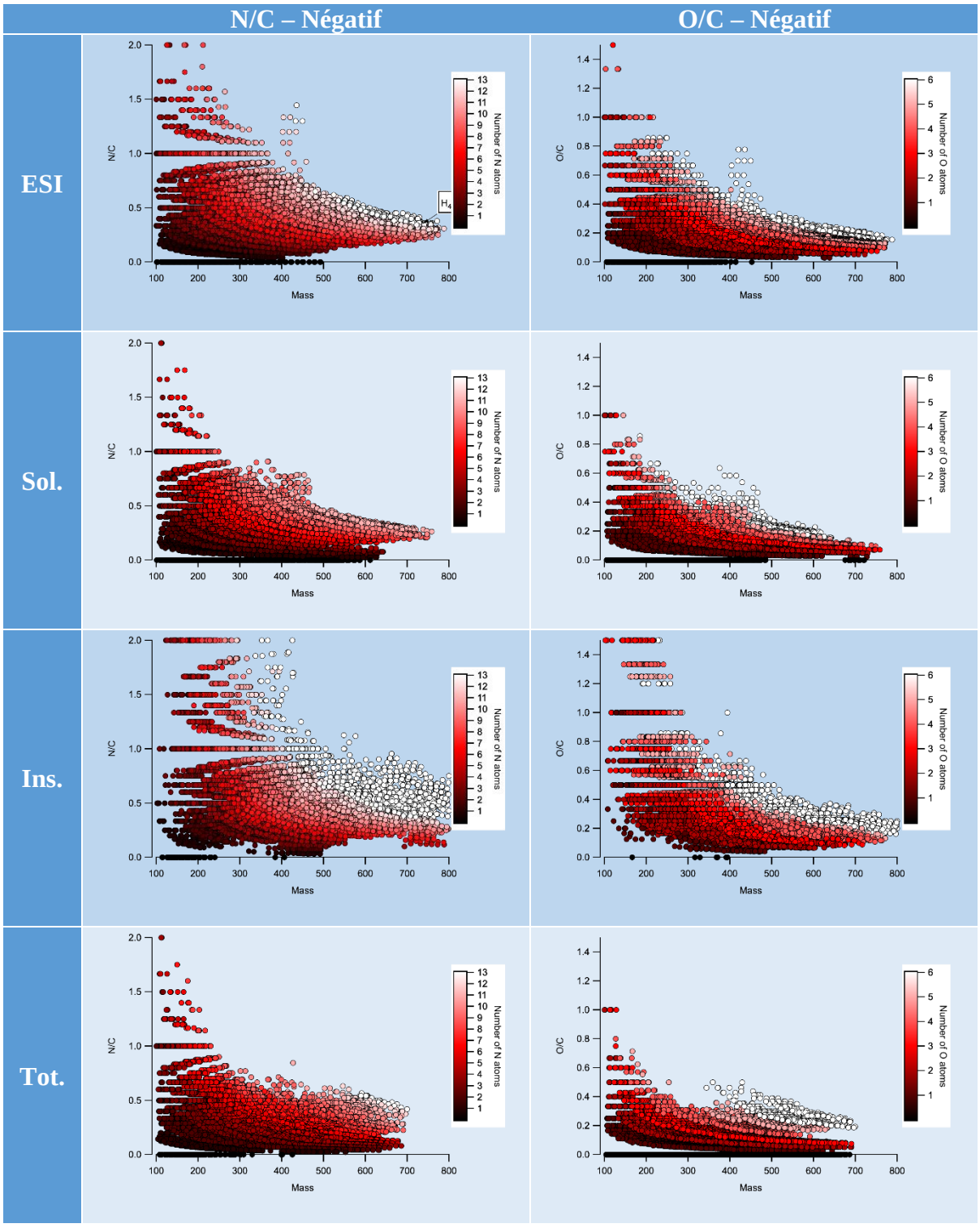


Figure 85 – Représentation des attributions en polarité négative des attributions ICR pour différentes fractions d'un même échantillon.

4.3. Chromatographie

L'intérêt de l'analyse en chromatographie pour les échantillons complexes réside dans sa capacité à pouvoir séparer des isomères, et à pouvoir identifier des composés que l'on trouve intéressant. Ces deux points ne sont pas si simples à mettre en œuvre et ont nécessité les développements décrits dans les chapitres 2.2, 2.3 et 3 pour être rendus possibles, à savoir le développement et la validation des méthodes chromatographiques, l'outil de prévision des temps de rétention pour réduire l'espace des molécules probables et le développement du

logiciel de traitement des données issues de la chromatographie. Le problème de coalescence des pics du fait du traitement de données effectuée, discuté en 3.2.2, peut cacher des résultats et l'ensemble de ses données devra être traité à nouveau une fois le traitement de données corrigé. Néanmoins, une bonne partie des données ne sont pas impactées par ce problème et peuvent donc être discutées ci-dessous.

Pour obtenir la meilleure sensibilité possible, la gamme de masse [50-550] Da est coupée en blocs de 50 Da chacun, engendrant pour l'analyse d'un échantillon pas moins de 48 cartes ioniques individuelles à analyser, sans compter les injections de mélange de contrôle pour vérifier que la méthode fonctionne de façon nominale et pouvoir effectuer la prédiction des temps de rétention. L'ensemble de ces analyses mis bout à bout représente plus de quatre jours d'analyses sans interruptions. Le Tableau 18 présente un récapitulatif de la séquence analytique réalisée.

AC POS	AC NEG	AcF NEG	AcF POS
Conditionnement	Conditionnement	Conditionnement	Conditionnement
Blanc	Blanc	Blanc	Blanc
Mélange Contrôle	*Mélange Contrôle*	*Mélange Contrôle*	*Mélange Contrôle*
Blanc	Blanc	Blanc	Blanc
SIM 50-100	SIM 50-100	SIM 50-100	SIM 50-100
SIM 80-130	SIM 80-130	SIM 80-130	SIM 80-130
SIM 120-170	SIM 120-170	SIM 120-170	SIM 120-170
SIM 160-220	SIM 160-220	SIM 160-220	SIM 160-220
SIM 210-260	SIM 210-260	SIM 210-260	SIM 210-260
SIM 250-300	SIM 250-300	SIM 250-300	SIM 250-300
SIM 290-340	SIM 290-340	SIM 290-340	SIM 290-340
SIM 330-380	SIM 330-380	SIM 330-380	SIM 330-380
SIM 370-420	SIM 370-420	SIM 370-420	SIM 370-420
SIM 410-460	SIM 410-460	SIM 410-460	SIM 410-460
SIM 450-500	SIM 450-500	SIM 450-500	SIM 450-500
SIM 490-550	SIM 490-550	SIM 490-550	SIM 490-550
Mélange Contrôle	*Mélange Contrôle*	*Mélange Contrôle*	*Mélange Contrôle*

Tableau 18 – Séquence analytique effectuée pour l'acquisition des données terminales en chromatographie. AC = méthode à pH basique ; AcF = méthode à pH acide.

Chaque analyse est effectuée en injectant 10µl de solution concentrée en mode « µl-PickUp » qui permet de ne consommer que les 10µl injectés, à la différence d'une injection de type « Full-Loop » qui consomme 160µl d'échantillon pour 50µl effectivement injectés. On réalise également une injection du mélange de prédiction des temps de rétentions au début et à la fin d'un bloc d'analyse pour estimer la déviation des temps de rétentions au cours du temps, et ainsi pouvoir s'assurer de pouvoir comparer les temps de rétention entre les injections en polarité positive et négative.

Les données sont traitées systématiquement et aucun traitement manuel n'est effectué avant l'étape de prédiction des temps de rétention. Seuls les signaux qui présentent une équivalence dans la base de données sont vérifiés manuellement pour s'assurer que les temps de rétention sont corrects avant d'effectuer la modélisation. L'ensemble des données est traité sur la gamme temporelle [2-40] minutes, et le niveau de bruit fixé à trois sigmas. On remarque a posteriori que les niveaux de bruits sont équivalents pour les analyses de l'échantillon complexe entre les différentes gammes de masses et entre les différentes méthodes.

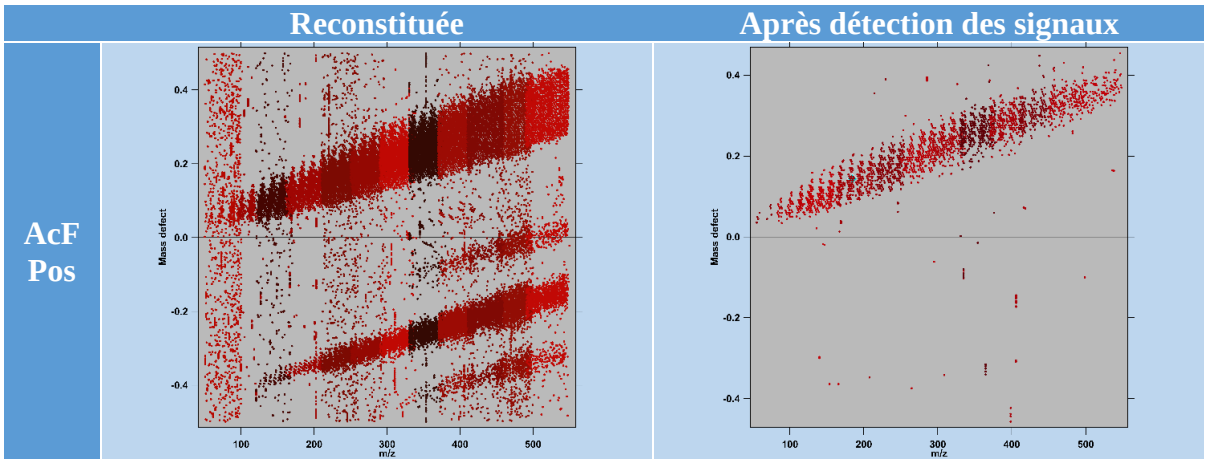
4.3.1. Exploration des données

On vérifie tout d'abord si l'information en masse est présente et pertinente. On va alors comparer les données après reconstitution et après détection des pics, comme développé en

partie 3.2. En effet, ces deux étapes sont cruciales pour pouvoir utiliser les données de chromatographie, car si les informations en masse ne sont pas cohérentes ou même absentes, il n'y a aucun intérêt à traiter les données temporelles. On présente en Figure 86 les données reconstituées ainsi que les données issues de la détection des signaux chromatographiques. Ces données sont générées en moyennant les spectres de masse sur l'ensemble de la plage temporelle considérée.

On peut remarquer directement que les données après détection des signaux sont très majoritairement incluses dans le fuseau de molécules organiques traditionnel. Cela signifie que l'ensemble des signaux détectés en chromatographie sont des molécules organiques carbonées, et que le traitement automatique effectué est capable de retirer efficacement une grande partie des signaux qui ne sont pas pertinents. Cependant, on note également que la densité d'information après détection des signaux est grandement diminuée comparée à la densité que l'on peut observer sans détection des signaux chromatographiques. Cela s'explique par le fait que la détection des signaux nécessite une quantité minimale d'information pour permettre la modélisation, et donc qu'une grande partie des signaux visible dans le spectre reconstitués ont une durée temporelle trop faible pour être conservés et traités par les algorithmes de détection des signaux.

On observe également que les données reconstituées en polarité positive présentent de multiples fuseaux, incluant les fuseaux des molécules chargées une fois et deux fois. D'autres fuseaux sont présents et sont explicables par la coalescence de pics dont la masse subit un changement majeur. Ces pics sont néanmoins filtrés par le système et ne sont pas retenus par l'algorithme de détection. Ils révèlent cependant qu'il y a un intérêt à résoudre le problème de coalescence et voir si l'on peut récupérer plus de signaux après traitement. On note également une sensibilité plus importante en polarité positive qu'en polarité négative. Cette différence est un biais instrumental qui ne peut être corrigé, ce genre de différence étant également visible en infusion directe.



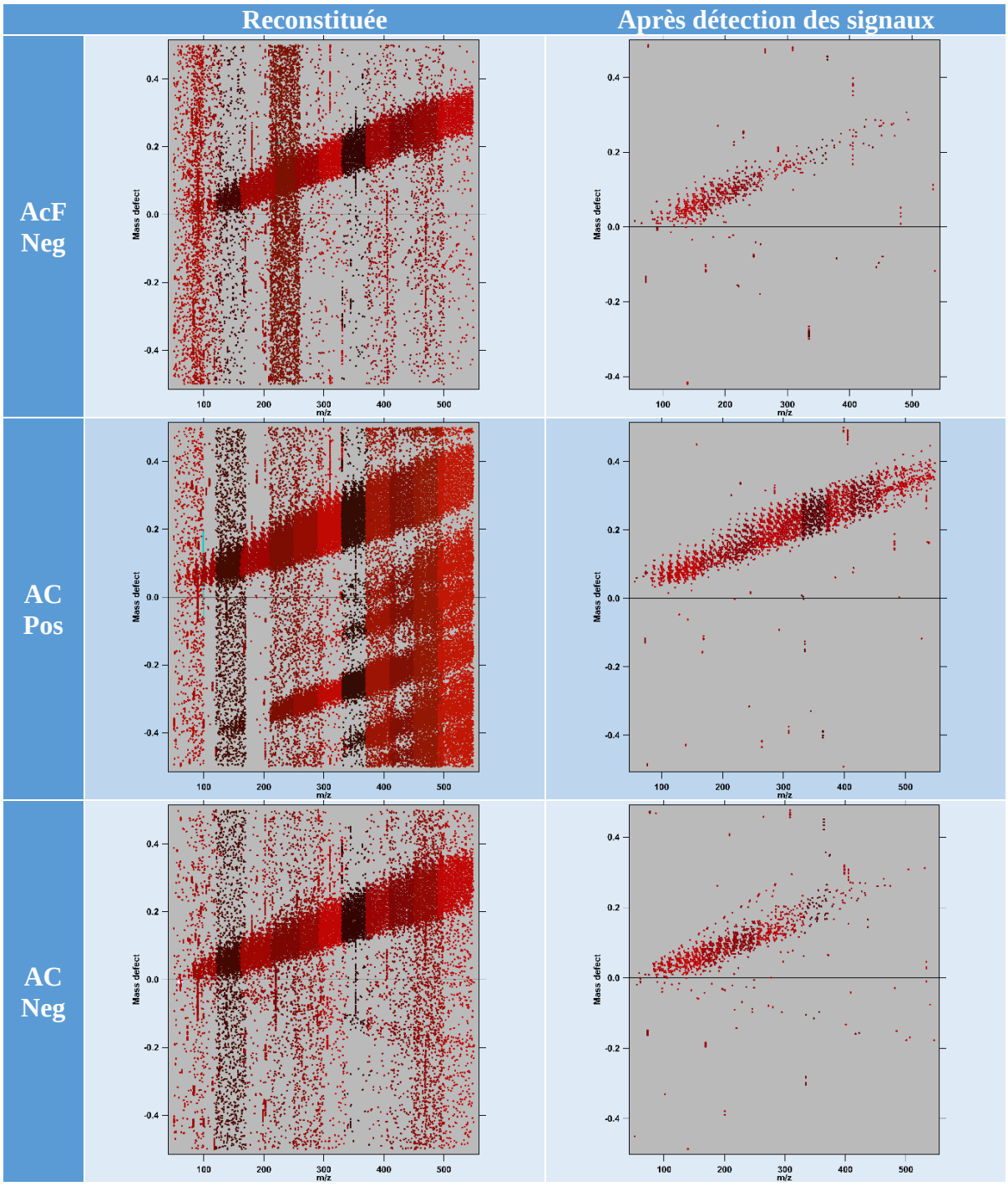


Figure 86 – Comparaison des données après reconstitution et après détection des signaux chromatographiques. AC = Ammonium Carbonate, méthode basique ; AcF = Acide Formique, méthode acide. Les différentes couleurs représentent les différentes gammes de masses effectuées.

On note également une différence importante de rétention entre les faibles gammes de masses et les gammes de masses plus élevées. En effet, comme illustré en Figure 87 où l'on présente la gamme de masse [120-170] Da et la gamme de masse [410-460] Da à titre d'exemple, les composés ayant des masses élevées présentent une séparation claire en plusieurs clusters bien délimités, alors que la faible gamme de masse présente une diversité de temps de rétentions.

