UNIVERSITÉ LAVAL

Développement d'outils computationnels pour une approche de métabolomique non ciblée par spectrométrie de masse à haut débit

Thèse

Pier-Luc Plante

Doctorat en médecine moléculaire
Philosophiæ doctor (Ph. D.)

Québec, Canada

Développement d'outils computationnels pour une approche de métabolomique non ciblée par spectrométrie de masse à haut débit

Thèse

Pier-Luc Plante

Sous la direction de :

Jacques Corbeil, directeur de recherche

Mario Marchand, codirecteur de recherche

Résumé

La métabolomique est l'étude des petites molécules produites par un système biologique. L'objectif principal des études en métabolomique non ciblées est la recherche d'une signature moléculaire, à base de biomarqueurs, permettant de distinguer deux phénotypes (ex. : malade et sain). Elle trouve des applications dans le domaine de la santé, de la nutrition, de l'agroalimentation et même de l'environnement. La spectrométrie de masse couplée à la chromatographie liquide est une des techniques les plus utilisées puisqu'elle offre sensibilité et spécificité lors de l'étude du métabolome. Par contre, le long temps d'analyse limite la taille et la portée des études métabolomiques. De nouvelles approches de métabolomique non ciblée à haut débit par spectrométrie de masse où un échantillon peut être analysé en quelques secondes peuvent cependant éliminer cette barrière. Ce changement de paradigme entraîne une complexification des différentes étapes de l'analyse de données (prétraitement, recherche de biomarqueurs et identification des métabolites).

Dans le cadre de cette thèse, nous proposons différents outils basés sur l'apprentissage automatique visant à résoudre les problèmes d'analyse de données causés par une accélération de la vitesse d'acquisition et une augmentation du nombre d'échantillons. Premièrement, nous proposons une série d'algorithmes de correction et d'alignement de spectres de masse visant à les rendre comparables afin de permettre les analyses statistiques et l'apprentissage automatique. Deuxièmement, nous présentons MetaboDashboard, un outil visant à simplifier et à démocratiser l'utilisation de l'apprentissage automatique pour la recherche de biomarqueurs en métabolomique non ciblée. Un exemple de son utilisation dans le contexte d'une infection virale des voies respiratoires est présenté. Finalement, un réseau de neurones appelé DeepCCS permettant la prédiction de la section efficace dans l'objectif de supporter l'identification des métabolites est exposé.

Nous démontrons, tout au long de cette thèse, l'utilité et la puissance de l'apprentissage automatique appliqué à la métabolomique non ciblée. Les outils computationnels présentés dans cette thèse sont le point de départ du développement d'une méthode de métabolomique non ciblée à haut débit. Nous espérons qu'ultimement, les contributions de cette thèse permettront la détection de biomarqueurs associés à différents phénotypes dans des populations entières avec un maximum de précision et à une vitesse encore jamais vue.

Abstract

Metabolomics is defined as the study of small molecules produced by a biological system. The main objective of metabolomic studies is the search of a molecular signature, constituted of biomarkers, that allow to distinguish two phenotypes (ex: sick and healthy). It can be applied to diverse fields such as health, nutrition, food and environment. Mass spectrometry coupled to liquid chromatography is the most common technique used in metabolomics since it offers sensibility and specificity. Unfortunately, the long running time of these analysis limits the size and impact of metabolomic studies. New approaches in high-throughput untargeted metabolomics, where a sample can be analyzed in seconds, try to overcome this limitation. This new paradigm increases the complexity of the different data analysis steps that follows that acquisition (data pre-treatment, biomarker discovery and metabolite identification).

In this thesis, we propose different tools based on machine learning that aim at solving the new data analysis issues that arise from the increased number of samples and throughput. First, we present new algorithms to correct and align mass spectra to make them comparable in order to enable statistical analysis and machine learning. Second, we present *MetaboDashboard*, a tool that aims at simplifying and democratizing the use of machine learning approach for biomarker discovery in the context of untargeted metabolomics. An example of its usage in the context of viral respiratory tract infection is then presented. Finally, a neural network tool called *DeepCCS*, that allow the prediction of collisional cross section for metabolite identification is reported.

Throughout this thesis, we demonstrate the use and impact of machine learning applied to different problems in untargeted metabolomics. The computational tools presented in this thesis are the first steps towards the development of new methods in high-throughput untargeted metabolomics. We hope that ultimately, the scientific contributions presented in this thesis will enable biomarker discovery for different phenotypes at the scale of whole population with a level of precision and speed never seen before.

Table des matières

Liste des figures

Liste des tableaux

Liste des équations

Liste des abréviations, sigles, acronymes

Da	Dalton (Unité de mesure de masse moléculaire)
uma	Unité de masse atomique
m/z	Ratio masse sur charge
DESI	*Desorption Electrospray Ionization*
LDTD	*Laser Diode Thermal Desorption*
MALDI	Désorption laser assistée par une matrice / *Matrix Assisted Laser Desoprtion Ionization*
ESI	Ionisation par électronébuliseur / *Electrospray Ionization*
APCI	Ionisation chimique à pression atmosphérique / *Atmospheric Pressure Chemical Ionization*
ToF	Temps de vol / *Time of Flight*
CCS	Section efficace / *Collisional Cross Section*
ADN	Acide désoxyribonucléique

À ma grenouille…

« *Legacy. What is a legacy?*

It's planting seeds in a garden you never get to see. »

- Lin-Manuel Miranda, *Hamilton*

Remerciements

Je remercie mon directeur de recherche, Pr Jacques Corbeil, de m'avoir fait confiance et de m'avoir toujours incité à viser plus haut. Je remercie également mon co-directeur, Pr Mario Marchand et son acolyte Pr François Laviolette qui m'ont introduit à l'apprentissage automatique et qui ont su guider et encadrer mon intuition. Merci au Pr Sylvain Moineau de m'avoir donné la chance de faire mon premier stage en bio-informatique, un point tournant de ma carrière, et pour son support durant toutes ces années.

Je remercie Pr Frédéric Raymond de m'avoir appris à faire de la belle science et à la communiquer ainsi qu'à Nancy Boucher pour sa confiance et son soutien.

Merci à tous mes collègues dans le laboratoire de Pr Corbeil, du GRAAL, du CRI et de l'Université Laval qui ont partagé une partie de cette incroyable aventure avec moi. Merci aussi à l'équipe de Phytronix qui m'a accueilli dans sa belle grande famille.

Je remercie également mes parents, Luc et Claire, ainsi que mes beaux-parents, Line et Claude, pour leurs encouragements et pour avoir montré autant d'intérêt envers mes recherches.