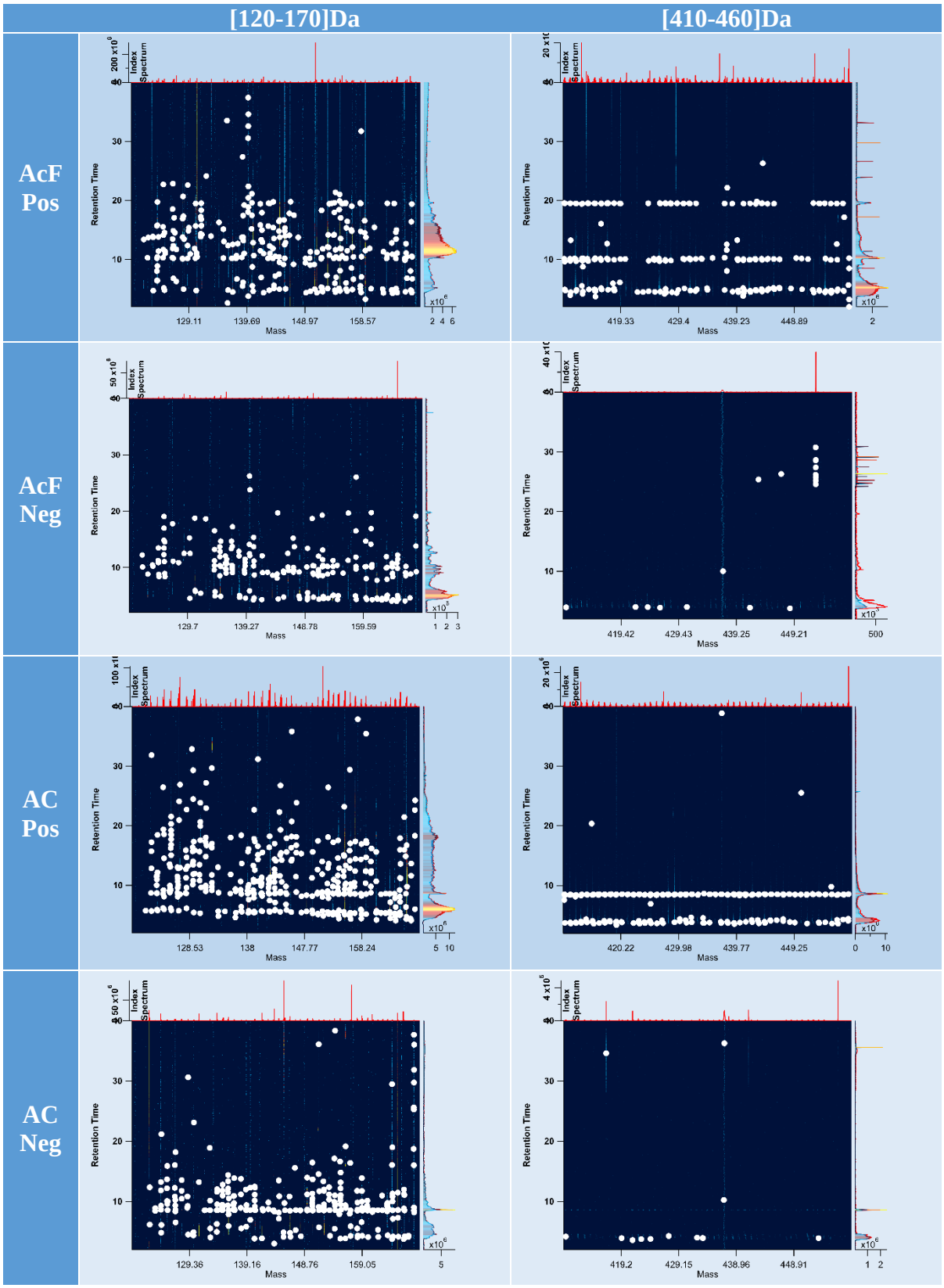


Figure 87 – Comparaison des données sur deux gammes de masses différentes. AC = Ammonium Carbonate, méthode basique ; AcF = Acide Formique, méthode acide.

Cette diversité à faible masse confirme que la colonne est pertinente pour l'analyse de ce type d'échantillon. Néanmoins, la perte de diversité des temps de rétentions à plus haute masse indique également un comportement différent des composés à faible masse comparé aux plus

hautes masses. *A priori*, on peut estimer que cette différence provient d'une différence de polarité des molécules entre la faible gamme de masse et la gamme plus importante, différence qui peut être induite par la taille importante des chaines carbonées comparé au nombre de groupements polaires pour les molécules ayant des masses élevées. À contrario, pour les masses les plus faibles, la bonne inclusion de groupements polaires comparé à la taille des chaines carbonées fait que les molécules sont plus polaires, et interagissent mieux avec la colonne, générant une meilleure séparation.

4.3.2. Recherche d'isomères

Rechercher des isomères revient à chercher des composés ayant des temps de rétention différents pour une même masse. Il est également d'intérêt d'identifier les isomères entre les différentes polarités et entre les différentes méthodes. En effet, avoir plusieurs détections en fonction de la polarité et de la méthode, permet de contraindre d'autant plus les structures moléculaires. Pour ce faire, il est nécessaire d'attribuer les signaux et d'associer à chaque molécule son temps de rétention. L'attribution des signaux est effectuée classiquement, comme présenté en partie 2.1.2 : les signaux chromatographiques sont convertis en spectre de masse et une attribution effectuée. Si plusieurs masses sont présentes, alors cette dernière est attribuée autant de fois que nécessaire et l'ensemble des attributions identiques sont conservées, permettant de réattribuer dans un second temps à chaque signal attribué son temps de rétention et de pouvoir exporter les données.

Le principal problème de cette étape est de ne pas considérer des molécules qui n'ont aucun sens, et donc de supprimer l'ensemble des molécules ayant des erreurs non-acceptables. Dans les faits, ce filtrage est fait manuellement sur l'erreur. En effet, comme présenté en Figure 88 où l'on représente la distribution des erreurs dans l'ordre croissant pour la gamme de masse [120-170], la distribution est visuellement partitionnée, avec un saut dans l'erreur observée. Ce saut est visible pour l'ensemble des données attribuées en chromatographie, et semble donc être caractéristique. Ce saut est également présent pour les données en infusion directe, mais a cependant une tendance continue qui fait qu'il n'est usuellement pas possible de manuellement placer une coupure de cette sorte. Ainsi, toute erreur supérieure à ce seuil est retirée et seules les attributions ayant une erreur inférieure sont alors discutées ; dans le cas présenté, toutes les erreurs supérieures à 1 ppm sont retirées. Une analyse à postériori non présentée ici confirme que ce filtrage grossier suffit à retirer la grande majorité (~95%) des signaux incorrects.

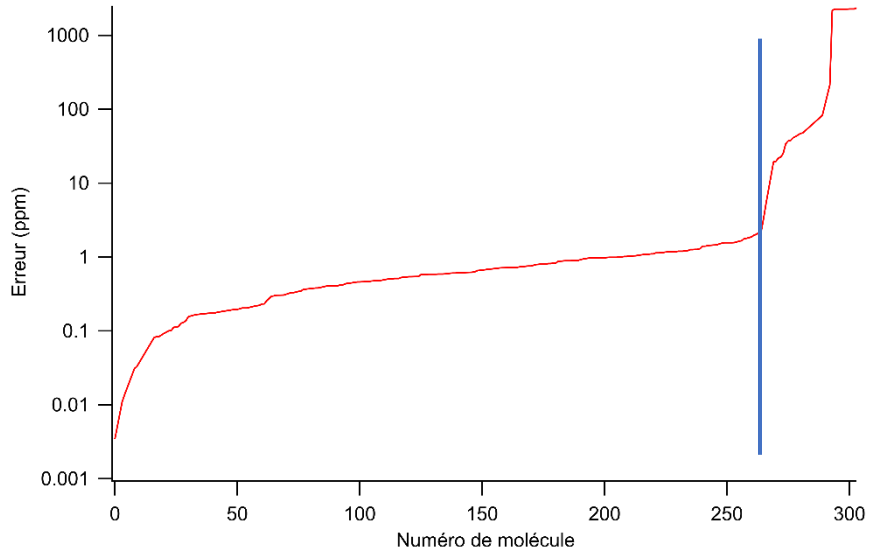


Figure 88 – Illustration de la distribution de la valeur absolue de l'erreur en échelle logarithmique pour les signaux attribués sur la gamme de masse [120-170]Da, analyse AcF polarité positive, données triées par erreur croissante,. La barre verticale bleue est placée arbitrairement au saut d'erreur observé, et toutes les attributions au-delà sont supprimée.

Les comparaisons comprenant l'ensemble des analyses, soit 48 chromatogrammes, donnent les résultats suivants :

- Nombre total de formules uniques
  - Infusion directe : 6023
  - Chromatographie : 4401
- Méthode Acide :
  - Positif : 1888 // Négatif : 532
- Méthode Basique :
  - Positif : 2405 // Négatif : 839
- Nombre de signaux communs
  - Infusion directe – chromatographie : 2090
  - Méthodes chromatographiques
    - Positif : 984
    - Négatif : 58
    - Acide/Basique : 1073

Ces résultats indiquent l'utilité des analyses sur différentes méthodes et polarités, puisqu'une grande majorité des composés ne sont visibles que dans une seule méthode et polarité. Un autre point intéressant est le faible recouvrement entre les attributions issues de la chromatographie et l'infusion directe. On note également que cela intervient majoritairement à hautes masses, indiquant un problème probable de calibration. Recalibrer les données par partie est malheureusement complexe à hautes masses, du fait de la perte de précision et de résolution de l'Orbitrap à hautes masses. De ce fait, les attributions à hautes masses ne sont pas pertinentes (i.e. après la masse 300Da), et les comparaisons doivent être effectuées à nouveau. L'ensemble des résultats qui suivent excluent alors l'ensemble des résultats qui concernent les masses supérieures à 300Da, soit 24 chromatogrammes :

- Nombre total de formules uniques
  - Infusion directe : 2093
  - Chromatographie : 1986
- Méthode Acide :
  - Positif : 780 // Négatif : 439
- Méthode Basique :
  - Positif : 907 // Négatif : 668
- Nombre de signaux communs
  - Infusion directe – chromatographie : 1110
  - Méthodes chromatographiques
    - Positif : 539
    - Négatif : 58
    - Acide/Basique : 623

L'ensemble de ces quatre méthodes permet de détecter plus de 554 isomères sur 1986 formules attribuées, ce qui représente près de 30% de molécules détectées ayant au moins un isomère de séparé par au moins une des analyses. Ce nombre est important puisque cette information isomérique n'est pas accessible autrement que par chromatographie, et donc le fait d'en détecter confirme que ce choix de colonne permet effectivement de résoudre des isomères dans cet échantillon. On peut également représenter le nombre d'isomères par masses détectés, comme présenté en Figure 89. On remarque que seule la méthode à pH basique en polarité positive présente des détections d'isomères multiples alors que les trois autres méthodes sont limitées à deux isomères détectés. Cette limitation a deux sources : la faible sensibilité de l'Orbitrap, ainsi que le problème de coalescence des ilots mentionnés en partie 3.2.2. Si le problème de sensibilité requiert un nouvel instrument, le problème de coalescence des signaux nécessite un peu plus de temps de développement de logiciel pour être corrigé.

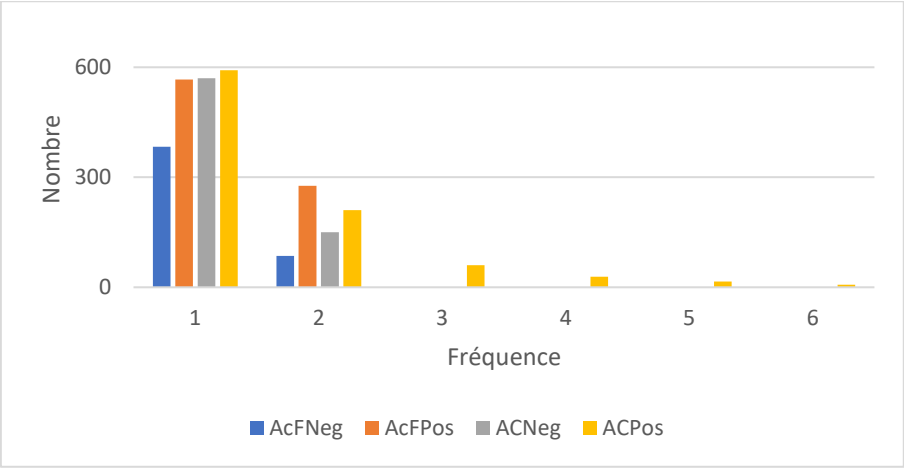


Figure 89 – Histogramme du nombre d'isomères par masse pour les quatre méthodes.

On présente également le pourcentage de formules stœchiométriques ayant au moins un isomère détecté par gamme de masse en Tableau 19. On note la détection de plus d'isomères sur les faibles gammes de masses que sur les gammes de masses plus élevées. Cela peut s'expliquer par une différence de sensibilité sur les hautes masses puisque fondamentalement, moins de signaux sont détectés. Mais cela peut également être une perte de résolution temporelle du fait de trop faibles différences de propriétés physico-chimiques des composés analysés, engendrant une interaction pauvre avec la colonne.

	AcF Neg	AcF Pos	AC Neg	AC Pos
Nb. Total de formules	440	781	669	908
Nb. d'isomères détectés	82	262	140	353
Nb. Moy. Isomères	18,3%	32,8%	20,8%	35,5%
50-100	50,0%	29,4%	0,0%	63,6%
80-130	20,7%	44,3%	44,6%	70,2%
120-170	19,7%	44,7%	29,6%	62,9%
160-220	19,2%	30,7%	19,3%	32,3%
210-260	14,6%	28,9%	8,8%	17,5%
250-300	10,3%	25,0%	11,9%	22,5%

Tableau 19 – Pourcentage de formules ayant au moins un isomère par gamme de masse.

4.3.3. Annotation de composés

Comme présenté en partie 2.3.1, annoter des composés est utile pour réduire la liste des composés possibles avant d'effectuer une éventuelle identification. L'information fournie par la chromatographie, en complément des formules stœchiométriques, permet de réduire la liste des composés possibles en se basant sur la prédiction des temps de rétention, développée en partie 2.3.2. Cette analyse est effectuée en deux étapes : extraction des molécules ayant un équivalent dans la base de données, puis prédiction des temps de rétention des composés ainsi extraits.

Une classification partielle entre les formules stœchiométriques détectées et la base de données de molécules organique carbonée est indiqué en Tableau 20, où 686 formules stœchiométriques détectées sont présentes dans la base de données [45]. Seules les familles moléculaires les plus représentées sont indiquées ici par soucis de clarté. Il convient également de mentionner que les nombres indiqués sont probablement sous-estimés puisque seul le

premier isomère de la liste des isomères possibles dans la base de données est considéré pour effectuer cette classification a priori. Autrement dit, si une formule possède 150 isomères dans la base de données, seule la classe moléculaire du premier isomère de la liste est considérée pour effectuer cette classification. On note avec intérêt la présence potentielle de nombreux acides aminés et peptides, ainsi que des base nucléiques. Il est également intéressant de remarquer que certains composés possèdent des isomères dans différentes méthodes, bien qu'aucun ne soit détecté dans les quatre analyses à la fois. On notera enfin que l'ensemble des occurrences ne présentent pas systématiquement des isomères, puisque par exemple, pour les 71 acides aminés éventuels, seuls 40 présentent des isomères.

	Nombre d'occurrences	Possède des isomères dans 1 analyse	Possède des isomères dans 2 analyses	Possède des isomères dans 3 analyses
Amino Acid Metabolism	71	28	10	2
Biosynthesis of Secondary Metabolites	30	12	2	0
Lipids: Fatty Acyls	31	9	0	0
Nucleotide Metabolism	24	7	5	1

Tableau 20 – Classification des formules stœchiométriques. Les nombres indiqués ne prennent en compte que le premier isomère possible de chaque formule stœchiométrique, les valeurs sont sous-estimées. Aucun isomère n'est détecté dans les 4 analyses à la fois ; de nombreux signaux ne présentent pas d'équivalent dans les autres analyses.

Une fois que la liste de l'ensemble des temps de rétentions associés aux formules stœchiométriques a été extraite, on injecte ces informations dans l'outil de prédiction des temps de rétention. On rappelle que la section 2.3 traite de cet outil, où il est présenté différentes calibrations des temps de rétentions. Les données utilisées étant issues de cette séquence analytique, les données de calibration déterminées en section 2.3 sont utilisées directement pour effectuer les prédictions de temps de rétentions de cette partie.

On présente en Tableau 21 le résultat synthétique de la prédiction des temps de rétention où l'on a extrait les familles ayant le plus de molécules. On peut potentiellement annoter au total 385 molécules en utilisant la méthode acide et 280 en utilisant la méthode basique, sachant que les listes initiales comportent respectivement 1038 et 961 temps de rétentions.

	Méthode Acide		Méthode basique	
	Match	No match	Match	No match
Amino Acid Metabolism	51	46	21	56
Biosynthesis of Secondary Metabolites	14	8	11	12
Lipids: Fatty Acyls	18	32	6	5
Nucleotide Metabolism	49	5	46	14

Tableau 21 – Résultats synthétiques de la prédiction des temps de rétention.

Le but de ce genre d'étude ciblée est souvent d'aller chercher les acides aminés protéinogènes et les base nucléiques. On présente un exemple parmi d'autres illustrant l'intérêt de la prédiction des temps de rétention en Tableau 22. Ici, les méthodes acides et basiques permettent d'identifier la 5-Methylcytosine sur les deux méthodes, et potentiellement d'exclure la 3-Methylcytosine comme étant présente dans le mélange. On note également que la base de données n'est pas complète puisque le dernier temps de rétention de la méthode basique ne présente pas de bon match du fait que la molécule potentielle a déjà été attribuée au temps de rétention précédent.

	tR expérimental (min)	tR prédiction (min)	Annoté	Nom
Méthode acide	10.29	13.19	Yes	5-Methylcytosine
	10.29	**10.11**	**Yes**	**2-O-Methylcytosine**
	10.29	11.47	Yes	3-Methylcytosine
	15.32	**13.19**	**Yes**	**5-Methylcytosine**
	15.32	10.11	No	2-O-Methylcytosine
	15.32	11.47	Yes	3-Methylcytosine
Méthode basique	**8.81**	**8.38**	**Yes**	**5-Methylcytosine**
	8.81	5.08	No	2-O-Methylcytosine
	8.81	5.31	No	3-Methylcytosine
	9.97	8.38	Yes	5-Methylcytosine
	9.97	5.08	No	2-O-Methylcytosine
	9.97	5.31	No	3-Methylcytosine

Tableau 22 – Illustration de la complémentarité entre méthodes pour une série de trois isomères de formule $C_5H_7N_3O$, dont seulement le 5-méthylcytosine est un dérivé de base nucléiques. En gras, les annotations retenues pour ce couple temps de rétention et masse.

Si l'on reprend l'article de Moran et al [26] qui dresse une liste des molécules potentielles, on peut aller chercher spécifiquement ces composés et vérifier si les temps de rétention prédis indiquent ou non la possibilité de présence de ces molécules. Parmi l'ensemble des composés suspectés dans la Table 7 [26], seuls quelques composés sont détectés. Sur l'ensemble, seul la 3-(Pyrazol-1-yl)-L-alanine est annotée, les autres présentant des variations trop importantes, même si les statistiques indiquent que ces composés sont dans le domaine des possibles. On peut également rejeter la présence d'histidine en se basant sur ces données.

tR exp. (min)	Formule stœchiométrique	tR pred. (min)	Annoté	Nom et famille moléculaire	
Méthode Acide					
9.05	$C_8H_9NO_2$	15.89	Yes	2-Phenylglycine	
10.14	$C_8H_9NO_2$	15.89	Yes	2-Phenylglycine	
10.19	$C_8H_{11}NO_2$	16.32	Yes	Dopamine	Amino Acid Metabolism
5.04	$C_6H_9N_3O_2$	23.06	No	L-Histidine	Amino Acid Metabolism
9.13	$C_6H_9N_3O_2$	23.06	No	L-Histidine	Amino Acid Metabolism
5.04	$C_6H_9N_3O_2$	23.06	No	L-Histidine	Amino Acid Metabolism
10.19	$C_6H_9N_3O_2$	23.06	No	L-Histidine	Amino Acid Metabolism
5.04	$C_6H_9N_3O_2$	18.53	No	3-(Pyrazol-1-yl)-L-alanine	
9.13	$C_6H_9N_3O_2$	18.53	No	3-(Pyrazol-1-yl)-L-alanine	
10.19	$C_6H_9N_3O_2$	18.53	Yes	3-(Pyrazol-1-yl)-L-alanine	
8.86	$C_7H_{11}N_3O_2$	21.66	No	α-methylhistidine	
4.82	$C_7H_{11}N_3O_2$	21.66	No	α-methylhistidine	
5	$C_{12}H_{21}N_3O_3$	20.69	No	L-Pyrrolysine	

Méthode basique					
5.47	$C_6H_9N_3O_2$	16.39	No	L-Histidine	Amino Acid Metabolism
9.84	$C_6H_9N_3O_2$	16.39	No	L-Histidine	Amino Acid Metabolism
11.03	$C_6H_9N_3O_2$	16.39	No	L-Histidine	Amino Acid Metabolism
8.78	$C_6H_9N_3O_2$	16.39	No	L-Histidine	Amino Acid Metabolism
5.47	$C_6H_9N_3O_2$	11.27	No	3-(Pyrazol-1-yl)-L-alanine	
9.84	$C_6H_9N_3O_2$	11.27	Yes	3-(Pyrazol-1-yl)-L-alanine	
11.03	**$C_6H_9N_3O_2$**	**11.27**	**Yes**	**3-(Pyrazol-1-yl)-L-alanine**	
8.78	$C_6H_9N_3O_2$	11.27	Yes	3-(Pyrazol-1-yl)-L-alanine	
5.01	$C_9H_{11}NO_2$	12.48	No	L-Phenylalanine	Amino Acid Metabolism

Tableau 23 – Extraction des molécules détectées en chromatographie à partir des données publiées dans Moran et al [26]. Seule la 3-(Pyrazol-1-yl)-L-alanine est possiblement présente, les autres sont rejetés basé sur les données disponibles.

Si l'on analyse en détail les données présentées dans les Tableau 22 et Tableau 23, on remarque également que plusieurs isomères ne présentent aucun match avec les composés prédis, indiquant la présence potentielle de composés non présents dans la base de données. Si l'on considère alors l'ensemble des données passées dans l'outil de prédiction des temps de rétention, on remarque également que près de 70% des composés détectés ne sont pas présents dans la base de données :

- Nombre de formules brutes entrées dans l'outil de prédiction des temps de rétention
  - Méthode acide : 437
  - Méthode basique : 401
- Nombre de formules brutes annotées :
  - Méthode acide : 91
  - Méthode basique : 63
- Nombre de formules brutes non présentes dans la base de données :
  - Méthode acide : 307
  - Méthode basique : 311

Ce manque de données était attendu puisque l'échantillon analysé n'est pas un échantillon biologique, et met en évidence la diversité importante de molécules et structures possibles dans les échantillons synthétiques. De nombreuses molécules d'intérêt peuvent potentiellement être présentes dans ces échantillons, et peuvent alors être ajoutées à la base de données si nécessaire.

4.4. Conclusion

On a présenté dans ce chapitre l'aboutissement du développement des méthodes analytiques et du logiciel de traitement des données issues du couplage entre chromatographie et spectrométrie de masse. On a présenté dans un premier temps les analyses en infusion directe ESI-Orbitrap, où trois échantillons différents ont été analysés. Leur diversité moléculaire a été comparée et des différences notables d'incorporation des hétéroatomes sont observées, différence qui peut être principalement explicable par la différence de composition élémentaire du gaz réactif. Du fait des limites de l'Orbitrap, à savoir une résolution et une précision limitée, un des trois échantillons a été analysé en ESI-ICR dans le but de comparer leurs résultats. On montre que les analyses Orbitrap n'accèdent qu'aux ions les plus intenses, mais que leur représentativité n'est pas significativement différente des analyses ICR puisque les

compostions élémentaires accédées sont équivalentes entre les deux instruments, et comparable avec les analyses en IRMS.

Les analyses en phase soluble effectuées en ESI-Spectrométrie de masse n'accèdent, par définition, qu'à une unique partie des échantillons. Ainsi, on a présenté les résultats obtenus pour l'analyse en LDI-ICR pour la fraction soluble, insoluble et totale d'un des trois échantillons où les données présentent des résultats intéressants entre phase soluble et insoluble, avec une partie des molécules qui ne sont pas communes entre les deux fractions, ainsi qu'une large partie qui est commune à la phase soluble, insoluble et totale. On note également que la phase insoluble semble présenter une quantité d'azote supérieure à la phase soluble, ainsi qu'une insaturation moyenne plus importante, ce qui explique que cette matière ne soit pas soluble *a priori*.

Enfin, cet échantillon a été analysé par chromatographie liquide couplée à l'Orbitrap, et les résultats complémentent les analyses en infusion directe, avec la détection de nombreux isomères possibles, incluant de multiples acides aminés, base nucléiques et leurs dérivés. On a utilisé l'outil de prédiction des temps de rétention pour réduire la liste des possibles, et de multiples annotations sont alors possibles, réduisant la liste des composés probables à analyser dans le cas d'une identification. En se basant sur les données publiées par Moran et al [26], on peut également *a priori* exclure la présence de quelques composés supposés, et par exemple annoter la 3-(Pyrazol-1-yl)-L-alanine comme potentiellement présente, et non son isomère structurel, l'Histidine. Les analyses indiquent également la présence de multiples isomères et molécules inconnus à la base de données.

5. Conclusion et perspectives

Problématique et démarche

En planétologie et astrophysique, presque tous les objets d'intérêt ne sont accessibles que par des observations et des analyses à distance à travers des méthodes spectroscopiques. Bien que l'information ainsi récupérée soit fondamentale, elle n'est pas suffisante pour caractériser complètement la complexité moléculaire des objets observés. Des expériences d'astrophysique de laboratoire visent alors à modéliser ces objets et ainsi à analyser grâce à des instruments analytiques de pointe des analogues issus de ces expériences. Spectrométrie de masse, microscopie, chromatographie et autres techniques analytiques vont alors permettre de caractériser chaque échantillon dans le but de contraindre les modes de formation des objets observés en fonction des contraintes qui ont été imposées lors des simulations.

Cependant, la complexité de ces échantillons est telle que l'information potentiellement récupérable nécessite des traitements et méthodes particulières pour s'assurer que l'interprétation effectuée repose sur des données solides et validées. Tout au long de cette thèse, nous avons travaillé l'aspect analytique de la caractérisation moléculaire d'échantillons complexes dans le but de pouvoir fournir les outils et garanties nécessaires à un traitement robuste et valide des données. Cette démarche est effectuée à travers le développement et l'optimisation de méthodes en spectrométrie de masse et en chromatographie liquide ainsi que par le développement d'un logiciel de traitement des analyses issues de la chromatographie. Ces développements ont enfin été testés sur quelques échantillons issus d'expériences d'astrophysique de laboratoire afin de s'assurer de la validité des protocoles et méthodes proposés.

Développements et applications

À l'IPAG, des analyses par infusion directe en Orbitrap sont effectuées depuis 2008. Néanmoins, aucune étude systématique concernant l'acquisition des données n'avait été réalisée pour étudier l'impact des micro-scans et des scans en profondeur. Ainsi, ce travail effectué dans le cadre de cette thèse a montré que scans et micro-scans ne donnent pas des résultats équivalents à temps d'analyse équivalents, et qu'il faut privilégier les micro-scans uniquement. Ce travail a fait l'objet d'une publication dans *Rapid Communication in Mass Spectrometry*. En complément des méthodes d'acquisition, des méthodes systématiques d'attribution et de validation des formules stœchiométriques obtenues ont été définies pour les analyses en ESI et en LDI. Ces méthodes systématiques permettent de s'assurer d'une qualité minimale nécessaire à l'utilisation des données par la suite, sans avoir à se poser la question de la pertinence des données sur lesquels les interprétations reposent.

L'analyse en infusion directe est une première étape du processus analytique, permettant d'établir la liste de l'ensemble des formules brutes attribuées à partir des signaux détectés pour l'échantillon analysé. Cette liste représente alors la diversité associée à chaque échantillon et peut alors être comparée à d'autres listes issues d'autres échantillons. Cependant, une formule brute ne permet pas de remonter de façon univoque à une structure moléculaire : il est alors nécessaire d'ajouter une autre dimension, la chromatographie. Le développement de méthodes capables de séparer la diversité importante de molécules observées est un défi analytique du fait de la diversité fonctionnelle importante supposée des échantillons analysés. Nous avons fait le choix de pouvoir a priori séparer des composés organiques biochimiques tels que ceux observés en métabolomique, et donc d'avoir une chimie de colonne centrée sur la séparation d'espèces oxygénées et peu azotées. Les deux méthodes développées utilisent une colonne HILIC, à pH acide et basique, permettant entre autres de séparer et de comparer entre méthodes les composés sensibles à la variation de pH. Un effort particulier a été effectué pour s'assurer que les méthodes développées soient capables de donner des résultats comparables dans le

temps, et tout particulièrement en effectuant des essais de répétabilité, de reproductibilité et en mettant en place un mélange de références jouant le rôle de contrôle qualité de la méthode.

Les données issues du couplage HPLC-HRMS lors de l'analyse d'échantillons organiques complexes doivent être traitées dans la même optique que les données en infusion directe, i.e. sans a priori ou supervision. Aucun logiciel commercial n'est développé à cet effet, et les quelques logiciels libres d'accès ne répondent pas aux critères et choix que l'on souhaite. Ainsi, notre propre logiciel de traitement des données issues de la HPLC-HRMS a été développé, depuis le traitement des données initiales (bruits, alignement en masse) jusqu'à la modélisation des signaux temporels, permettant d'établir pour chaque échantillon une liste de couple [masse ; temps de rétention] qui permet d'identifier de façon unique chaque signal détecté. Ces signaux ainsi détectés sont traditionnellement comparés à des standards purs pour effectuer leur identification. Cependant, les analyses effectuées présentent des milliers de formules brutes, soit des centaines de milliers de structures potentielles. Pour réduire cet espace des solutions, un outil de prédiction des temps de rétention a été développé dans le but d'annoter des molécules, et d'exclure des molécules dont les propriétés physico-chimiques sont trop différentes des propriétés observées. Des étapes supplémentaires permettant de réduire la liste des annotations et de permettre finalement l'identification nécessitent des développements et techniques supplémentaires.

L'application des protocoles ainsi développés pour l'analyse d'échantillons d'analogues d'aérosols atmosphériques pour des exoplanètes de type super-Terres ou mini-Neptunes révèle une diversité de formules brutes et de molécules intéressantes. L'analyse par ESI-Orbitrap sur la gamme de masse [150-450] Da montre que les trois échantillons, bien que présentant des spectres de masses similaires, présentent des différences en terme : (1) de leur diversité apparente après attribution, avec une quantité non négligeable de formules stœchiométriques uniques à chacun et (2) d'incorporation d'hétéroatomes et de leur insaturation avec une différence en termes d'implantation d'oxygène et d'insaturation portée par l'azote. Ces différences, révélées par les analyses de la fraction soluble de chaque échantillon, sont surement le reflet de la différence de composition initiale de chaque échantillon. Cependant, deux d'entre eux possèdent des compositions élémentaires initiales présentant des variations peu importantes, et présentent des variations en termes de composition moléculaire importantes. Du fait du nombre important de changements entre les deux échantillons, remonter à la source de cette variation n'est pas possible et nécessiterai des synthèses complémentaires pour investiguer l'origine de cette différence.

L'Orbitrap étant limité en résolution et précision, des analyses complémentaires ont été effectuées en ESI-ICR dans le but de déterminer la pertinence des analyses Orbitrap. Il en ressort que les analyses Orbitrap effectuées sur la phase soluble représentent les signaux les plus intenses des mêmes analyses effectuées en ICR. Différentes explications sont possibles pour expliquer cette différence, l'une d'entre elles étant l'avance technologique de l'ICR utilisé comparé à l'Orbitrap, avec plusieurs générations d'améliorations entre les deux instruments. En effet, l'Orbitrap de Grenoble est la première version commerciale de cet instrument, alors que l'ICR utilisé est l'une des dernières versions commerciales. Néanmoins, même si les analyses ne sont pas équivalentes, les analyses Orbitrap et ICR de la fraction soluble sont toutes deux équivalentes à l'analyse élémentaire de l'échantillon total, indiquant que l'analyse de la fraction soluble est représentative de l'ensemble de l'échantillon, que ce soit pour l'Orbitrap ou l'ICR, la différence de sensibilité n'ajoutant alors que de la diversité comparable à la diversité moyenne. Enfin, la puissance résolutive ne commence à avoir d'intérêt, en CHNO, qu'à hautes masses où les solutions possibles pour attribuer un signal en masse sont de plus en plus nombreuses. Ainsi, sur la gamme considérée, la puissance résolutive de l'ICR n'a aucun avantage comparé à l'Orbitrap.

Même si l'analyse de la fraction soluble est l'unique moyen d'accéder à la composition moléculaire des échantillons à travers l'utilisation de la chromatographie, il est néanmoins intéressant de s'intéresser à la comparaison de la fraction soluble, insoluble et totale d'un échantillon. Ce type d'analyse est effectué en LDI-ICR puisque c'est une analyse effectuée en phase solide. Dans cette analyse, toute la puissance de l'ICR est utilisée et permet d'observer des différences notables entre fractions solubles et insolubles, et plus particulièrement sur leur degré d'insaturation moyen, avec la fraction soluble qui présente des degrés d'insaturation moindres que ceux observés pour la fraction insoluble. On note également que les signaux observés pour la fraction totale sont a priori la somme de la fraction soluble et insoluble, comme attendu. Les variations mineures observées sont attribuables à des différences de sensibilité entre les différentes analyses, ainsi qu'aux éventuelles modifications chimiques dues à l'ionisation. On observe enfin une diversité similaire en polarité positive, alors que la diversité observée est très différente en polarité négative. Cette différence de rendement d'ionisation est difficilement explicable sans essais complémentaires, mais peut éventuellement être liée à la structure des molécules qui se fragmenteraient préférentiellement sous forme d'ions négatifs.

Enfin, l'analyse chromatographique présente certes un manque de sensibilité, mais la majorité des signaux observés sont inclus dans les signaux détectés en infusion directe, bien qu'une quantité non négligeable de signaux ne sont pas communs. Cette différence est peut-être due au problème logiciel, et est donc ignoré pour le moment. On note que des signaux commun et différent sont observés sur les quatre méthodes, validant l'utilisation des quatre méthodes pour l'analyse complète des échantillons. L'analyse détaillée chromatogramme par chromatogramme indique également une bonne séparation temporelle à faible masse, et un alignement à quelques temps de rétention à plus haute masses, indiquant peut-être une différence de polarité des molécules à faibles masses comparées aux plus hautes masses. L'ensemble des quatre méthodes permet également de détecter des isomères pour plus de 30% des signaux détectés en moyenne, ce qui valide aussi l'utilisation de ce type de colonne pour l'analyse de cet échantillon. Enfin, les attributions des signaux sont comparées à la base de données de composés utilisés en métabolomique, et indiquent la présence potentielle d'acides aminés et de base nucléiques par exemple, dont certains sont potentiellement détectés par plusieurs analyses. Pour annoter ces composés, l'outil de prédiction des temps de rétention est utilisé et permet de calculer un temps de rétention théorique pour chacun des composés de la base de données. Ces temps sont alors comparés aux temps observés, et permettent alors de limiter la liste des molécules potentielles. Cette comparaison et prédiction révèle qu'une grande partie des composés détectés ne sont pas présents dans la base de données, indiquant alors la nécessité potentielle de pouvoir ajouter des composés dans la base de données si cela s'avère nécessaire.