Sarah-Anne, « merci » n'est pas assez pour exprimer toute ma gratitude. Je n'y serais jamais arrivé sans toi…

Avant-propos

La bio-informatique appliquée à la métabolomique est un domaine de recherche à l'intersection de la biologie, de la chimie et de l'informatique. De ce fait, la bio-informatique est une science collaborative qui ne peut être développée sans l'apport précieux de chercheurs dans plusieurs domaines. Durant mes études doctorales, j'ai collaboré avec différents groupes de recherche vers un objectif commun : le développement d'une approche de métabolomique à haut débit employant l'apprentissage automatique. Cette thèse présente trois articles ainsi que d'autres contributions qui pavent la voie vers cet objectif ambitieux.

Publication incluses

MetaboDashboard: simplified machine learning in metabolomics

Pier-Luc Plante, Francis Brière, Nancy Boucher, Élina Francovic-Fontaine, Didier Brassard, Benoit Lamarche, Jacques Corbeil

Statut : Soumis à *Bioinformatics* le 6 décembre 2020

Inclusion : Chapitre 2

Changement pour l'inclusion : Aucun

Contribution des auteurs : P.-L.P a conçu le projet, construit l'outils et a rédigé le manuscrit. F.B., N.B., D.B. et E.F.-F. ont participé à la conception et la construction de l'outil. D.B. et B.L ont fourni des échantillons. B.L. et J.C. ont participé à la conception du projet et à la révision du manuscrit.

Predicting ion mobility collision cross-sections using a deep neural network: DeepCCS

Pier-Luc Plante, Élina Francovic-Fontaine, Jody C May, John A McLean, Erin S Baker, François Laviolette, Mario Marchand, Jacques Corbeil

Analytical chemistry volume 91 (8), 5191-5199 (2019)

Statut : Publié le 1er avril 2019

Inclusion : Chapitre 6

Changement pour l'inclusion : Aucun

Contribution des auteurs : P.-L.P a conçu le projet, coordonné la collaboration, effectué les expériences, construit l'outil et rédigé le manuscrit. E.F.-F. a participé à la réalisation des expériences, à la construction de l'outil et à la révision du manuscrit. J.C.M., J.A.M. et E.S.B. ont fourni des données, révisé le manuscrit et effectué des mesures analytiques supplémentaires pour valider certaines hypothèses. F.L., M.M. et J.C. ont participé à la conception du projet et à la révision du manuscrit.

Publication en annexe

Puisque je ne suis pas le premier auteur sur cette publication, elle ne peut être insérée dans cette thèse. Toutefois, elle a eu un impact important dans mon cheminement scientifique et dans les travaux présentés dans cette thèse. Elle est donc discutée au chapitre 1 et peut être retrouvée à l'annexe A.

Mass spectra alignment using virtual lock-masses

Francis Brochu, Pier-Luc Plante, Alexandre Drouin, Dominic Gagnon, Dave Richard, Francine Durocher, Caroline Diorio, Mario Marchand, Jacques Corbeil, François Laviolette
Scientific reports volume 9 (1), 1-15 (2019)

Statut : Publié le 11 juin 2019
Inclusion : Annexe A
Changement pour l'inclusion : La notation employée a été mise à jour pour correspondre à celle employée dans la thèse.
Contribution des auteurs : P.-L.P. et J.C. ont conçu et effectué les expériences de métabolomique. F.B., P.-L.P., A.D., M.M. et F.L. ont conçu l'algorithme. F.B., M.M. et F.L. ont conçu les expériences computationnelles. F.B., P.-L.P., A.D. et M.M. ont programmé l'algorithme. F.B. a conduit les expériences computationnelles et analysé les résultats. D.G., D.R., F.D. et C.D. ont validé les résultats et révisé le manuscrit. J.C. a participé au design initial et a révisé le manuscrit. F.B., P.-L.P., A.D., M.M. et F.L. ont écrit et révisé le manuscrit.

Autres publications

La bio-informatique étant une science de collaboration, j'ai participé et collaboré durant mes études à de nombreux projets qui sont plus ou moins éloignés du sujet de cette thèse. Les

publications liées à ces projets ne sont donc pas incluses à l'intérieur de celle-ci. Voici tout de même une liste de ces travaux et de mes contributions :

Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter

Moïra B. Dion, Pier-Luc Plante, Edwige Zufferey, Shiraz A. Shah, Jacques Corbeil and Sylvain Moineau

Nucleic Acids Research (2021)

Statut: Publié le 2 mars 2021

Contribution des auteurs : M.B.D. a conçu l'étude, supervisé par S.M.. M.B.D., P.-L.P. et E.Z. ont effectué les analyses et rédigé le manuscrit, supervisé par S.M. et J.C.. S.A.S. a contribué à l'analyse des résultats et à la conceptualisation de l'étude. Tous les auteurs ont lu et approuvé le manuscrit.

A lactococcal phage protein promotes viral propagation and alters the host proteomic response during infection

Marie-Laurence Lemay, Sandra Maaß, Andreas Otto, Jérémie Hamel, Pier-Luc Plante, Geneviève M. Rousseau, Denise M. Tremblay, Rong Shi, Jacques Corbeil, Stéphane M. Gagné, Dörte Becher, Sylvain Moineau

Viruses (2020)

Statut: Publié le 24 juillet 2020

Contribution des auteurs : Conception : M.-L.L., R.S., S.M.G. et S.M. (Sylvain Moineau) ; Nettoyage des données : M.-L.L., S.M. (Sandra Maaß), A.O., J.H., P.-L.P., G.M.R., D.M.T., R.S, S.M.G. et D.B. ; Analyse formelle : M.-L.L., J.H., P.-L.P., G.M.R., D.M.T., R.S, S.M.G., D.B. et S.M. (Sylvain Moineau) ; Acquisition des fonds de recherche : R.S, S.M.G. et S.M. (Sylvain Moineau) ; Investigation : M.-L.L., S.M. (Sandra Maaß), A.O., P.-L.P., G.M.R., D.M.T., R.S., J.C., S.M.G., D.B. et S.M. (Sylvain Moineau) ; Méthodologie : M.-L.L., S.M. (Sandra Maaß), A.O., J.H., P.-L.P., G.M.R., D.B., R.S., J.C., S.M.G. et D.B. ; Administration du projet : S.M. (Sylvain Moineau) ; Ressources : M.-L.L., R.S., J.C., S.M.G., D.B. et S.M. (Sylvain Moineau) ; Logiciels : M.-L.L., S.M. (Sandra Maaß), A.O., P.-L.P, R.S., J.C., S.M.G. et D.B. ; Supervision : R.S, S.M.G., D.B. et S.M. (Sylvain Moineau) ; Validation : M.-L.L.,

S.M. (Sandra Maaß), A.O., G.M.R., D.M.T., R.S. et S.M.G. ; Visualisation : M.-L.L., S.M. (Sandra Maaß), A.O., G.M.R., D.M.T., R.S. et S.M.G. ; Écriture—première version : M.-L.L. ; Écriture—révision et édition : M.-L.L., G.M.R., D.M.T., R.S, J.C., S.M.G., D.B. et S.M. (Sylvain Moineau). Tous les auteurs ont lu et accepté la version finale du manuscrit.