Apports et limites des travaux

Les développements effectués au cours de cette thèse ne concernent que les analyses HRMS et HPLC-HRMS, qui ne représentent qu'une petite partie d'une chaine analytique globale. En effet, ces analyses présentent des faiblesses saillantes faisant qu'elles ne peuvent être uniquement considérées pour caractériser aussi complètement que possible un échantillon organique complexe. Si l'on ne s'intéresse qu'à la caractérisation moléculaire de tels échantillons, il est nécessaire de s'intéresser à la phase liquide soluble des échantillons puisque les techniques chromatographiques ne permettent pas d'autres formes d'échantillons. Dès lors, l'avantage des techniques développées et présentées dans ce travail de thèse est leur versatilité dans le sens où : (1) elles permettent d'accéder à une diversité importante, avec aussi peu de biais d'observation que possible, (2) elles permettent d'effectuer un premier tri dans les annotations en vue d'identifications potentielles et (3) elles permettent d'effectuer de multiples traitements de données et interprétations à partir du même set de données issues de l'instrumentation, ce qui n'est pas autorisé par toutes les techniques. D'autres techniques, telles

que la chromatographie en phase gazeuse par exemple, sont extrêmement sélectives et permettent une identification complète des composés volatils thermiquement stables à partir d'un échantillon. Cette information, bien que précieuse, ne permet pas d'avoir une information globale de la diversité présente, comme on peut l'obtenir avec la chromatographie en phase liquide. Ainsi, les méthodes proposées dans ce travail s'insèrent dans un processus global, depuis l'analyse générale par spectroscopie jusqu'à l'analyse spécifique permettant des identifications très fines et spécifiques.

Une limitation majeure de l'Orbitrap est sa précision et sa résolution insuffisante pour analyser les échantillons jusqu'à des masses élevées, i.e. après 400 Da. Néanmoins, de nouvelles techniques d'attributions telles que *Graphttribution* permettent de mitiger ce problème si l'échantillon considéré présente une diversité suffisante, permettant alors d'ajuster les techniques d'acquisitions en Orbitrap sur des gammes de masses utiles plus importantes. Une autre limitation est l'analyse de la phase soluble seule par la source ESI, nécessitant l'ajout de techniques analytiques autres tel que le LDI pour analyser les autres fractions. En plus de cette limitation de phase accessible, la source ESI a tendance à favoriser les molécules polaires acido-basiques, incluant alors un biais systématique dans les molécules accessibles avec cette technique. A priori, ce biais n'est pas forcément critique du fait du couplage avec la chromatographie lorsque la colonne sélectionnée est adaptée à ce biais. Au contraire, la source LDI a l'avantage de pouvoir analyser l'ensemble des fractions et ne semble a priori pas biaisée sur des types de composés et fonctions chimiques. On note néanmoins la problématique de l'ionisation par plasma qui peut engendrer des réactions chimiques annexes et générer de la diversité non originaire de l'échantillon en lui-même, tels que la génération de fullerènes dans l'ensemble des spectres traités. Des techniques alternatives, telles que le MALDI (*Matrix Assisted LDI*) permettent de solutionner une partie des problème mentionnés précédemment. Cependant, développer une matrice capable d'ioniser correctement et sans biais un échantillon complexe nécessite un développement analytique complet, très certainement échantillon dépendant, et n'est alors pas réalisable aisément dans le cadre d'analyses sur des échantillons complexes synthétiques qui sont, par conception, tous plus ou moins uniques.

Il convient également de sélectionner de façon appropriée le spectromètre de masse en fonction des résultats attendus. En effet, les ICR et leur possibilité de faire des spectres en ESI et LDI à haute résolution et sensibilité est intéressant, mais le coût important de l'instrument à l'achat, les contraintes de mise en œuvre (champs magnétiques intenses, liquides cryogéniques) et sa faible fréquence d'acquisition sont autant de limitations importantes qui rendent les analyses Orbitrap compétitives pour l'analyse d'échantillons complexes à faibles masses. Ainsi, effectuer des analyses Orbitrap d'échantillons complexes à faible masses, typiquement pour des masses inférieures à 400 Da, est plus efficace que d'effectuer des analyses ICR. De plus, les contraintes de l'Orbitrap permettent un couplage chromatographique beaucoup plus aisé qu'en ICR, rendant les analyses isomériques possibles en routine en Orbitrap.

Exploitations et perspectives

Continuité des développements analytiques

La priorité majeure est de corriger le logiciel de traitement des données de l'erreur de conception effectuée et qui a inclus un biais important dans le traitement des données. Ce travail est prioritaire et sera effectué dès début 2021. En complément de ce travail de correction, l'inclusion de l'outil de prédiction des temps de rétentions doit être intégré à Attributor, et un outil d'ajout de composés dans la base de données ajouté. Ce dernier point nécessite l'interaction avec un logiciel de calcul théorique, ainsi que l'ajout de formules de calcul des charges partielles en fonction du pH, formules déjà publiées dans les travaux qui ont servi de support au développement de l'outil de prédiction des temps de rétentions.

De multiples autres projets peuvent être envisagés pour poursuivre les développements analytiques présentés dans ce travail. L'un d'entre eux est le développement de nouvelles

méthodes sur des chimies de colonnes différentes, travail déjà initié par l'encadrement d'une stagiaire de M1 durant l'été 2020. Entre les méthodes présentées dans ce travail, plutôt sensibles aux composés oxygénés, et la méthode plutôt sensible aux composés azotés développée lors du stage, une séquence analytique complète peut être imaginée pour l'analyse d'échantillons complexes et alors obtenir plusieurs jeux de données plus ou moins complémentaires sur lesquels effectuer des comparaisons entre échantillons concernant l'isomérie observée.

L'instrumentation possédée à l'IPAG permet, à l'issue d'une séparation, de faire du fractionnement et ainsi d'échantillonner ce qui sort de la colonne dans le but de l'analyser sur d'autres méthodes analytiques. Cette option peut être particulièrement intéressante d'un point de vue de la préparation d'échantillon par exemple avec la possibilité de purifier une fraction temporelle non résolue en vue d'une analyse sur une chimie de colonne différente par exemple. On peut également introduire lors de l'analyse en masse des étapes de spectrométrie de masse tandem, où les composés les plus intenses sont fragmentés dans le but de pouvoir contraindre encore plus leur structure moléculaire. Du fait de la version de l'Orbitrap, cette étape n'est malheureusement pas automatisable, mais est néanmoins intéressante si l'on s'intéresse à quelques composés bien définis.

Enfin, l'ensemble des travaux présentés dans cette thèse peuvent faire parti d'une chaine analytique globale, où les données issues de l'Orbitrap et de son couplage chromatographique permettent de contraindre les analyses effectuées dans l'étape suivante, comme par exemple des analyses en chromatographie en phase gazeuse, où les composés annotés et leur classe chimique vont diriger les analyses et identifications à effectuer.

Application astrophysique

Ces travaux ont abouti sur la définition d'un projet LabEx, qui a été financé en décembre 2020. Ce projet vise à conforter le groupe Orbitrap dans son implication dans le groupe de travail préliminaire d'analyse de la fraction soluble pour la mission de retour d'échantillon Hayabusa 2. En effet, le poste de CDD ainsi financé sera chargé du traitement des données issues des analyses en chromatographie-Orbitrap effectuées par Hiroshi Naraoka à l'université de Kyushu. Le traitement des données sera supporté par le logiciel développé au cours de cette thèse, et mettra ainsi en application les systématiques et méthodes développées à cet effet. L'ensemble de ces travaux, rendu possible par le travail effectué dans cette thèse, fera l'objet d'au moins une publication dans une revue majeure.

En complément des travaux sur Hayabusa 2, le projet souhaite développer la capacité du laboratoire, et plus particulièrement du pôle Spectrométrie de masse (i.e. l'Orbitrap et le spectromètre de masse à rapport isotopique (IRMS)), à pouvoir répondre aux appels d'offres internationaux permettant d'obtenir des échantillons issus de la mission Hayabusa 2. Ce projet vise alors à conforter les protocoles analytiques développés par la constitution d'une base de données d'échantillons dont leur composition élémentaire et moléculaire a été caractérisée. Ainsi, ces données pourront être comparées aux échantillons issus de cette mission spatiale et des contraintes et interprétations effectuées. En outre de toutes les acquisitions et traitements de données qui devront être effectués dans le cadre de ce projet, un travail important de développement de la base de données sera également effectué, puisque à ce jour, aucune base de données permettant de savoir ce qui a été fait ou non sur tel échantillon n'est en place à l'IPAG. Ce travail a été initié dans le courant de l'automne 2020 au sein de l'IPAG en collaboration avec le service informatique et devra être poursuivi, mis en œuvre et éventuellement valorisé par son ouverture au-delà du pôle de spectrométrie de masse.

Ces travaux ont également un intérêt prospectif puisque divers développements instrumentaux pour les missions spatiales sont à l'étude tel que par exemple le projet COSMO-Orbitrap. Ce projet vise à spatialiser un Orbitrap, et celui-ci sera doté d'une source LDI pour l'analyse des matériaux échantillonnés. Le groupe Orbitrap de l'IPAG est impliqué dans ce projet dans le cadre de l'analyse et le traitement des données, et donc la définition de protocoles

et systématiques de traitement et d'attributions des données en LDI est une étape importante pour être capable d'analyser les données issues de cet instrument.

Bibliographie

[1] P. Schmitt-Kopplin, D. Hemmler, F. Moritz, R.D. Gougeon, M. Lucio, M. Meringer, C. Müller, M. Harir, N. Hertkorn, Systems chemical analytics: introduction to the challenges of chemical complexity analysis, Faraday Discussions. 218 (2019) 9–28. https://doi.org/10.1039/C9FD00078J.

[2] B. Marty, K. Altwegg, H. Balsiger, A. Bar-Nun, D.V. Bekaert, J.-J. Berthelier, A. Bieler, C. Briois, U. Calmonte, M. Combi, others, Xenon isotopes in 67P/Churyumov-Gerasimenko show that comets contributed to Earth9s atmosphere, Science. 356 (2017) 1069–1072.

[3] S. Bartlett, M.L. Wong, Defining Lyfe in the Universe: From Three Privileged Functions to Four Pillars, Life. 10 (2020) 42. https://doi.org/10.3390/life10040042.

[4] S.E. Cummins, P. Thaddeus, R.A. Linke, A survey of the millimeter-wave spectrum of Sagittarius B2, The Astrophysical Journal Supplement Series. 60 (1986) 819. https://doi.org/10.1086/191102.

[5] S.M. Hörst, Titan's atmosphere and climate: TITAN'S ATMOSPHERE, Journal of Geophysical Research: Planets. 122 (2017) 432–482. https://doi.org/10.1002/2016JE005240.

[6] A. Bardyn, D. Baklouti, H. Cottin, N. Fray, C. Briois, J. Paquette, O. Stenzel, C. Engrand, H. Fischer, K. Hornung, R. Isnard, Y. Langevin, H. Lehto, L. Le Roy, N. Ligier, S. Merouane, P. Modica, F.-R. Orthous-Daunay, J. Rynö, R. Schulz, J. Silén, L. Thirkell, K. Varmuza, B. Zaprudin, J. Kissel, M. Hilchenbach, Carbon-rich dust in comet 67P/Churyumov-Gerasimenko measured by COSIMA/Rosetta, Monthly Notices of the Royal Astronomical Society. 469 (2017) S712–S722. https://doi.org/10.1093/mnras/stx2640.

[7] P. Caselli, C. Ceccarelli, Our astrochemical heritage, The Astronomy and Astrophysics Review. 20 (2012). https://doi.org/10.1007/s00159-012-0056-x.

[8] V. Vuitton, R.V. Yelle, S.J. Klippenstein, S.M. Hörst, P. Lavvas, Simulating the density of organic species in the atmosphere of Titan with a coupled ion-neutral photochemical model, Icarus. 324 (2019) 120–197. https://doi.org/10.1016/j.icarus.2018.06.013.

[9] O. Poch, I. Istiqomah, E. Quirico, P. Beck, B. Schmitt, P. Theulé, A. Faure, P. Hily-Blant, L. Bonal, A. Raponi, M. Ciarniello, B. Rousseau, S. Potin, O. Brissaud, L. Flandinet, G. Filacchione, A. Pommerol, N. Thomas, D. Kappel, V. Mennella, L. Moroz, V. Vinogradoff, G. Arnold, S. Erard, D. Bockelée-Morvan, C. Leyrat, F. Capaccioni, M.C. De Sanctis, A. Longobardo, F. Mancarella, E. Palomba, F. Tosi, Ammonium salts are a reservoir of nitrogen on a cometary nucleus and possibly on some asteroids, Science. 367 (2020) eaaw7462. https://doi.org/10.1126/science.aaw7462.

[10] F.-R. Orthous-Daunay, L. Piani, L. Flandinet, R. Thissen, C. Wolters, V. Vuitton, O. Poch, F. Moynier, I. Sugawara, H. Naraoka, S. Tachibana, Ultraviolet-photon fingerprints on chondritic large organic molecules, GEOCHEMICAL JOURNAL. 53 (2019) 21–32. https://doi.org/10.2343/geochemj.2.0544.

[11] B.N. Khare, C. Sagan, H. Ogino, B. Nagy, C. Er, K.H. Schram, E.T. Arakawa, Amino acids derived from Titan Tholins, Icarus. 68 (1986) 176–184. https://doi.org/10.1016/0019-1035(86)90080-1.

[12] G. Danger, F.-R. Orthous-Daunay, P. de Marcellus, P. Modica, V. Vuitton, F. Duvernay, L. Flandinet, L. Le Sergeant d'Hendecourt, R. Thissen, T. Chiavassa, Characterization of laboratory analogs of interstellar/cometary organic residues using very high resolution mass spectrometry, Geochimica et Cosmochimica Acta. 118 (2013) 184–201. https://doi.org/10.1016/j.gca.2013.05.015.

[13] G. Danger, A. Fresneau, N. Abou Mrad, P. de Marcellus, F.-R. Orthous-Daunay,

F. Duvernay, V. Vuitton, L. Le Sergeant d'Hendecourt, R. Thissen, T. Chiavassa, Insight into the molecular composition of laboratory organic residues produced from interstellar/pre-cometary ice analogues using very high resolution mass spectrometry, Geochimica et Cosmochimica Acta. 189 (2016) 184–196. https://doi.org/10.1016/j.gca.2016.06.014.

[14] A. Fresneau, Simulations expérimentales en laboratoire pour la préparation à l'analyse des données issues de missions spatiales, ainsi que pour l'étude de l'impact en exobiologie de l'évolution de la matière organique au sein d'environnements astrophysiques, Sciences chimiques, Université d'Aix-Marseille, 2016.

[15] A. Fresneau, N.A. Mrad, L. LS d?Hendecourt, F. Duvernay, L. Flandinet, F.-R. Orthous-Daunay, V. Vuitton, R. Thissen, T. Chiavassa, G. Danger, Cometary Materials Originating from Interstellar Ices: Clues from Laboratory Experiments, The Astrophysical Journal. 837 (2017) 168. https://doi.org/10.3847/1538-4357/aa618a.

[16] C. Szopa, G. Cernogora, L. Boufendi, J.J. Correia, P. Coll, PAMPRE: A dusty plasma experiment for Titan's tholins production and study, Planetary and Space Science. 54 (2006) 394–404. https://doi.org/10.1016/j.pss.2005.12.012.

[17] J. Maillard, N. Carrasco, I. Schmitz-Afonso, T. Gautier, C. Afonso, Comparison of soluble and insoluble organic matter in analogues of Titan's aerosols, Earth and Planetary Science Letters. 495 (2018) 185–191. https://doi.org/10.1016/j.epsl.2018.05.014.

[18] L. Jovanović, T. Gautier, V. Vuitton, C. Wolters, J. Bourgalais, A. Buch, F.-R. Orthous-Daunay, L. Vettier, L. Flandinet, N. Carrasco, Chemical composition of Pluto aerosol analogues, Icarus. 346 (2020) 113774. https://doi.org/10.1016/j.icarus.2020.113774.

[19] P. de Marcellus, C. Meinert, I. Myrgorodska, L. Nahon, T. Buhse, L.L.S. d'Hendecourt, U.J. Meierhenrich, Aldehydes and sugars from evolved precometary ice analogs: Importance of ices in astrochemical and prebiotic evolution, Proceedings of the National Academy of Sciences. 112 (2015) 965–970. https://doi.org/10.1073/pnas.1418602112.

[20] C. Meinert, I. Myrgorodska, P. de Marcellus, T. Buhse, L. Nahon, S.V. Hoffmann, L. Le Sergeant d'Hendecourt, U.J. Meierhenrich, Ribose and related sugars from ultraviolet irradiation of interstellar ice analogs, Science. 352 (2016) 208–212. https://doi.org/10.1126/science.aad0371.

[21] C. Meinert, J.-J. Filippi, P. de Marcellus, L. Le Sergeant d'Hendecourt, U.J. Meierhenrich, N-(2-Aminoethyl)glycine and Amino Acids from Interstellar Ice Analogues, ChemPlusChem. 77 (2012) 186–191. https://doi.org/10.1002/cplu.201100048.

[22] P. Pernot, N. Carrasco, R. Thissen, I. Schmitz-Afonso, Tholinomics—Chemical Analysis of Nitrogen-Rich Polymers, Analytical Chemistry. 82 (2010) 1371–1380. https://doi.org/10.1021/ac902458q.

[23] S.M. Hörst, R.V. Yelle, A. Buch, N. Carrasco, G. Cernogora, O. Dutuit, E. Quirico, E. Sciamma-O'Brien, M.A. Smith, Á. Somogyi, C. Szopa, R. Thissen, V. Vuitton, Formation of Amino Acids and Nucleotide Bases in a Titan Atmosphere Simulation Experiment, Astrobiology. 12 (2012) 809–817. https://doi.org/10.1089/ast.2011.0623.

[24] C. He, S.M. Hörst, S. Riemer, J.A. Sebree, N. Pauley, V. Vuitton, Carbon Monoxide Affecting Planetary Atmospheric Chemistry, The Astrophysical Journal. 841 (2017) L31–L37. https://doi.org/10.3847/2041-8213/aa74cc.

[25] S.M. Hörst, C. He, N.K. Lewis, E.M.-R. Kempton, M.S. Marley, C.V. Morley, J.I. Moses, J.A. Valenti, V. Vuitton, Haze production rates in super-Earth and mini-Neptune atmosphere experiments, Nature Astronomy. 2 (2018) 303–306. https://doi.org/10.1038/s41550-018-0397-0.

[26] S.E. Moran, S.M. Hörst, V. Vuitton, C. He, N.K. Lewis, L. Flandinet, J.I. Moses, N. North, F.-R. Orthous-Daunay, J. Sebree, C. Wolters, E.M.-R. Kempton, M.S. Marley, C.V. Morley, J.A. Valenti, Chemistry of Temperate Super-Earth and Mini-Neptune Atmospheric Hazes from Laboratory Experiments, The Planetary Science Journal. 1 (2020) 17.

https://doi.org/10.3847/PSJ/ab8eae.

[27] C. He, S.M. Hörst, N.K. Lewis, X. Yu, J.I. Moses, E.M.-R. Kempton, P. McGuiggan, C.V. Morley, J.A. Valenti, V. Vuitton, Laboratory Simulations of Haze Formation in the Atmospheres of Super-Earths and Mini-Neptunes: Particle Color and Size Distribution, The Astrophysical Journal. 856 (2018) L3. https://doi.org/10.3847/2041-8213/aab42b.

[28] P.J. Amorim Madeira, P. A., C. M., High Resolution Mass Spectrometry Using FTICR and Orbitrap Instruments, in: S. Salih (Ed.), Fourier Transform - Materials Analysis, InTech, 2012. https://doi.org/10.5772/37423.

[29] M. Scigelova, M. Hornshaw, A. Giannakopulos, A. Makarov, Fourier Transform Mass Spectrometry, Molecular & Cellular Proteomics. 10 (2011) M111.009431. https://doi.org/10.1074/mcp.M111.009431.

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

[31] E.M. Schmidt, M.A. Pudenzi, J.M. Santos, C.F.F. Angolini, R.C.L. Pereira, Y.S. Rocha, E. Denisov, E. Damoc, A. Makarov, M.N. Eberlin, Petroleomics *via* Orbitrap mass spectrometry with resolving power above 1 000 000 at *m* / *z* 200, RSC Advances. 8 (2018) 6183–6191. https://doi.org/10.1039/C7RA12509G.

[32] C.C. Gu, B. Lin, P. Yehl, J. Pease, N. Chetwyn, Mass spectrometry in small molecule drug development, Small Molecule Development. (2015) 27.

[33] B. Monégier, Electrospray, Techniques de l'ingénieur. P3350 (1997).

[34] M.M. Houck, J.A. Siegel, Light and Matter, in: Fundamentals of Forensic Science, Elsevier, 2015: pp. 93–119. https://doi.org/10.1016/B978-0-12-800037-3.00005-4.

[35] V. Krevelen, Graphical-statistical method for the study of structure and reaction processes of coal, Fuel. 29 (1950) 269–284.

[36] L.R. Snyder, J.J. Kirkland, J.W. Dolan, Introduction to modern liquid chromatography, 3rd ed, Wiley, Hoboken, N.J, 2010.

[37] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics. 11 (2010) 395. https://doi.org/10.1186/1471-2105-11-395.

[38] C. Wolters, L. Flandinet, C. He, J. Isa, F. Orthous-Daunay, R. Thissen, S. Hörst, V. Vuitton, Enhancing data acquisition for the analysis of complex organic matter in direct-infusion Orbitrap mass spectrometry using micro-scans, Rapid Communications in Mass Spectrometry. 34 (2020). https://doi.org/10.1002/rcm.8818.

[39] T. Zhang, D.J. Creek, M.P. Barrett, G. Blackburn, D.G. Watson, Evaluation of Coupling Reversed Phase, Aqueous Normal Phase, and Hydrophilic Interaction Liquid Chromatography with Orbitrap Mass Spectrometry for Metabolomic Studies of Human Urine, Analytical Chemistry. 84 (2012) 1994–2001. https://doi.org/10.1021/ac2030738.

[40] T. Gautier, I. Schmitz-Afonso, D. Touboul, C. Szopa, A. Buch, N. Carrasco, Development of HPLC-Orbitrap method for identification of N-bearing molecules in complex organic material relevant to planetary environments, Icarus. 275 (2016) 259–266. https://doi.org/10.1016/j.icarus.2016.03.007.

[41] HILIC Separations - A Practical Guide to HILIC Mechanisms, Method Development and Troubleshooting, (2014).

[42] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.-L. Veuthey, Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments, European Journal of Pharmaceutics and Biopharmaceutics. 68 (2008) 430–440. https://doi.org/10.1016/j.ejpb.2007.06.018.

[43] D.V. McCalley, Is hydrophilic interaction chromatography with silica columns a viable alternative to reversed-phase liquid chromatography for the analysis of ionisable

compounds?, Journal of Chromatography A. 1171 (2007) 46–55. https://doi.org/10.1016/j.chroma.2007.09.047.

[44] L.G. Gagliardi, C.B. Castells, C. Ràfols, M. Rosés, E. Bosch, δ Conversion Parameter between pH Scales in Acetonitrile/Water Mixtures at Various Compositions and Temperatures, Analytical Chemistry. 79 (2007) 3180–3187. https://doi.org/10.1021/ac062372h.

[45] D.J. Creek, A. Jankevics, R. Breitling, D.G. Watson, M.P. Barrett, K.E.V. Burgess, Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction, Analytical Chemistry. 83 (2011) 8703–8710. https://doi.org/10.1021/ac2021823.

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

[47] B. Eddhif, A. Allavena, S. Liu, T. Ribette, N. Abou Mrad, T. Chiavassa, L.L.S. d'Hendecourt, R. Sternberg, G. Danger, C. Geffroy-Rodier, P. Poinot, Development of liquid chromatography high resolution mass spectrometry strategies for the screening of complex organic matter: Application to astrophysical simulated materials, Talanta. 179 (2018) 238–245. https://doi.org/10.1016/j.talanta.2017.11.008.

[48] J. Hoshen, R. Kopelman, Percolation and cluster distribution. I. Cluster multiple labeling technique and critical concentration algorithm, Physical Review B. 14 (1976) 3438–3445. https://doi.org/10.1103/PhysRevB.14.3438.

[49] Eli. Grushka, Characterization of exponentially modified Gaussian peaks in chromatography, Analytical Chemistry. 44 (1972) 1733–1738. https://doi.org/10.1021/ac60319a011.

[50] Y. Kalambet, Y. Kozmin, K. Mikhailova, I. Nagaev, P. Tikhonov, Reconstruction of chromatographic peaks using the exponentially modified Gaussian function, Journal of Chemometrics. 25 (2011) 352–356. https://doi.org/10.1002/cem.1343.

[51] A. Verri, C. Uras, P. Frosini, M. Ferri, On the use of size functions for shape analysis, Biological Cybermetics. 70 (1993) 99–107.

[52] C. He, S.M. Hörst, N.K. Lewis, X. Yu, J.I. Moses, E.M.-R. Kempton, M.S. Marley, P. McGuiggan, C.V. Morley, J.A. Valenti, V. Vuitton, Photochemical Haze Formation in the Atmospheres of Super-Earths and Mini-Neptunes, The Astronomical Journal. 156 (2018) 38. https://doi.org/10.3847/1538-3881/aac883.

[53] S.M. Hörst, Post-Cassini Investigations of Titan Atmospheric Chemistry, University of Arizona, 2011.

[54] Á. Somogyi, M.A. Smith, V. Vuitton, R. Thissen, I. Komáromi, Chemical ionization in the atmosphere? A model study on negatively charged "exotic" ions generated from Titan's tholins by ultrahigh resolution MS and MS/MS, International Journal of Mass Spectrometry. 316–318 (2012) 157–163. https://doi.org/10.1016/j.ijms.2012.02.026.

[55] S. Pizzarello, G.W. Cooper, G.J. Flynn, The Nature and Distribution of the Organic Material in Carbonaceous Chondrites and Interplanetary Dust Particles, (n.d.) 27.

ANNEXES

Annexe I. Publications

a. Liste des publications dans les revues à comité de lecture

- **CL01** : F.-R. Orthous-Daunay, L. Piani, L. Flandinet, R. Thissen, **C. Wolters**, V. Vuitton, O. Poch, F. Moynier, I. Sugawara, H. Naraoka, S. Tochibana, " Ultraviolet-photon fingerprints on chondritic large organic molecules", [Geochemical Journal] Publié (2019) ; doi:10.2343/geochemj.2.0544

Participation au traitement des données.

- **CL02** : L. Jovanović, T. Gautier, V. Vuitton, **C. Wolters**, J. Bourgalais, A. Buch, F.-R. Orthous-Daunay, L. Vettier, L. Flandinet, N. Carrasco, " Chemical composition of Pluto aerosol analogues ", [Icarus] Publié (2020) ; doi:10.1016/j.icarus.2020.113774

Participation au traitement des données et discussions sur la rédaction de l'article.

- **CL03** : SE Moran, SM Hörst, V Vuitton, C He, N K Lewis, L Flandinet, JI Moses, N North, FR Orthous-Daunay, J Sebree, **C Wolters**, EMR Kempton, MS Marley, CV Morley, JA Valenti, "Chemistry of Temperate Super-Earth and Mini-Neptune Atmospheric Hazes from Laboratory Experiments", [The Planetary Science Journal] Publié (2020) ; doi:10.3847/PSJ/ab8eae

Participation à l'acquisition des données.

- **CL04** : S Potin, S Manigand, P Beck, **C Wolters**, B Schmitt, "A model of the 3-µm hydration band with Exponentially Modified Gaussian (EMG) profiles: Application to hydrated chondrites and asteroids", [Icarus] Publié (2020) ; doi: 10.1016/j.icarus.2020.113686

Proposition de l'utilisation du modèle EMG et de comment le mettre en place.

- **CL05** : **C Wolters**, L Flandinet, C He, J Isa, FR Orthous-Daunay, R Thissen, S Hörst, V Vuitton, "Enhancing data acquisition for the analysis of complex organic matter in direct-infusion Orbitrap mass spectrometry by using micro-scans", [Rapid Communications in Mass Spectrometry] Publié (2020) ; doi:10.1002/rcm.8818

Définition des plans d'expériences, traitement des données intégral, rédaction de l'article, revue des commentaires des co-auteurs, gestion du processus de soumission et de revue des éditeurs et reviewers.

- **CL06** : V Vuitton, SE Moran, C He, **C Wolters**, L Flandinet, FR Orthous-Daunay, JI Moses, JA Valenti, NK Lewis, SM Hörst, "H2SO4 and organosulfur compounds in laboratory analogue aerosols of warm high metallicity exoplanet atmospheres", [Planetary Science Journal] Accepté (2020)

Participation au traitement des données, propositions de représentations, revue et commentaires sur l'article.

- **CL07** : RG Urso, V Vuitton, G Danger, LLS d'Hendecourt, L Flandinet, Z Djouadi, O Mivumbi, FR Orthous-Daunay, A Ruf, V Vinogradof, **C Wolters**, R Brunetto, "Irradiation dose affects the composition of organic refractory materials in space: results from laboratory analogues"; [Astronomy & Astrophysics] Publié (2020) ; doi:10.1051/0004-6361/202039528

Participation à l'acquisition des données, commentaires sur l'article.

b. Autres publications et communications :

- **Autre01** : S. E. Moran, S. Hörst, C. He, L. Flandinet, J. I. Moses, F.-R. Orthous-Daunay, V. Vuitton, **C. Wolters**, N. Lewis, "Laboratory Studies of Planetary Hazes: composition of cool exoplanet atmospheric aerosols with very high-resolution mass spectrometry", [Communication - 49th Meeting of the Division for Planetary Sciences of the American Astronomical Society] (2017)
- **Autre02** : **C. Wolters**, V. Vuitton, F.-R. Orthous-daunay, L. Flandinet, U. Meierhenrich, P. Poinot, G. Danger, "Sélection, mise en place et validation de méthodes HPLC-HRMS en vue de l'analyse d'analogues de glaces cométaires", [Communication - Exobiologie Jeunes Chercheurs 2017] (2017).
- **Autre03** : **C. Wolters**, V. Vuitton, F.-R. Orthous-daunay, L. Flandinet, G. Danger, "Caractérisation par HPLC-HRMS de molécules biologiques dans des matériaux organiques complexes d'intérêt pour la planétologie", [Communication – Spectrométrie de masse à transformée de Fourier (FT-ICR et Orbitrap), École Thématique du CNRS] (2018)
- **Autre04** : F.-R. Orthous-Daunay, **C. Wolters**, L. Flandinet, V. Vuitton, F. Moynier, D. Voisin, M. Kuga, S. Horst, L. Bonal, G. Danger, L. Piani, S. Tachibana, R. Thissen, "Molecular growth in the solar system", [Communication – 3S Sapporo] (2018)
- **Autre05** : S. E. Moran, S. M. Hörst, N. K. Lewis, N. E. Batalha, N. Bishop, C. He, L. Flandinet, J. I. Moses, F.-R. Orthous-Daunay, J. Sebree, V. Vuitton, **C. Wolters**, "Super-Earth and Mini-Neptune laboratory haze analogues and their effects on exoplanetary atmospheric modeling", [Communication – Cambridge] (2018)
- **Autre06** : V. Vuitton, C. He, S. Moran, **C. Wolters**, L. Flandinet, F.-R. Orthous-Daunay, R. Thissen, S. Horst, "Titan's Oxygen Chemistry and its Impact on Haze Formation", [Communication – AAS] (2018)
- **Autre07** : **C. Wolters**, V. Vuitton, F.-R. Orthous-daunay, L. Flandinet, G. Danger, "Caractérisation par HPLC-HRMS de molécules biologiques dans des matériaux organiques complexes d'intérêt pour la planétologie", [Poster – Rencontres de la SFE] (2018)
- **Autre08** : L. Jovanovic, G. Thomas, N. Carasco, **C. Wolters**, L. Flandinet, F.-R. Orthous-Daunay, V. Vuitton, " Pluton, un exemple de chimie organique oxygénée", [Communication – Rencontres de la SFE] (2018)
- **Autre09** : F.-R. Orthous-daunay, **C. Wolters**, L. Flandinet, V. Vuitton, P. Beck, L. Bonal, J. Isa, F. Moynier, D. Voisin, S. Moran, S. Horst, G. Danger, V. Vinogradoff, L. Piani, DV. Bekaert, L. Tissandier, Y. Isono, S. Tachibana, H. Naraoka, L. Remusat, R. Thissen, " Comparison of Molecular Complexity Between Chondrites, Martian Meteorite and Lunar Soils", [Communication - 82nd Annual Meeting of The Meteoritical Society, LPI Contribution No. 2157] (2019)

- **Autre10 :** **C. Wolters**, V. Vuitton, F.-R. Orthous-Daunay, L. Flandinet, C. He, S. Moran, S. Horst, DV. Bekaert, L. Tissandier, B. Marty, L. Piani, "", [Communication - 82nd Annual Meeting of The Meteoritical Society, 2157] (2019)
- **Autre11 :** L. Jovanovic, T. Gautier, N. Carrasco, V. Vuitton, E Quirico, ***C* Wolters**, F.-R. Orthous-Daunay, L. Vettier, L Flandinet, "Laboratory Simulation of Pluto's Atmosphere and Aerosols", [Journal Letters, 848-L5] (2019)
- **Autre12 :** **C. Wolters**, V. Vuitton, L. Flandinet, S. Moran, C. He, H. Ayoub, F.-R. Orthous-Daunay, S. Hörst, "Orbitrap mass spectrometry of synthetic exoplanetary particles", [Communication - EPSC-DPS2019-2029] (2019)
- **Autre13 :** SE Moran, S Horst, V Vuitton, C He, N Lewis, L Flandinet, J Moses, F Orthous-Daunay, J Sebree, ***C* Wolters**, "Chemistry of Temperate Exoplanet Hazes from the Laboratory", [Communication – AAS 248.04] (2020)
- **Autre14 :** SE Moran, SM Hörst, V Vuitton, C He, NK Lewis, N Bishop, L Flandinet, JI Moses, F-R Orthous-Daunay, J Sebree, ***C* Wolters**, "Chemistry of Laboratory Exoplanet Hazes", [Communication – LPICo, 2195, p3030] (2020)
- **Autre15 :** F-R Orthous-Daunay, ***C* Wolters**, V Vuitton, J Isa, H Naraoka, R Thissen, "Orbitrap-MS and Chromatography in Preparation for Hayabusa2 Molecular Complexity Analyses", [Communication – LPICo, 2326, p2551] (2020)

Annexe II. Analyse statistique pour les modèles de prédiction des temps de rétention

Méthode acide

Dans l'idée d'obtenir un modèle acceptable, une étude systématique est réalisée. Pour commencer, on considère un modèle strictement linéaire impliquant l'ensemble des paramètres physico-chimiques disponibles, ainsi qu'une constante :

- Coefficient 1 : logP
- Coefficient 2 : logD à pH 3,5
- Coefficient 3 : charge nette positive à pH 3,5
- Coefficient 4 : charge nette négative à pH 3,5
- Coefficient 5 : nombre de liaisons autorisant une rotation
- Coefficient 6 : masse exacte
- Coefficient 7 : nombre d'atomes acceptant des liaisons hydrogènes
- Coefficient 8 : nombre d'atomes des donnant liaisons hydrogènes
- Coefficient 9 : constante

Ce premier modèle, disponible en Figure 90, bien que démontrant une qualité de régression acceptable, présente un nombre important de paramètres ayant des probabilités (*p)* élevés, non-significatifs.