The initial state of the human gut microbiome determines its reshaping by antibiotics
Frédéric Raymond, Amin A Ouameur, Maxime Déraspe, Naeem Iqbal, Hélène Gingras, Bédis Dridi, Philippe Leprohon, Pier-Luc Plante, Richard Giroux, Ève Bérubé, Johanne Frenette, Dominique K Boudreau, Jean-Luc Simard, Isabelle Chabot, Marc-Christian Domingo, Sylvie Trottier, Maurice Boissinot, Ann Huletsky, Paul H Roy, Marc Ouellette, Michel G Bergeron & Jacques Corbeil
The ISME Journal volume 10 (3), 707–720 (2016)

Statut : Publié le 11 septembre 2015
Contribution des auteurs : F.R. a contribué au design expérimental, préparé les librairies de séquençage, séquencé les échantillons, effectué les analyses bio-informatiques et l'exploration de données, analysé les résultats et écrit le manuscrit avec M.B. F.R., A.A.O., N.I. et H.G. ont préparé les librairies de séquençage. E.B. a géré les échantillons fécaux. A.A.O. et R.G. ont extrait les acides nucléiques. M.D. a créé et nettoyé les bases de données de référence. M.D., P.-L.P., D.K.B. et P.H.R. ont effectué les analyses bio-informatiques. I.C et S.T. ont réalisé l'étude clinique. B.D., P.L., J.F., J.-L.S. et M.-C.D. ont offert le soutien technique et scientifique au projet. S.T., M.B., A.H., P.H.R., M.O., M.G.B. et J.C. ont conçu l'étude et supervisé le projet. Tous les auteurs ont révisé et accepté le manuscrit.

Culture-enriched human gut microbiomes reveal core and accessory resistance genes
Frédéric Raymond, Maurice Boissinot, Amin Ahmed Ouameur, Maxime Déraspe, Pier-Luc Plante, Sewagnouin Rogia Kpanou, Ève Bérubé, Ann Huletsky, Paul H Roy, Marc Ouellette, Michel G Bergeron, Jacques Corbeil
Microbiome volume 7 (1), 56 (2019)

Statut : Publié le 5 avril 2019
Contribution des auteurs : F.R. et M.B. ont contribué au design expérimental, séquencé les échantillons, effectué les analyses bio-informatiques et l'exploration de données,

analysé les résultats et écrit le manuscrit. A.A.O. a préparé les librairies de séquençage. E.B. a géré les échantillons fécaux. A.A.O. et E.B. ont cultivé les échantillons fécaux et extrait les acides nucléiques. M.D., P.-L.P., S.R.K. et P.H.R. ont effectué les analyses bio-informatiques. A.H., P.H.R., M.O., M.G.B. et J.C. ont offert le soutien technique et scientifique au projet. M.B., A.H., P.H.R., M.O., M.G.B. et J.C. ont conçu l'étude et supervisé le projet. Tous les auteurs ont révisé et accepté la version finale du manuscrit.

Prédiction de l'observance à court terme d'une diète méditerranéenne grâce à une approche métabolomique non ciblée et des études d'alimentation contrôlées

Didier Brassard, Francis Brière, Nancy Boucher, Pier-Luc Plante, Jacques Corbeil, Simone Lemieux, Benoit Lamarche

Nutrition Clinique et Métabolisme volume 34 (1), 89-90 (2020)

Statut : Publié en avril 2020 (conférence)

Contribution des auteurs : D.B., S.L. et B.L. ont conçu et réalisé l'étude nutritionnelle. F.B, N.B. et P.-L.P. ont réalisé l'expérience de métabolomique par spectrométrie de masse. F.B., P.-L.P. et D.B. ont analysé les données de métabolomique. D.B. a écrit le manuscrit. Tous les auteurs ont révisé le manuscrit.

Phenotypic and Genetic Characterization of the Cheese Ripening Yeast Geotrichum candidum

Vincent Perkins, Stéphanie Vignola, Marie-Hélène Lessard, Pier-Luc Plante, Jacques Corbeil, Eric Dugat-Bony, Michel Frenette, Steve Labrie

Frontiers in Microbiology volume 11, 737 (2020)

Statut : Publié le 7 mai 2020

Contribution des auteurs : V.P., S.V., M.-H.L., E.D.-B., M.F., et S.L. ont participé à la planification de l'étude et à l'écriture du manuscrit. S.V. a effectué les caractérisations phénotypiques, l'extraction de l'ADN, l'assemblage des génomes et les analyses de génomique comparative. M.-H.L. a préparé les librairies pour le séquençage par HiSeq. V.P. a extrait l'ADN pour l'analyse MLST, effectué l'analyse MLST, assemblé les génomes et effectué des analyses de génomique comparative. P.-L.P. et J.C. ont participé au développement d'outils bio-informatiques pour l'analyse des génomes.

Genomic characterization of a large outbreak of Legionella pneumophila serogroup 1 strains in Quebec City, 2012

Simon Lévesque, Pier-Luc Plante, Nilmini Mendis, Philippe Cantin, Geneviève Marchand, Hugues Charest, Frédéric Raymond, Caroline Huot, Isabelle Goupil-Sormany, François Desbiens, Sébastien P Faucher, Jacques Corbeil, Cécile Tremblay

PLoS One volume 9 (8), e103852 (2014)

Statut : Publié le 8 août 2014

Contribution des auteurs : Conception des expériences : S.L., S.F., J.C., C.T. Réalisation des expériences : S.L., P.-L.P., N.M., P.C., G.M., F.R. Analyse des données : S.L., P.-L.P., N.M., P.C., G.M., H.C., F.R., C.H., I.G.-S., F.D., S.F., J.C., C.T. Contribution aux réactifs/matériels/outils d'analyse : S.L., P.C., G.M., S.F., J.C., C.T. Contribution à l'écriture du manuscrit : S.L., P.-L.P., N.M., S.F., J.C., C.T.

Improving the safety of Staphylococcus aureus polyvalent phages by their production on a Staphylococcus xylosus strain

Lynn El Haddad, Nour Ben Abdallah, Pier-Luc Plante, Jeannot Dumaresq, Ramaz Katsarava, Steve Labrie, Jacques Corbeil, Daniel St-Gelais, Sylvain Moineau

PLoS One volume 9 (7), e102600 (2014)

Statut : Publié le 25 juillet 2014

Contribution des auteurs : Conception des expériences : S.M., D.S.-G., J.C., S.L., L.E.H. Réalisation des expériences : L.E.H., N.B.A., P.-L.P. Analyse des données : L.E.H., N.B.A., P.-L.P., J.D., R.K., S.L., J.C., D.S.-G., S.M. Contribution aux réactifs/matériels/outils d'analyse : L.E.H., N.B.A., P.-L.P., R.K., J.C., D.S.-G., S.M. Contribution à l'écriture du manuscrit : L.E.H., N.B.A., P.-L.P., J.D., R.K., S.L., J.C., D.S.-G., S.M.

Targeted proteomics of human metapneumovirus in clinical samples and viral cultures

Matthew W Foster, Geoff Gerhardt, Lynda Robitaille, Pier-Luc Plante, Guy Boivin, Jacques Corbeil, M Arthur Moseley

Analytical chemistry volume 87 (20), 10247-10254 (2015)

Statut : Publié le 16 septembre 2015

Contribution des auteurs : M.W.F et G.G. ont développé et effectué les expériences de protéomique. L.R., P.-L.P. et J.C. ont effectué les analyses génomiques. J.C. et M.A.M. ont conçu l'étude. G.B. a fourni les échantillons cliniques. M.W.F. a écrit le manuscrit. Tous les auteurs ont révisé et accepté la version finale du manuscrit.

Introduction

La fin des « omiques »

La génomique, le champ de recherche ayant pour but d'étudier le matériel génétique, a été la première branche des « omiques ». Pour la première fois, on cherchait à caractériser l'entièreté d'une des molécules de la vie : l'ADN. De la génomique est venu le premier brouillon du génome humain qui a été complété au début de l'année 2001 [1]. Cette longue séquence de A, C, T et G venait avec une promesse immense :

> « *Humankind is on the verge of gaining immense new power to heal. Genome science [...] will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases.* » [2] - Bill Clinton, 2010

Effectivement, c'est la connaissance du génome qui permet le développement d'études d'associations pangénomiques (*genome-wide association study*, GWAS) et pharmacogénomiques, la création du *Cancer Genome Atlas* et l'étude des maladies rares et monogéniques [1]. Ces approches et ces projets ont eu, et continuent d'avoir, un impact monumental sur la santé des populations et ont vu le jour grâce au travail colossal des centaines de chercheurs qui ont séquencé le premier génome. Cependant, après plus de 20 ans et plusieurs milliers de génomes humains séquencés, il est maintenant clair que la connaissance de l'ADN humain n'est qu'une partie de la clé permettant d'accéder aux remèdes à toutes les maladies. L'héritage génétique, certes important, n'est qu'un facteur parmi d'autres [3].

Avec la génomique sont ensuite venues la transcriptomique et la protéomique. Tel que leurs noms le suggèrent, ces dernières étudient l'ARN transcrit à partir du génome et les protéines traduites à partir des ARN par le ribosome. La motivation derrière l'exploration des transcrits était évidente : la transcription de certaines régions de l'ADN permet à ce dernier d'avoir un impact sur l'organisme. En étudiant l'abondance des transcrits, on obtient aussi une idée des processus de régulation de la transcription. Avec l'arrivée des puces à ADN et des séquenceurs à haut débit, l'étude des transcrits est devenue beaucoup plus simple et accessible. La transcriptomique a permis, par exemple, de détecter un nouvel exon dans le gène RPGR, qui est lié aux rétines pigmentaires [4]. Elle est aussi utilisée pour aider au diagnostic de certains cancers afin d'évaluer la réponse au traitement [5]. Malheureusement, l'instabilité de certains ARN et la corrélation imparfaite entre l'abondance des transcrits et des protéines obligent aussi l'étude de ces dernières [6–8]. Encore ici, l'information obtenue

est directement dérivée du génome : l'ADN encode des transcrits qui sont ensuite traduits en protéines. L'implication des différentes fonctions des protéines sur les maladies est au cœur de la recherche en protéomique. Il est relativement simple de détecter un ARN et sa protéine, mais il est complexe d'en déterminer la fonction. Les bases de données en sont la preuve : elles contiennent les séquences d'une infinité de gènes ayant une fonction toujours inconnue. La complexité est encore plus grande quand on étudie les protéines puisqu'elles sont modifiées au cours du temps (ex. : phosphorylation, acétylation, etc.) pour effectuer diverses fonctions.

On étudie aujourd'hui de multiples couches d'information dérivées directement du fameux code génétique, mais nous sommes encore loin de comprendre et de pouvoir guérir toutes les maladies. Après tout cela, que pouvait-il rester? Depuis des siècles, on mesure l'abondance de petites molécules dans le sang et l'urine afin de détecter certaines maladies. Par exemple, au XI^{e} siècle, avant la découverte des gènes et protéines, on goûtait l'urine pour diagnostiquer le diabète, car elle prenait un goût sucré [9]. Aujourd'hui, on cherche à identifier et à quantifier l'ensemble des molécules se trouvant dans un système biologique. L'étude de ces petites molécules, appelée la métabolomique, est à la fois la plus vieille et la plus récente des sciences « omiques ». Alors que pour les autres « omiques » l'ensemble des molécules pouvant être mesurées sont déterminées à partir du génome, il n'existe pas encore aujourd'hui de liste complète des petites molécules pouvant se retrouver dans un organisme complexe tel que l'humain. En effet, le métabolome est directement affecté par les autres « omiques », mais aussi par des facteurs externes tels que l'alimentation et les conditions de vie qui ajoutent de nouvelles molécules très diversifiées dans le métabolome; c'est ce qu'on appelle l'exposomique. Les métabolites interagissent directement avec le système biologique que ce soit pour la production d'énergie, la communication cellulaire ou la réaction aux stimuli. La métabolomique est donc très près du phénotype observé. De ce fait, il est naturel de vouloir étudier l'ensemble de ces petites molécules afin d'obtenir une photo instantanée de l'état d'un système biologique à un moment donné. L'ADN n'est finalement que la première couche d'un réseau d'interactions complexe entre génomique, transcriptomique, protéomique et métabolomique. Le métabolome est donc l'état de transformation final de l'information dans un organisme, et la métabolomique est l'ultime science « omique ».

Métabolomique

La métabolomique est définie comme étant l'étude des petites molécules contenues dans un échantillon biologique. Ces petites molécules, qui ont généralement une masse moléculaire située entre 50 et 1200 daltons (Da), peuvent se trouver dans le sang, dans une cellule et voire même dans un fromage! L'étude du métabolome présente beaucoup d'intérêts pour plusieurs domaines dans lesquels les micro- et macroorganismes ont un impact tels que la santé, la nutrition, l'agroalimentation et l'environnement.