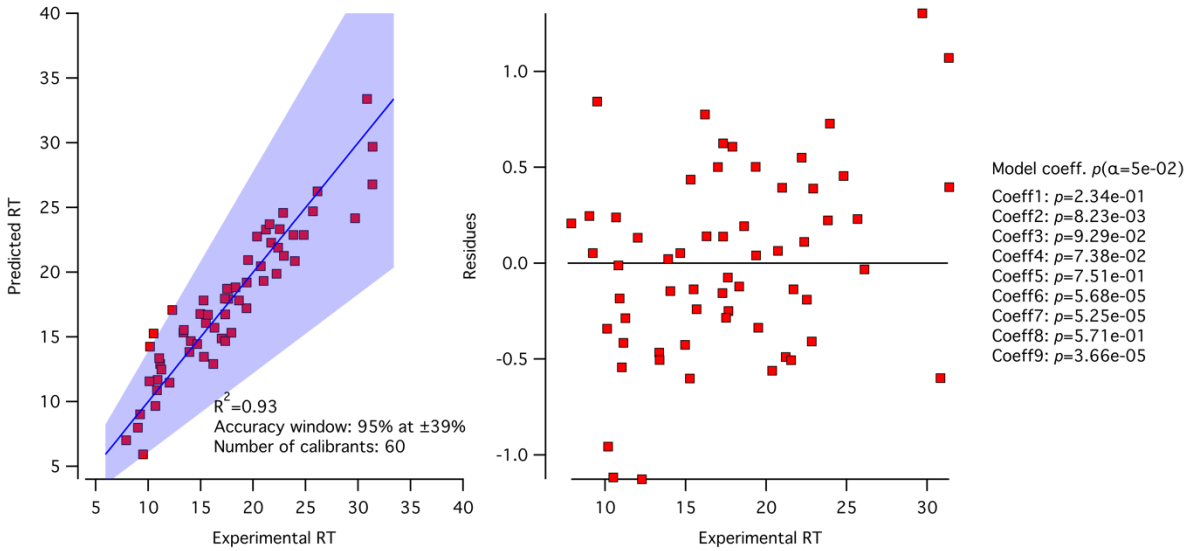


Figure 90 – Modèle linéaire à 8 coefficients et une constante pour la méthode acide.

Pour tenter de trouver le meilleur modèle, on retire le plus mauvais coefficient, *i.e.* le coefficient ayant le *p* le plus grand, et on itère jusqu'à obtenir un modèle dont tous les coefficients sont statistiquement significatifs. Cette itération, non présentée ici, nous conduit à un modèle à trois paramètres présentés en Figure 91 :

- Coefficient 1 : logD à pH 3,5
- Coefficient 2 : charge nette négative à pH 3,5
- Coefficient 3 : constante

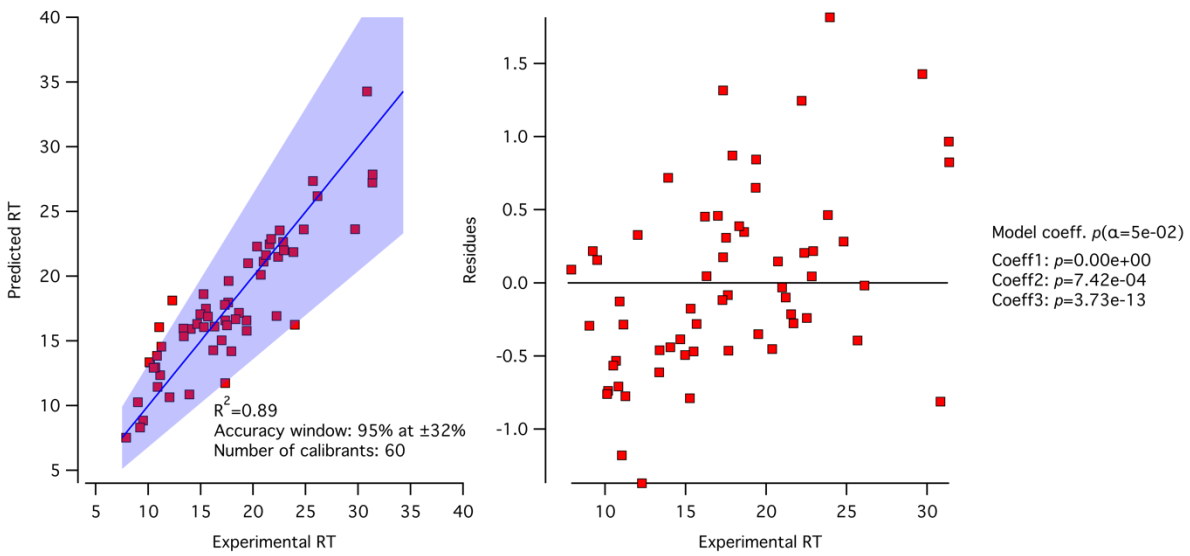


Figure 91 – Modèle linéaire à 2 coefficients et une constante pour la méthode acide. Les p indiqués à 0 signifient que leur valeur est inférieure à 10^{-16}, non qu'ils soient strictement nuls.

Le fait que le meilleur modèle soit basé sur le logD et la charge nette négative au pH de travail est révélateur des interactions majoritaires qui se produisent au niveau de la colonne. Ainsi, le coefficient de partage et les interactions ioniques semblent privilégiées sur cette colonne et à ce pH.

Même si ce modèle est satisfaisant d'un point de vue statistique, on se doit d'ajouter la connaissance de ce que l'on cherche à modéliser. Ainsi, on note qu'il manque un type d'interaction dans le modèle : les interactions de type liaisons hydrogènes. Il convient donc de les ajouter au modèle pour ne pas négliger des effets qui peuvent être important pour certaines molécules. Des tests sur des échantillons complexes indiquent également une distribution hyperbolique des temps de rétention avec la masse, pouvant indiquer que la masse est influente avec un terme de type hyperbolique. Ces ajouts nous conduisent à un modèle à 6 coefficients, montré en Figure 92 :

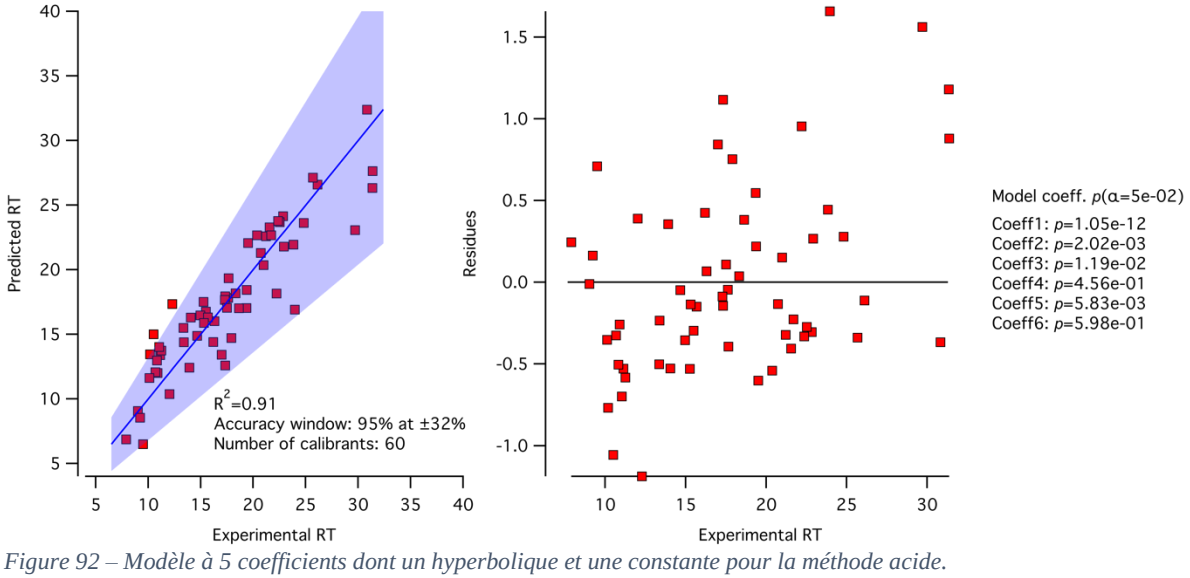


Figure 92 – Modèle à 5 coefficients dont un hyperbolique et une constante pour la méthode acide.

Les coefficients retenus pour ce modèle sont les suivants :

- Coefficient 1 : logD à pH 3,5
- Coefficient 2 : charge nette négative à pH 3,5
- Coefficient 3 : nombre d'atomes donnant des liaisons hydrogènes

- Coefficient 4 : nombre d'atomes acceptant des liaisons hydrogènes
- Coefficient 5 : 1 ÷ masse exacte
- Coefficient 6 : constante

On notera également, entre ce dernier modèle et celui à trois coefficients, que la qualité de la régression a augmentée (R^2 est passé de 0,89 à 0,91) tout en conservant une fenêtre de précision égale. Ce modèle possède également des coefficients que l'on pourrait considérer non-significatif, mais le sens physique du modèle a plus d'importance dans ce cas. Ce modèle est donc celui retenu comme représentant au mieux les interactions se déroulant au sein de la colonne, et est utilisé par la suite pour effectuer des prédictions de composés inconnus.

Méthode basique

De la même manière que pour la méthode acide, une étude statistique systématique est effectuée dans le but d'obtenir un modèle statistiquement acceptable. Pour commencer, on considère un modèle strictement linéaire impliquant l'ensemble des paramètres physico-chimiques disponibles, ainsi qu'une constante :

- Coefficient 1 : logP
- Coefficient 2 : logD à pH 9
- Coefficient 3 : charge nette positive à pH 9
- Coefficient 4 : charge nette négative à pH 9
- Coefficient 5 : nombre de liaisons autorisant une rotation
- Coefficient 6 : masse exacte
- Coefficient 7 : nombre d'atomes donnant des liaisons hydrogènes
- Coefficient 8 : nombre d'atomes acceptant des liaisons hydrogènes
- Coefficient 9 : constante

Ce premier modèle, disponible en Figure 93, bien que démontrant une qualité de régression acceptable, présente un nombre important de paramètres ayant des *p* élevés, non-significatifs.

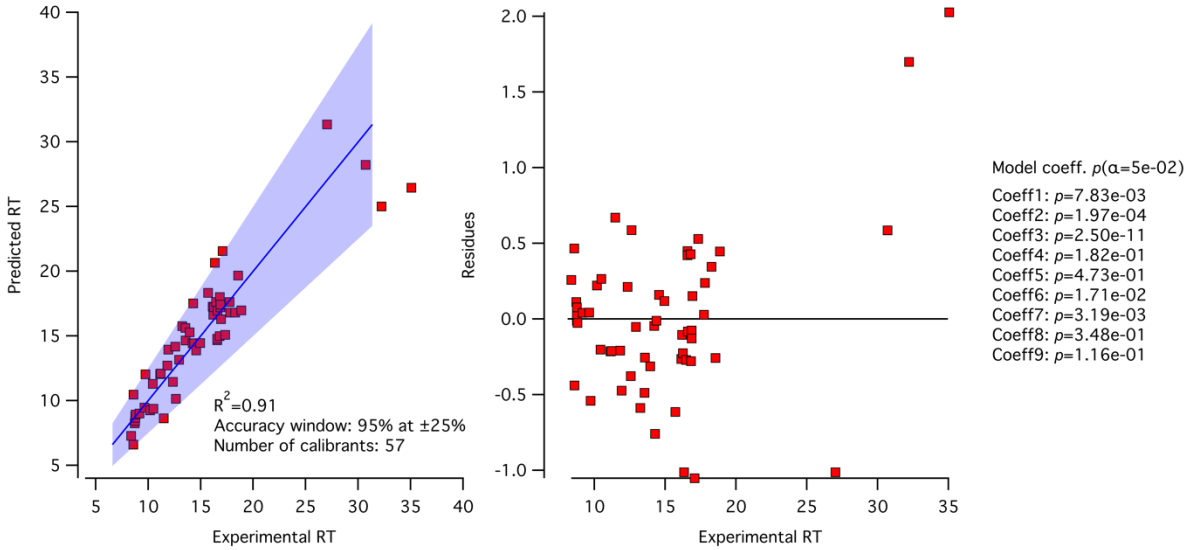


Figure 93 – Modèle linéaire à 8 coefficients et une constante pour la méthode basique.

Pour tenter de trouver le meilleur modèle, on retire le plus mauvais coefficient, *i.e.* le coefficient ayant le *p* le plus grand, et on itère jusqu'à obtenir un modèle dont tous les coefficients sont statistiquement significatifs. Cette itération, non présentée ici, nous conduit à un modèle à six paramètres présentés en Figure 94 :

- Coefficient 1 : logP
- Coefficient 2 : logD à pH 9
- Coefficient 3 : charge nette positive à pH 9

- Coefficient 4 : masse exacte
- Coefficient 5 : nombre d'atomes acceptant des liaisons hydrogènes
- Coefficient 6 : constante

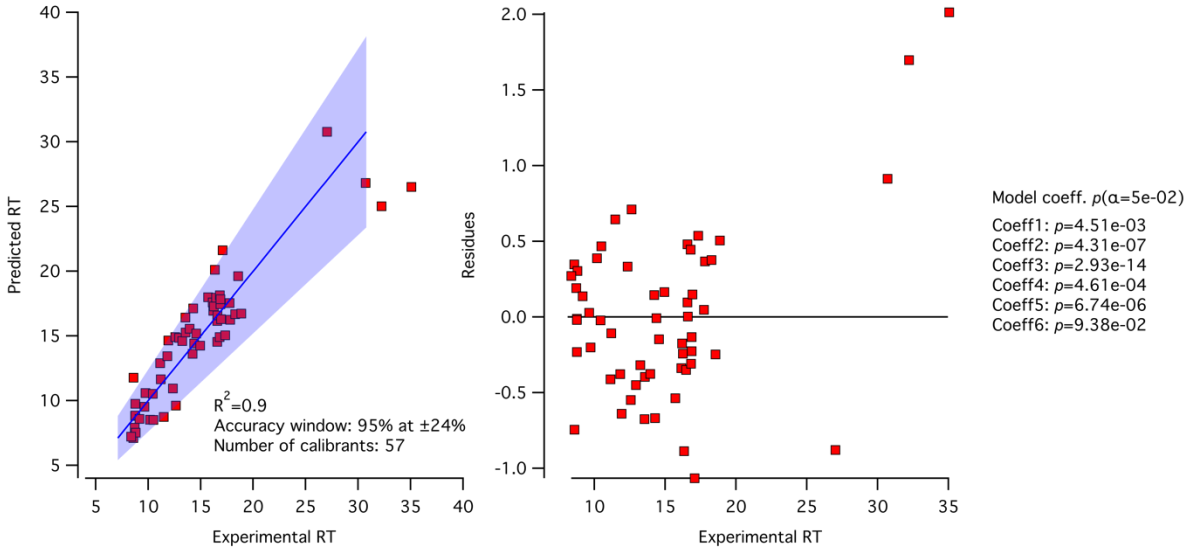


Figure 94 – Modèle linéaire à 5 coefficients et une constante pour la méthode basique.

La significativité de cinq paramètres physico-chimiques à pH basique comparé aux trois à pH acide révèle que les critères de séparation à pH basique sont plus complexes à interpréter qu'à pH acide. On notera également un manque de composés éluant entre 20 et 30 minutes, limitant la gamme de paramètres physico-chimiques pris en compte par ce modèle. Ceci explique également l'erreur plus importante de prédiction des quatre composés très retenus.

De la même manière que pour la méthode acide, on rajoute au modèle la seconde interaction hydrogène et on change la masse exacte en paramètre hyperbolique. Ces ajouts nous conduisent à un modèle à sept coefficients, montré en Figure 95 :

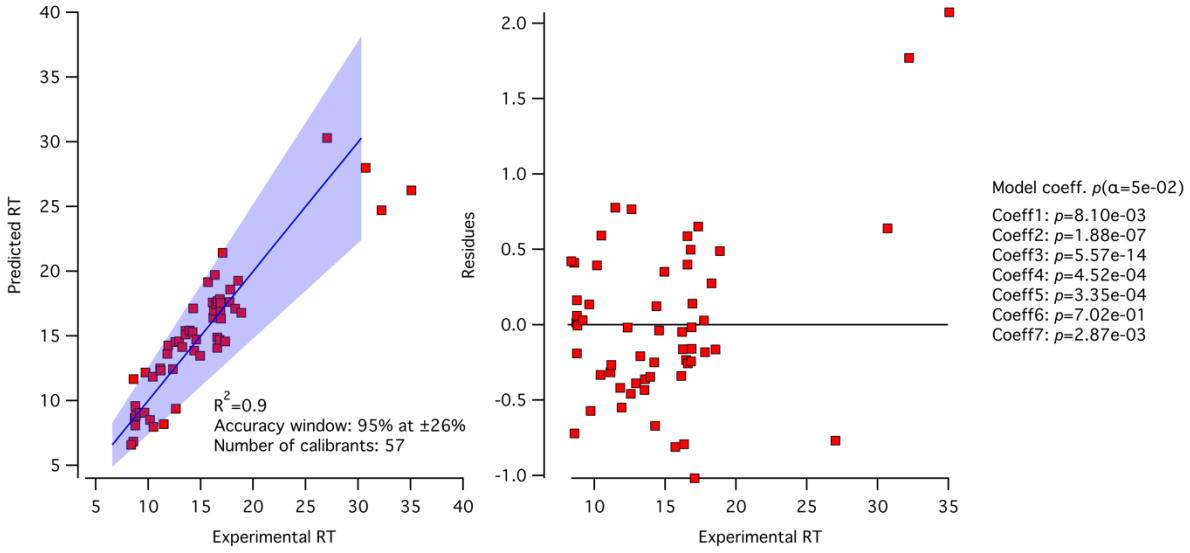


Figure 95 – Modèle à 6 coefficients dont un hyperbolique et une constante pour la méthode basique.