La métabolomique appliquée à la santé a comme principal objectif d'identifier des biomarqueurs pour le diagnostic et la prédiction de maladies [10]. Par exemple, en étudiant le profil métabolique de l'urine d'enfants atteints d'une bronchiolite, il a été remarqué que les sujets à risque de développer un sifflement pulmonaire récurrent présentaient une différence au niveau du métabolisme de l'acide citrique et des acides aminés. Il devenait donc possible d'identifier dès la première hospitalisation quels enfants étaient susceptibles d'avoir ce sifflement et d'en prévenir la récurrence. La métabolomique peut aussi être employée afin de diagnostiquer et de sélectionner le meilleur traitement pour différents cancers en plus d'évaluer la réponse au dit traitement [11–13]. Il est également possible de faire des pronostics, c'est-à-dire prédire l'apparition de maladies [13,14]. En dehors du diagnostic, l'étude du métabolome de patients malades permet d'obtenir de l'information sur les mécanismes et chemins métaboliques impliqués dans une maladie. À partir des résultats de ces études, il est possible d'identifier de nouvelles cibles thérapeutiques ou des médicaments existants pouvant aider au traitement de la maladie [15]. L'effet des médicaments sur le métabolisme peut aussi être mesuré afin de comprendre leurs mécanismes d'actions et de déterminer le taux de métabolisation des molécules pour un individu. Par exemple, la PharmGKB* est une base de données qui définit les liens entre les différences dans le génome humain et le métabolisme des médicaments [16]. À partir de cette information, il devient possible d'adapter un traitement en fonction de la présence de certaines mutations génétiques chez un individu. La métabolomique est donc aujourd'hui un outil de premier plan dans le domaine de la santé pour l'amélioration des diagnostics, la création d'outils de support au pronostic, l'étude des mécanismes moléculaires ainsi que le développement et le suivi de nouveaux traitements.

* https://www.pharmgkb.org/

La métabolomique est aussi couramment utilisée dans les études nutritionnelles afin de mesurer l'effet de la diète sur le métabolisme. Par exemple, Jin *et al.* ont utilisé la métabolomique afin de mesurer l'effet d'une diète méditerranéenne sur les chemins métabolomiques reconnus comme ayant un impact dans différentes maladies [17]. D'autres emploient la métabolomique afin de personnaliser les recommandations nutritionnelles [18]. L'avantage des études métabolomiques en nutrition vient du lien intime entre la consommation de métabolites et leur présence dans l'organisme. Par exemple, la FooDB*contient une liste de métabolites trouvés dans différents aliments ainsi que ceux pouvant être retrouvés dans un système biologique à la suite de leur consommation.

Dans le domaine agroalimentaire, la métabolomique peut servir à l'étude des processus de transformation biologiques tels que la fermentation ainsi qu'à l'identification de la fraude alimentaire. Le profilage métabolique du fromage durant son vieillissement a permis d'identifier de fines différences au niveau de la quantité de certains acides aminés en fonction de la taille des colonies impliquées dans l'affinage [19]. La métabolomique a aussi été utilisée pour contrer la fraude alimentaire pour de nombreux produits tels que le miel, le vinaigre balsamique, la pâte de tomates, le safran et le café [20]. En analysant la composition en petites molécules des produits, il devient possible de distinguer la signature métabolique liée aux produits frauduleux de celle liée à des produits de bonne qualité. Un autre exemple dans le domaine agroalimentaire est la quantification d'androstérone et de scatol dans la viande de porc. La quantification de ces hormones est maintenant effectuée dans certains pays européens où la castration des porcs est illégale. En quantifiant ces molécules, les abattoirs espèrent réduire les pertes liées à la forte odeur que ces composés peuvent produire et, par le fait même, au rejet des produits par les consommateurs [21].

Finalement, dans le domaine environnemental, l'analyse des métabolites produits par les microorganismes présents dans le sol permet d'évaluer l'effet de perturbateurs, tels que la température et la pollution, sur différentes communautés écologiques [22–24]. Un groupe de chercheurs a utilisé la métabolomique afin de caractériser les biofilms dans des systèmes de filtration d'eau [25]. Ces résultats ont permis d'établir un lien entre la composition des tuyaux, la qualité de l'eau et l'activité microbienne. Quoique l'analyse du métabolome de populations microbiennes complexes telles que celles retrouvées dans l'environnement soit

* https://foodb.ca/

difficile et moins répandue, il est tout de même possible d'obtenir une vue d'ensemble des processus biologiques présents et d'en tirer des conclusions ayant un grand potentiel environnemental et économique.

Dans plusieurs exemples mentionnés précédemment, on cherche à évaluer l'effet d'un facteur, comme une maladie ou un changement d'environnement, sur le métabolisme d'un organisme ou d'une population microbienne. Ce facteur a souvent un effet sur l'expression génique et protéique, mais ces dernières sont sujettes à un contrôle homéostasique strict et direct qui n'est pas aussi présent au niveau du métabolome [26,27]. De ce fait, la métabolomique peut être un indicateur plus sensible de la réponse à un facteur de stress comparativement aux autres « omiques », ce qui la rend très intéressante. De plus, contrairement à des marqueurs génomiques et protéomiques qui sont sujets aux modifications génétiques, la détection de métabolites est peu affectée par les changements liés à l'évolution. La structure moléculaire d'un métabolite restera toujours la même.

Métabolomique ciblée et non ciblée

Plusieurs méthodes analytiques permettent de détecter et de quantifier des métabolites avec plus ou moins de spécificité. La quantification du glucose dans le sang par spectrophotométrie [28] et la quantification de la vitamine D par immunochimie [29] sont des exemples de méthodes quantifiant des métabolites précis. L'analyse métabolomique est historiquement et plus communément effectuée de manière ciblée, c'est-à-dire qu'une ou plusieurs molécules sont sélectionnées a priori et cette sélection affecte l'ensemble de l'expérience, de la préparation de l'échantillon à l'analyse instrumentale (Figure 0.1). On débute par le développement d'une méthode analytique permettant la quantification des molécules purifiées. Une fois la méthode en place, on l'utilise pour quantifier les métabolites dans une matrice biologique comme le plasma, l'urine ou un extrait cellulaire. Chaque étape est optimisée afin de favoriser la sélection des molécules ciblées et l'élimination des interférents. Ce type de méthode est particulièrement intéressant lorsqu'on désire effectuer la quantification de certains composés. Afin de valider les résultats, la précision de la quantification peut être confirmée par la mesure d'échantillons contrôles ayant une concentration prédéfinie. On peut aussi caractériser la méthode afin de déterminer différentes métriques telles que la linéarité, les limites inférieure et supérieure de quantification, la précision, la reproductibilité, la résistance aux interférents et la limite de détection. Des organismes réglementaires tels que la *Food and Drug Administration* (FDA)

encadrent le processus de validation de telles méthodes par la publication de recommandations [30]. Cette validation est critique dans certaines applications où il est nécessaire d'avoir une grande confiance dans les résultats obtenus et où la reproductibilité est cruciale, par exemple pour des tests cliniques. Quoique très utile pour quantifier avec précision un métabolite d'intérêt, ce type de méthodes donne toutefois une vision très limitée de la complexité du métabolome plasmatique où interagissent des milliers de petites molécules.

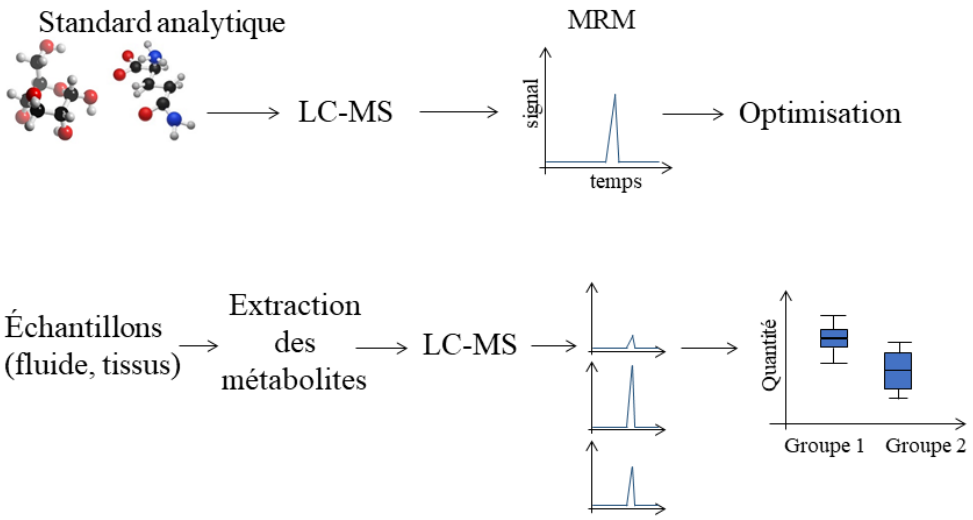


Figure 0.1 Schématisation d'une expérience de métabolomique ciblée par chromatographie liquide et spectrométrie de masse (LC-MS). On débute avec l'analyse d'un standard contenant les molécules d'intérêt purifiées. On optimise ensuite les étapes de préparation d'échantillons et d'analyse instrumentale. Finalement, on procède à l'expérience où l'on compare le signal obtenu pour les molécules désirées seulement à travers les différents groupes d'échantillons. En présence d'une courbe standard, il est possible d'effectuer une quantification absolue des métabolites d'intérêt.

D'un autre côté, les avancées technologiques permettent aujourd'hui d'effectuer des analyses en ayant aucune hypothèse de départ : la métabolomique non ciblée. On obtient ainsi une vue beaucoup plus élargie du métabolome. L'utilisation de méthodes analytiques polyvalentes, telles que la résonnance magnétique nucléaire (NMR) et la spectrométrie de masse (MS), permettent de mesurer des centaines, voire des milliers de métabolites, dans un seul échantillon. L'objectif est d'effectuer une analyse quantitative non biaisée d'un grand nombre des métabolites trouvés dans un échantillon biologique [27]. Les extractions métaboliques employées sont les plus inclusives possibles, mais permettent d'éliminer les contaminants comme les protéines pouvant interférer avec l'analyse. La méthode analytique, que ce soit par MS ou NMR, produit une grande quantité de données

correspondant à la structure moléculaire et à l'abondance d'une multitude de métabolites. Ce type d'expérience est aussi caractérisé par une analyse de données beaucoup plus complexe (Figure 0.2). De même, seule une quantification relative est possible, c'est-à-dire la comparaison de l'intensité du signal entre les échantillons pour un métabolite donné, puisque le signal produit en spectrométrie de masse pour une molécule varie en fonction de son taux d'ionisation. La métabolomique non ciblée est particulièrement appropriée lorsqu'on cherche à identifier des métabolites inconnus permettant de différencier des phénotypes, tels que des patients malades et sains. La caractérisation et l'évaluation de la qualité des résultats en métabolomique non ciblée sont cependant plus difficiles. Toutefois, certaines recommandations telles que l'évaluation de l'état de l'instrumentation avant l'analyse et l'incorporation d'échantillons contrôles facilitent ce processus [31]. Les conclusions d'études métabolomiques non ciblées peuvent être validées par une analyse de métabolomique ciblée employant une méthode bien caractérisée. On passe donc de la phase de découverte à la phase de validation.

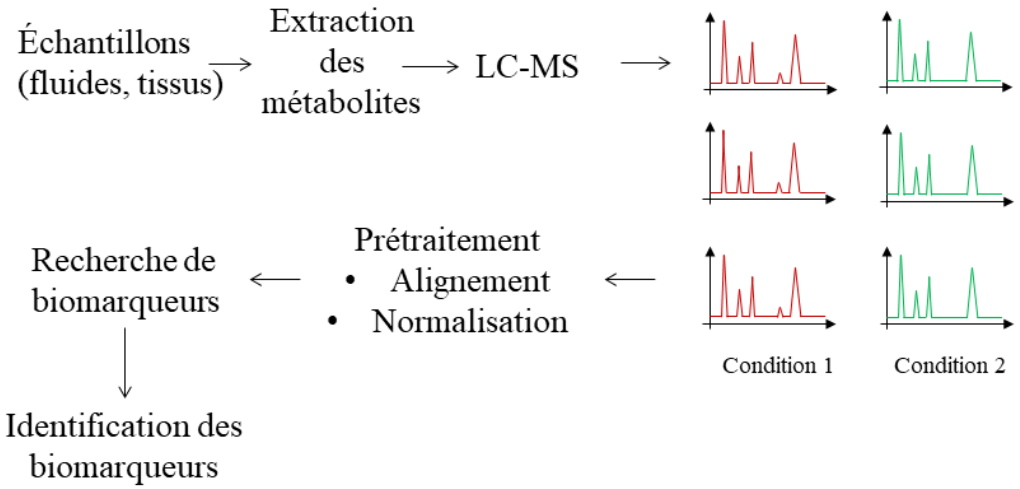


Figure 0.2 Schématisation des étapes d'une expérience de métabolomique non ciblée par chromatographie liquide et spectrométrie de masse (LC-MS). On débute par l'extraction de façon à mettre en solution une grande variété de métabolites provenant de l'échantillon de départ. L'acquisition des données par LC-MS sur le mélange permet d'obtenir des chromatogrammes pour toutes les masses comprises à l'intérieur d'un intervalle de masses prédéfini. Les données produites sont ensuite alignées et normalisées pour les rendre comparables. La recherche de biomarqueurs est ensuite effectuée par analyse statistique. Finalement, on cherche l'identification des molécules correspondant aux biomarqueurs sélectionnés à l'étape précédente.