Les coefficients retenus pour ce modèle sont les suivants :

- Coefficient 1 : logP
- Coefficient 2 : logD à pH 9
- Coefficient 3 : charge nette positive à pH 9
- Coefficient 4 : 1 ÷ masse exacte

- Coefficient 5 : nombre d’atomes donnant des liaisons hydrogènes
- Coefficient 6 : nombre d’atomes acceptant des liaisons hydrogènes
- Coefficient 7 : constante

Ce modèle possède également des coefficients que l’on pourrait considérer non-significatif, mais le sens physique du modèle a plus d’importance dans ce cas. Ce modèle est donc celui retenu comme représentant au mieux les interactions se déroulant au sein de la colonne, et est utilisé par la suite pour effectuer des prédictions de composés inconnus.

Annexe III. Résumé des méthodes techniques et instrumentales

Préparation d'échantillon pour l'analyse en spectrométrie de masse

Matériel :

- Portoir à tube
- Tubes de 2mL (verre, polypropylène…)
- Micropipette : 1 000µL
- Coupelle de pesée
- Balance de précision
- Vortex mixeur
- Centrifugeuse

Protocole expérimental :

- Placer la coupelle sur la balance de précision
- Appuyer sur "Tare"
- Peser environ 2 mg (soit sur la balance : 0.00200g)
- Appuyer sur "Tare"
- Placer la poudre dans un tube de 2mL
- Placer la coupelle vide sur la balance et noter la masse
- Ajouter 1 000 µL de méthanol dans le tube avec une micropipette
- Agiter pendant 10 min à l'aide du vortex mixeur (Speed 5). La présence d'une suspension dans la solution est normale si un échantillon complexe est utilisé (l'ensemble n'est pas forcément totalement soluble)
- Centrifugeuse :
    - Placer le tube dans la centrifugeuse en position 1
    - Placer un tube rempli de 1 000 µL de méthanol en position 13 (pour équilibrer les masses)
    - Centrifuger pendant 10 min à 10 000 RPM
- Récupérer 500µL de surnageant à l'aide d'une micropipette, transvaser dans un nouveau tube et ajouter 500µl de méthanol. Agitez.

Note : Cette solution est utilisée directement pour les analyses en spectrométrie de masse. Pour les analyses couplées avec la chromatographie, l'étape finale de dilution n'est pas effectuée.

Méthodes Orbitrap

Instrumentation

- LTQ-Orbitrap-XL, Grenoble
    - Calibration externe avec solution de calibration fournie par Thermo Scientific
        - CalMix Positif : caféine, L-methionine-arginylphenylalanyl-alanine et Ultramak 1621
        - CalMix Négatif : sodium dodecyl sulfate, sodium taurocholate et Ultramark 1621

En ESI-Orbitrap

Polarité	Positif	Négatif
Position de la source	B 1	C 1,5
Débit de liquide (µl/min)	1 à 10	1 à 10
Débit de gaz Sheath (arb)	3	5
Débit de gaz Aux (arb)	0	1
Débit de gaz Sweep (arb)	0	0
Tension de spray (kV)	3,5	-3,5

Température du tube de transfert (°C)	275	275
Tension du tube de transfert (V)	40	-50

Tableau 24 – Paramètres en ESI-Orbitrap

Le paramètre de « Tube Lens » est dépendant de la gamme de masse choisie et n'est donc pas renseigné ici. Le nombre de scans et de micro-scans est dépendant de l'analyse et est renseigné lorsque nécessaire. Les tensions appliquées sur les optiques de transfert ioniques sont héritées des méthodes de calibrations et ne sont pas modifiées.

En couplage HPLC-Orbitrap

Polarité	**Positif**	**Négatif**
Position de la source	C 1	C 1
Débit de liquide (µl/min)	Imposé par HPLC (275µl/min)	Imposé par HPLC (275µl/min)
Débit de gaz Sheath (arb)	50	60
Débit de gaz Aux (arb)	15	30
Débit de gaz Sweep (arb)	0	0
Tension de spray (kV)	3,5	3,0
Température du tube de transfert (°C)	300	350
Tension du tube de transfert (V)	35	35

Tableau 25 – Paramètres en couplage HPLC-Orbitrap

Le paramètre de « Tube Lens » est dépendant de la gamme de masse choisie et n'est donc pas renseigné ici. Les tensions appliquées sur les optiques de transfert ioniques sont héritées des méthodes de calibrations et ne sont pas modifiées.

Analyse d'échantillon en ESI-Orbitrap

Les analyses des échantillons présentés dans ce travail ont été effectuées avant les optimisations des protocoles d'acquisitions, et utilisent une méthode d'acquisition utilisant la moyenne de 4 scans de 128 micro-scans, sur trois gammes de masse différentes : [50-300] Da, [150-450] Da et [400-1000] Da. Ce sont ces analyses qui sont d'ailleurs en partie à l'origine des travaux sur l'optimisation des conditions d'acquisitions des données. Chaque échantillon est également acquis en polarité positive et en polarité négative.

Analyse d'échantillon en HPLC-Orbitrap

Les analyses en HPLC-Orbitrap sont réalisées en segmentant les analyses par gamme de 50Da. Cette segmentation est disponible en partie 4.3.

Méthodes FT-ICR

Instrumentation

- FT-ICR Solarix XR, 12T, Rouen
    - ESI source
    - LDI source (laser NdYAg 355mm)
    - Calibration externe avec une solution de tri-fluoroacétate

Préparation des plaques LDI

- Les fractions solubles, insolubles et totales sont analysées.
- La fraction soluble est préparée en utilisant la méthode des gouttes sèches (5x1µl pour avoir une bonne concentration) [17]
- Les fractions insolubles et totales sont déposées en aplatissant une pointe de spatule sur la plaque, et en retirant le surplus non-cohérant avec la plaque en utilisant un flux d'azote léger.

En ESI-FT-ICR

Polarité	**Positif**	**Négatif**
Débit de liquide (µl/min)	3µl/min	3µl/min

Débit de gaz (L/min)	4	4
Tension de spray (kV)	3,5	-3,0
Funnel 1 (V)	150	-150
Skimmer 1 (V)	10	-10

Tableau 26 – Paramètres en ESI-FT-ICR

La gamme de masse et le temps de stockage sont des paramètres dépendant de l'échantillon et mentionnés ci-nécessaire. Les traitements de données spécifiques au FT-ICR (AMP…) pour générer des spectres ont été effectués à Rouen par du personnel qualifié et les détails sont mentionnés ci-nécessaire sur les spectres.

En LDI-FT-ICR

Polarité	Positif	Négatif
Fréquence du laser (Hz)	1000	1000
Puissance du laser (%)	21	21
Nombre de tirs par spectres	100	20
Tension sur la plaque (V)	100	-100
Tension plaque deflective (V)	150	-200
Tension Funnel 1 (V)	150	-150
Tension Skimmer 1 (V)	25	-25

Tableau 27 – Paramètres en LDI-FT-ICR

Les traitements de données spécifiques au FT-ICR (AMP…) pour générer des spectres ont été effectués à Rouen par du personnel qualifié et les détails sont mentionnés ci-nécessaire sur les spectres.

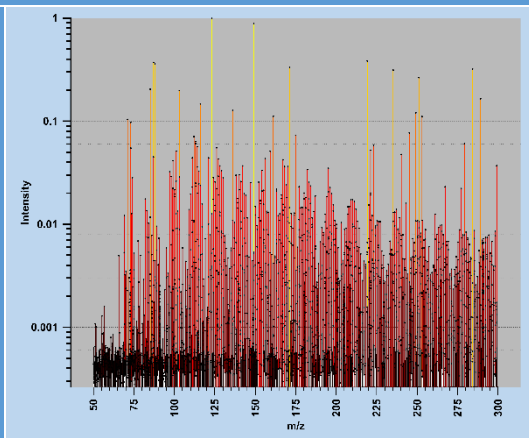
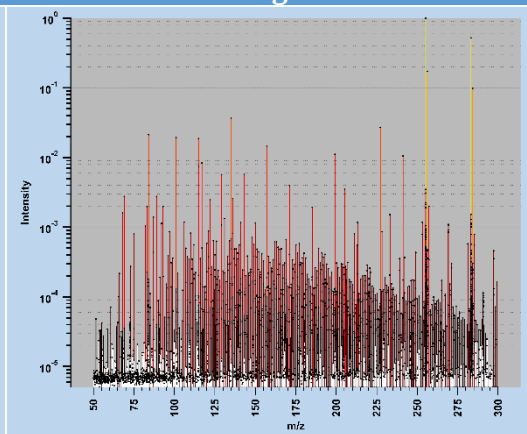
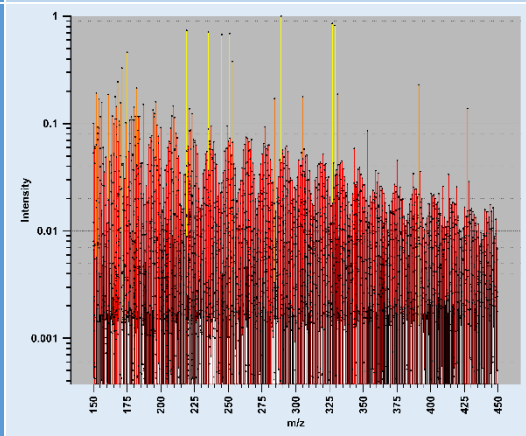
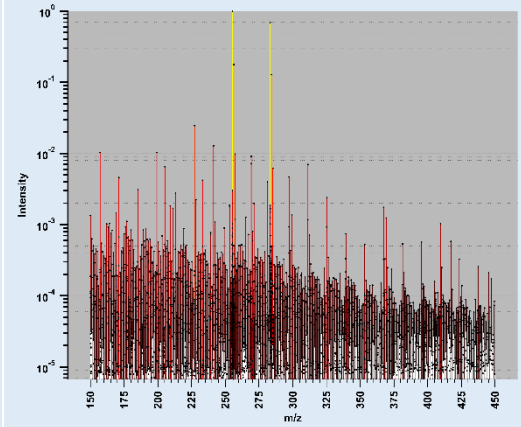
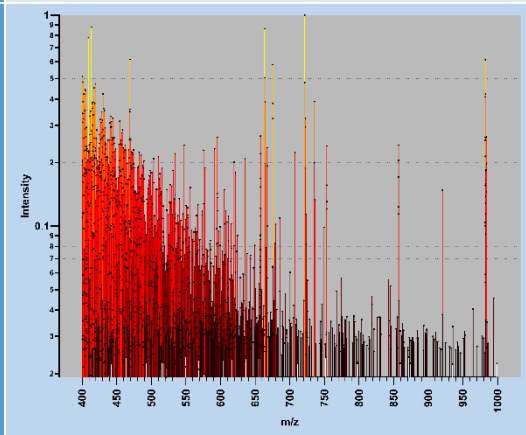
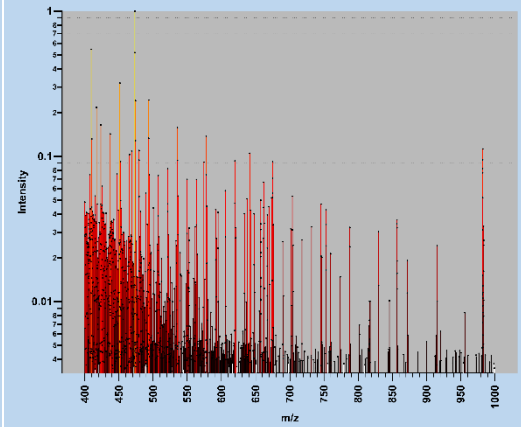
Analyse d'échantillon par ICR

Les échantillons sont analysés sur une gamme de masse unique qui est légèrement en variable en fonction du paramétrage. Par soucis de comparaison, l'ensemble des analyses peuvent être considérées sur la gamme [100-800]Da.

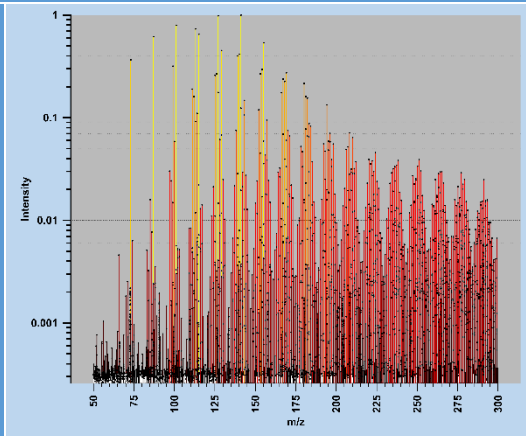
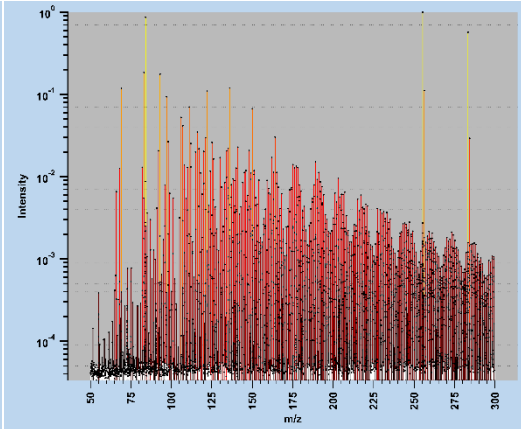
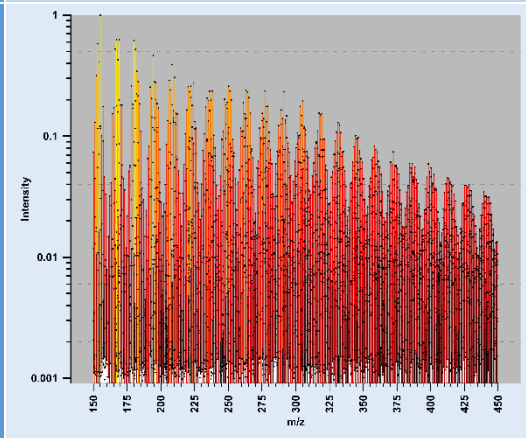
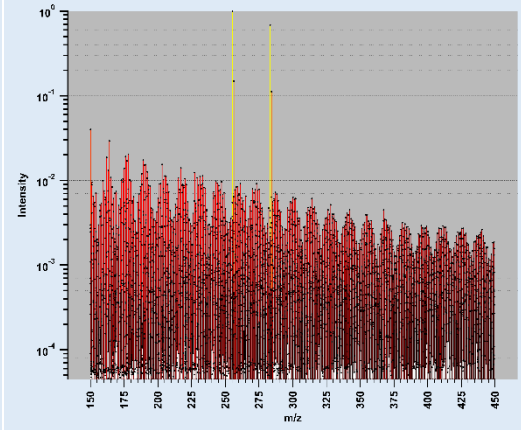
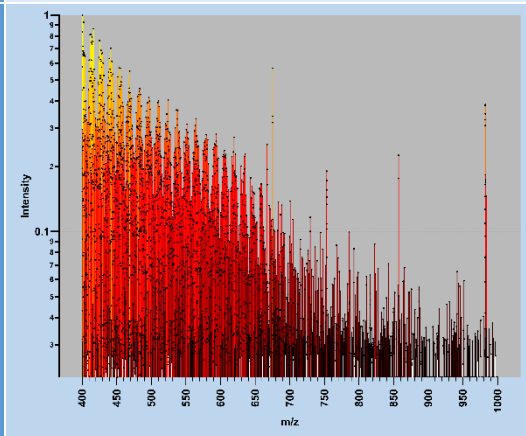
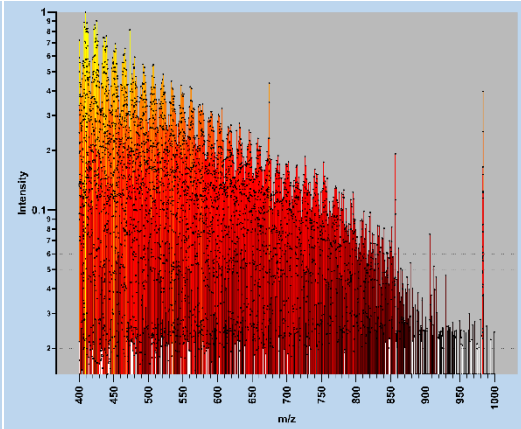
Annexe IV. Ressources graphiques pour les échantillons analysés par ESI-Orbitrap

Par soucis de concision, seuls les spectres sur la gamme [150-450] Da sont discutés dans le corps du texte. L'ensemble des données graphiques concernant les données acquises sont présentées ci-après.