La distinction entre la métabolomique ciblée et non ciblée peut être comparée à la célèbre analogie de l'aiguille dans la botte de foin. En utilisant la spectrométrie de masse, il est généralement possible de trouver l'aiguille cachée dans l'amas de foin, c'est-à-dire qu'on

réussit habituellement à trouver les molécules recherchées, malgré la complexité du mélange où elles se trouvent. Toutefois, le foin contient une grande quantité d'information et il peut être très intéressant d'aller fouiller afin d'y trouver autre chose qu'une aiguille. Cette analogie représente bien la différence entre la métabolomique ciblée et non ciblée : la première est dirigée par une hypothèse et l'autre cherche à générer des hypothèses.

Spectrométrie de masse

Comme mentionné précédemment, la spectrométrie de masse est une méthode analytique permettant l'investigation du métabolome. Que ce soit pour une analyse ciblée ou non ciblée, il est possible d'utiliser une approche basée sur la spectrométrie de masse. Celle-ci est particulièrement puissante puisqu'elle permet d'obtenir une information qualitative et quantitative pour un ensemble de molécules. L'avantage de la spectrométrie de masse par rapport à plusieurs autres méthodes est le gain de spécificité lié à la mesure de la masse moléculaire et à la potentielle fragmentation. De plus, les spectromètres de masses sont considérés comme beaucoup plus sensibles qu'un NRM, ce qui en fait le choix parfait pour l'étude du métabolome.

Définition de termes courants en spectrométrie de masse

Sensibilité : Capacité de l'instrument à détecter une petite quantité d'analyte. Cette caractéristique est directement corrélée au ratio signal/bruit. On estime que le signal produit par un analyte doit être au moins trois fois supérieur au bruit pour être considéré. La sensibilité est directement liée à la limite de détection de l'instrument.

Résolution : Capacité de l'instrument à distinguer deux signaux (pics) rapprochés. Les signaux peuvent être des pics de masses, des pics chromatographiques et même de mobilité ionique. La résolution de masse est calculée selon l'équation 1.1

Précision : Capacité de l'instrument à répéter la même mesure.

Spécificité : Caractéristique d'une mesure qui permet de la distinguer des interférents. Par exemple, une molécule possédant une structure particulière pourra produire un fragment spécifique qui permet de distinguer cette mesure des autres molécules ayant une masse identique.

Exactitude : Capacité de l'instrument à mesurer la masse d'un ion avec peu de différence par rapport à la valeur théorique.

Puisque cette thèse porte sur la métabolomique par spectrométrie de masse, il est nécessaire d'expliquer le fonctionnement général de l'instrumentation afin de comprendre les différents facteurs influençant la production des données. Un spectromètre de masse (MS) est un instrument couramment utilisé en chimie

analytique et qui a pour fonction de mesurer le ratio masse sur charge (m/z) de molécules chargées.

> «The basic principle of mass spectrometry (MS) is to generate ions … by any suitable method, to separate these ions by their mass-to-charge ratio (m/z) and to detect them qualitatively and quantitatively by their respective m/z and abundance. The analyte may be ionized thermally, by electric fields or by impacting energetic electrons, ions or photons. The ... ions can be single ionized atoms, clusters, molecules or their fragments or associates. Ion separation is effected by static or dynamic electric or magnetic fields.» [32]

Selon cette définition, on peut diviser un spectromètre de masse en trois éléments : une source d'ions, un analyseur permettant d'évaluer le m/z des ions et un détecteur permettant de quantifier ces ions. Le MS permet donc d'obtenir minimalement une information qualitative (c.-à-d. le m/z de l'ion atteignant le détecteur) et une information quantitative (c.-à-d. le nombre d'ions atteignant le détecteur).

Sources d'ionisation Afin d'entrer dans l'analyseur pour ensuite être détectée et produire un signal, une molécule doit respecter au minimum deux conditions : elle doit être sous forme gazeuse et elle doit être chargée. C'est la source d'ionisation qui a comme fonction de placer les molécules dans cet état.

Une des sources d'ionisation utilisée le plus couramment en métabolomique est la source d'ionisation à électronébuliseur (ESI). Son fonctionnement suit le processus suivant. Les analytes sont d'abord dilués dans un solvant, puis cette solution est vaporisée en fines gouttelettes. Ensuite, le solvant s'évapore et les ions sont éjectés des gouttelettes chargées par une fission de Coulomb [32,33]. Les analytes précédemment en solution se retrouvent donc en phase gazeuse et chargés. Les principaux avantages d'utiliser ce type de source d'ionisation en métabolomique sont la douceur de l'ionisation et sa polyvalence. De plus, contrairement à d'autres méthodes d'ionisation, l'ESI cause très peu de fragmentation à même la source ce qui permet de facilement détecter l'ion moléculaire. On doit aussi mentionner sa compatibilité avec la chromatographie liquide. Cette dernière est une méthode d'introduction de l'échantillon basée sur la différence d'affinité des analytes pour un support solide (la colonne) et un solvant (la phase mobile). L'échantillon d'intérêt est d'abord placé à l'entrée de la colonne, puis les phases mobiles l'entraînent à l'intérieur. La colonne est remplie de fines particules pouvant posséder différentes propriétés physico-

chimiques. Les molécules contenues dans l'échantillon interagissent avec la phase solide, ce qui retarde leur expulsion vers la sortie. En appliquant un changement dans la composition de la phase mobile, les molécules de l'échantillon sont éventuellement entraînées par la phase liquide pour laquelle elles ont une meilleure affinité. Le liquide finit son parcours dans le capillaire de la source ESI où il est vaporisé. Comme les analytes sont expulsés de la colonne à un moment précis, le signal est concentré sur une courte période de temps ce qui augmente la sensibilité d'un tel système. En plus d'être compatible avec la chromatographie liquide, la source ESI permet l'injection en flux et l'infusion, ce qui lui donne un avantage indéniable en métabolomique pour l'analyse des molécules biologiques, qui sont majoritairement solubles.

La source MALDI (*Matrix Assisted Laser Desorption Ionisation*) est une autre méthode d'ionisation employée en métabolomique [34]. Elle est apparue presque au même moment que la source ESI mais elle fonctionne complètement différemment. Les analytes sont d'abord mélangés à une matrice qui peut être excitée par un laser. La matrice utilisée doit pouvoir absorber l'énergie d'un laser, UV ou IR, et la transformer en charge (ex. : acide picolinique (PA), acide dihydroxybenzoïque (DHB) et dihydroxyacetophenone (DHAP)). La plaque contenant les échantillons est placée sous vide, puis un faisceau laser est dirigé vers le mélange où la matrice absorbe l'énergie du laser, ce qui cause une désorption et une ionisation. Le réglage de la puissance et de la pulsation à grande vitesse du laser permet un contrôle de la désorption. Étant donné la taille du faisceau laser, sa résolution spatiale est autour de 0,1mm. Cette caractéristique permet au MALDI d'être utilisé pour effectuer des analyses d'imagerie par spectrométrie de masse. Contrairement à la source ESI qui peut facilement produire des ions multichargés (ex. : M+2, M+3, etc.), la source MALDI produit principalement des ions simplement chargés selon le principe du « *lucky survivor* » [32,35]. Le MALDI est donc très intéressant pour l'analyse de polymères et de peptides afin de ne pas distribuer le signal d'une molécule en plusieurs ions possédant différentes charges. La source MALDI peut aussi être employée pour différents modes d'ionisation tels que la désorption/ionisation par impact laser exaltée de surface (*surface enhanced laser desorption ionization*, SELDI), la désorption/ionisation laser (*laser desorption ionization*, LDI) et plusieurs autres dérivés.