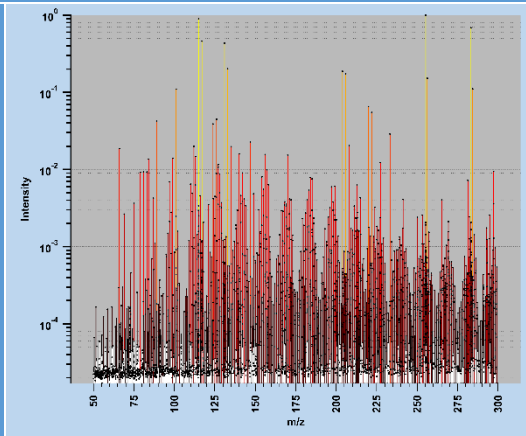
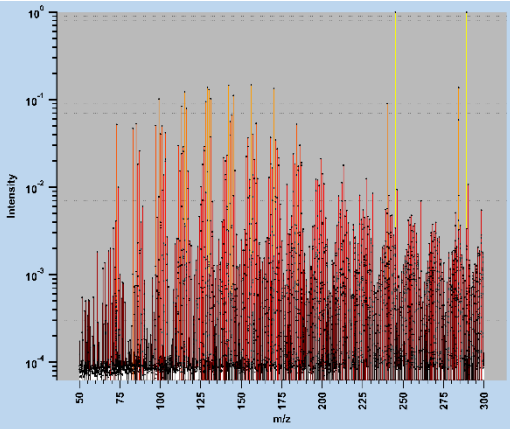
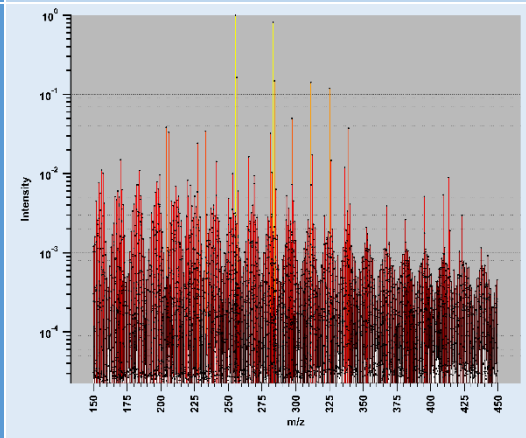
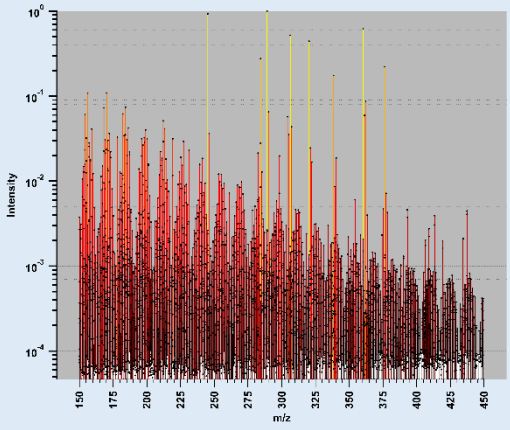
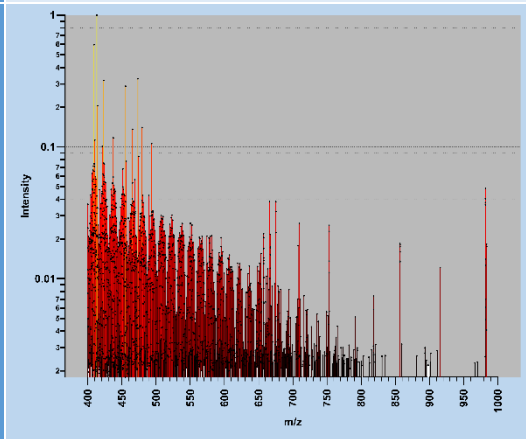
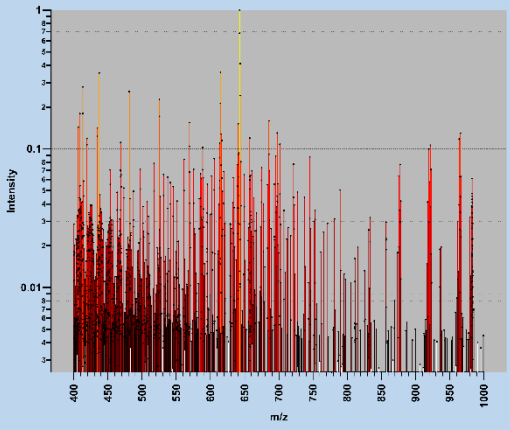
300K - Spectres

	Positif	Négatif
50-300		
150-450		
400-1000		

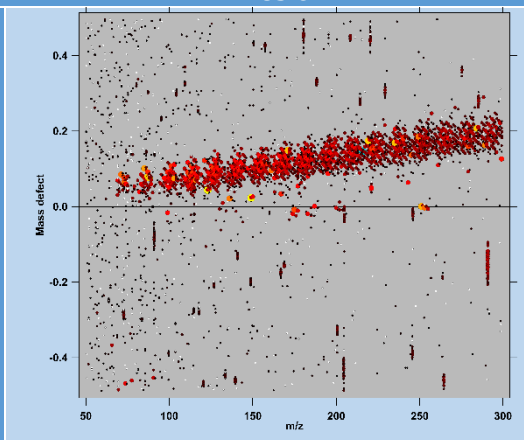
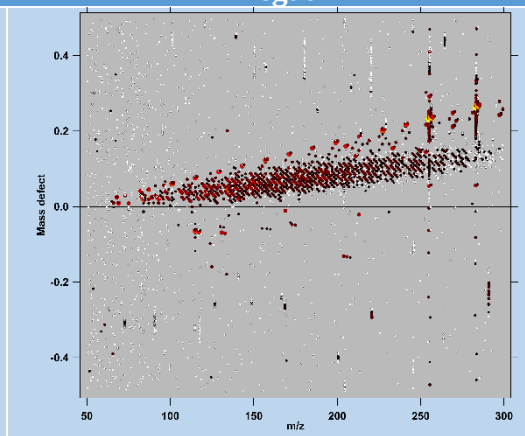
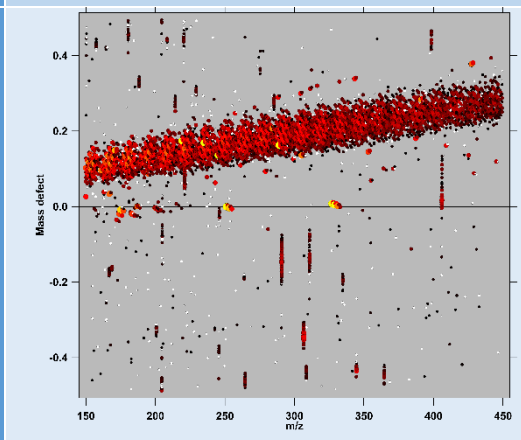
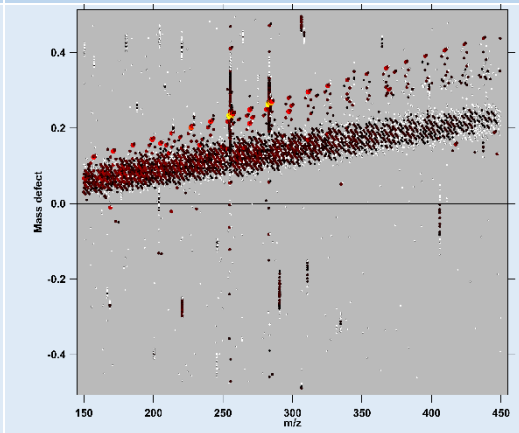
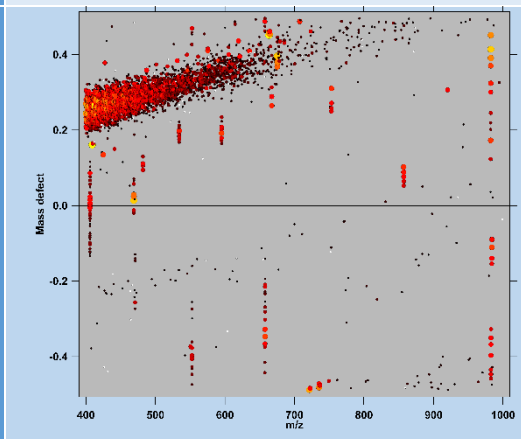
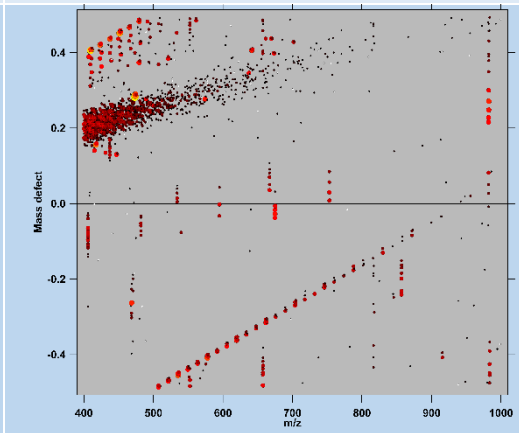
400K - Spectres

	Positif	Négatif
50-300		
150-450		
400-1000		

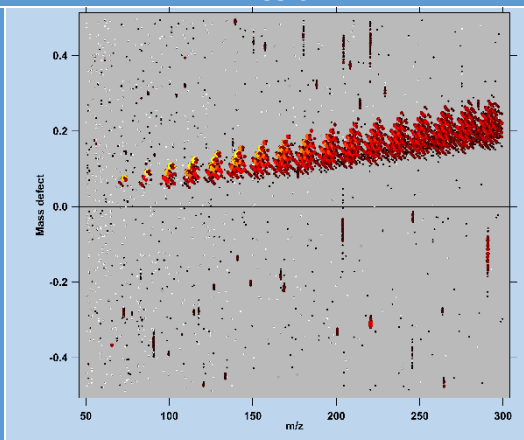
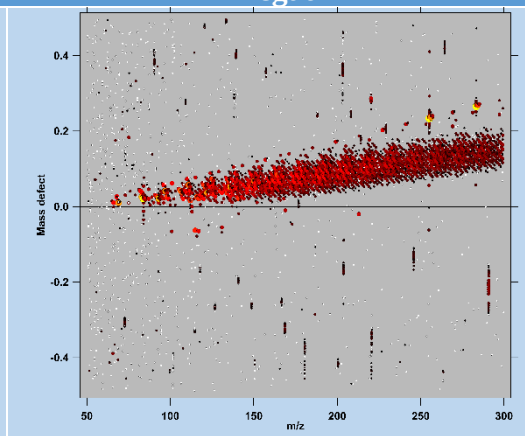
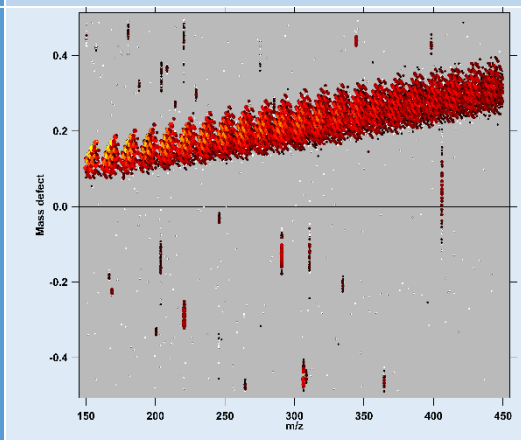
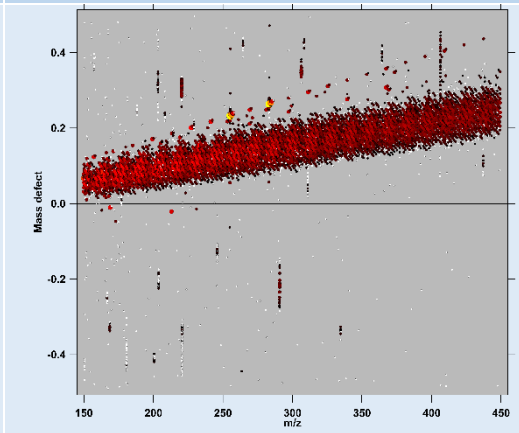
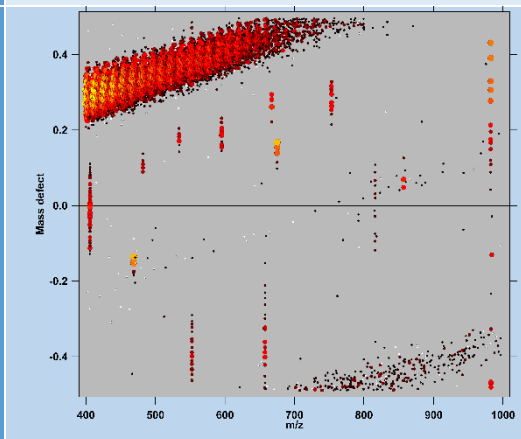
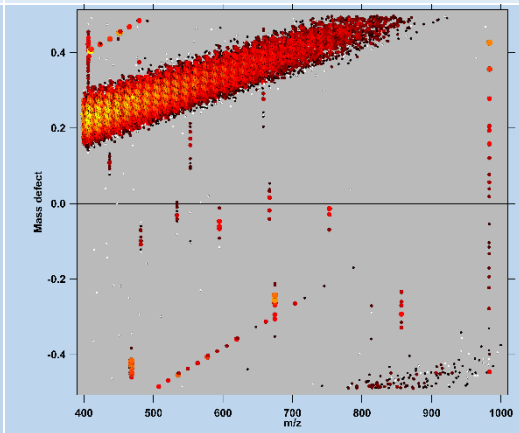
600K - Spectres

	Positif	Négatif
50-300		
150-450		
400-1000		

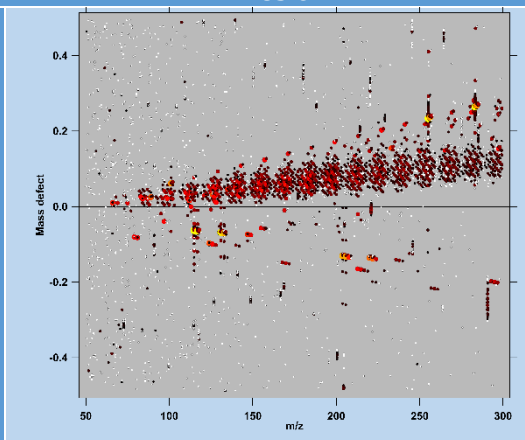
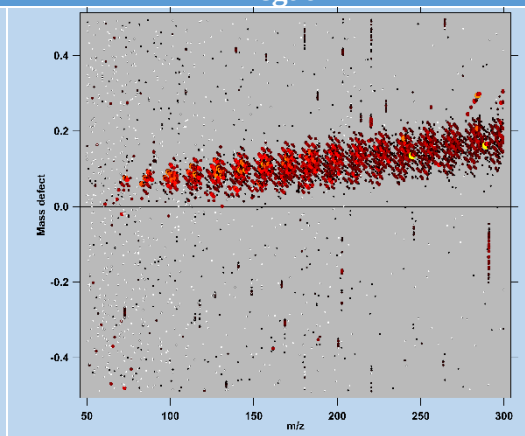
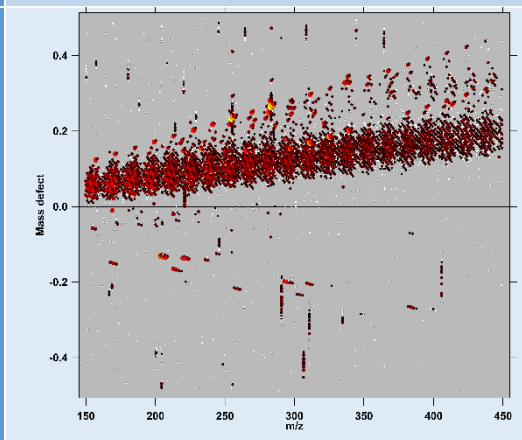
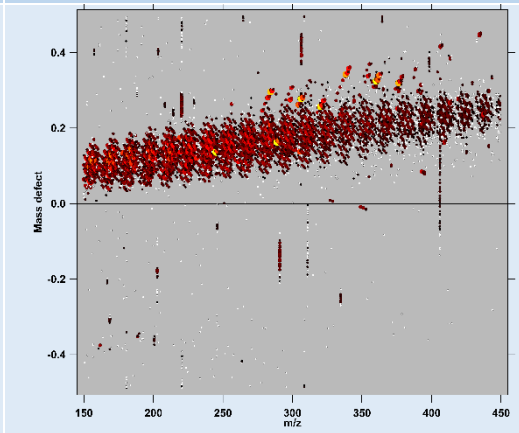
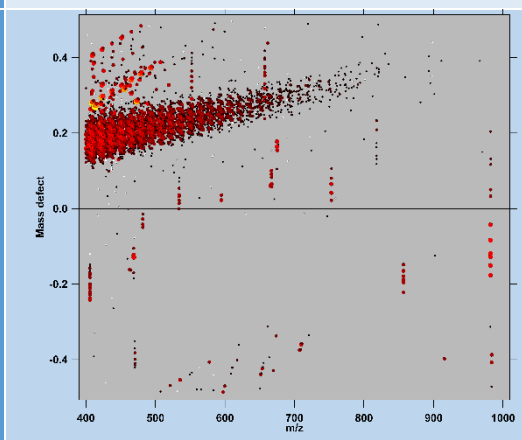
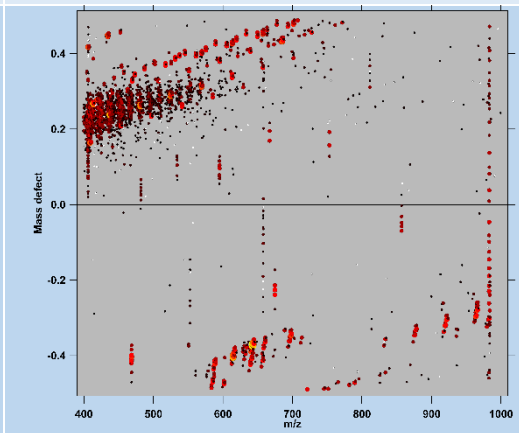
300K – Défaut de masse vs Masse

	Positif	Négatif
50-300		
150-450		
400-1000		

400K - Défaut de masse vs Masse

	Positif	Négatif
50-300		
150-450		
400-1000		

600K - Défaut de masse vs Masse

	Positif	Négatif
50-300		
150-450		
400-1000		

SIGILLUM UNIVERSITATIS GRATIANOPOLITANAE
VERI TAS
LIBE RABIT
1339