La source LDTD (*Laser Diode Thermal Desorption*) est un autre type de source d'ionisation pouvant être employée en spectrométrie de masse. La source LDTD utilise l'énergie

produite par un laser afin d'augmenter la température d'une feuille métallique (Figure 0.3). De l'autre côté de cette feuille métallique se trouvent les métabolites cristallisés. L'absorption de l'énergie par les molécules cristallisées les fait passer sous forme gazeuse. Le laser est contrôlé grâce à un patron de puissance permettant d'obtenir une désorption optimale pour les molécules étudiées. Un tube de transfert dans lequel circule un gaz porteur, typiquement de l'air ambiant, est employé pour favoriser le transfert des molécules sous forme gazeuse vers la zone d'ionisation chimique à pression atmosphérique (APCI). Contrairement à la source ESI qui peut rendre difficile l'analyse de plus d'un échantillon par minute, la technologie LDTD permet d'analyser un échantillon en quelques secondes. Cependant, l'énergie utilisée pour la désorption a tendance à favoriser la fragmentation de certains composés à la source (ex. : les phosphatidylglycérols) alors que d'autres composés sont presque impossibles à désorber (ex. : les protéines). Ce faisant, l'ion moléculaire ne peut être détecté. La dérivatisation chimique, qui consiste à modifier une molécule en lui ajoutant un nouveau groupe fonctionnel, peut favoriser la désorption. Cette approche permet, par exemple, d'analyser des acides aminés qui sont autrement très peu désorbés [36].

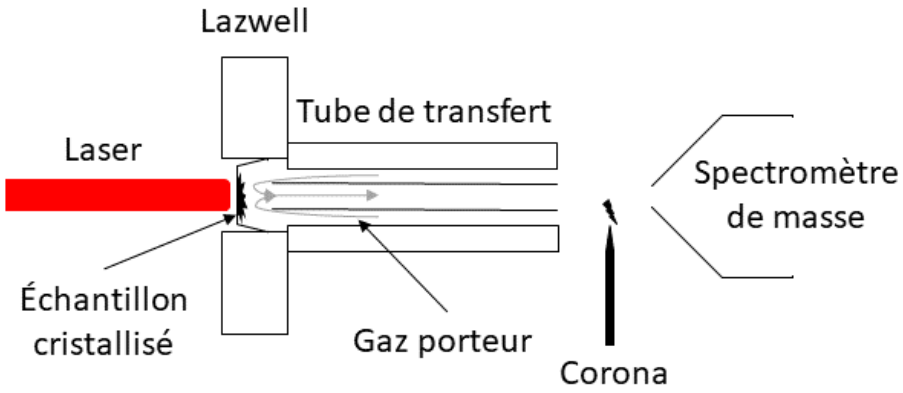


Figure 0.3 Représentation du fonctionnement d'une source LDTD. Les analytes sont cristallisés sur la surface métallique au fond du puits d'une Lazwell. Une diode laser chauffe ensuite l'arrière de la surface métallique. Le transfert d'énergie cause une désorption des molécules cristallisées. Les molécules maintenant sous forme gazeuse sont entraînées dans le tube de transfert vers le spectromètre de masse par le gaz porteur. Avant d'entrer dans le spectromètre de masse, les molécules sont ionisées par ionisation chimique à pression atmosphérique (APCI) grâce à l'aiguille corona.

La source DESI (*Desorption Electrospray Ionisation*) est une source d'ionisation à pression ambiante qui envoie des gouttelettes chargées provenant d'un électronébuliseur, similaire à celui d'une source ESI, vers une surface où se trouve l'échantillon [37]. Le solvant formant les gouttelettes arrive au nébuliseur à un débit de 1 à 20 µl/min de façon à produire un jet à haute pression sur une petite surface. À l'impact, les gouttelettes entraînent une partie de

l'échantillon en phase gazeuse où elles sont aspirées par le bazooka vers le spectromètre de masse (Figure 0.4). Le processus d'ionisation est très similaire à celui d'une source ESI et, de ce fait, produit des spectres très similaires. En changeant la constitution du solvant employé, il est possible de favoriser la désorption de certaines classes de métabolites. Comme pour le MALDI, on peut séparer les composés dans l'espace bidimensionnel de la surface analysée de façon à effectuer de l'imagerie par spectrométrie de masse. Comparativement au MALDI, la résolution de l'image est moins élevée (environ 3 fois moins), car la dimension du jet de solvant est plus large que la dimension du laser. Par contre, aucune matrice n'est nécessaire et l'analyse s'effectue à pression ambiante.

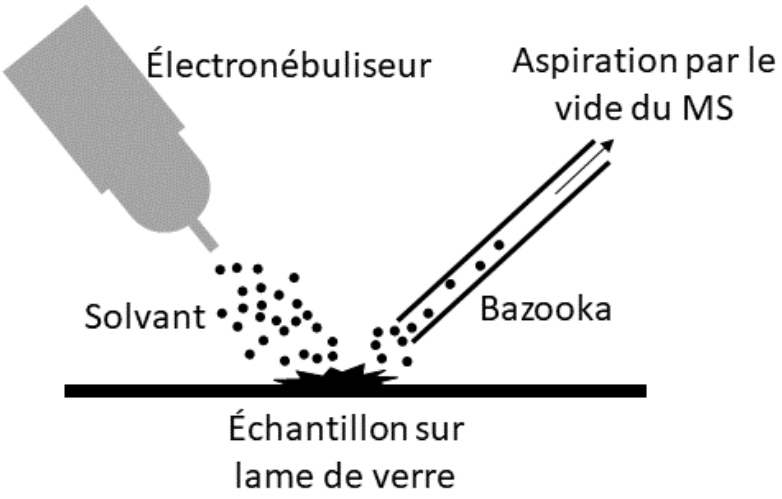


Figure 0.4 Représentation du fonctionnement d'une source DESI. Les molécules sont d'abord cristallisées sur une surface de verre. Dans le cas d'un tissu, ce dernier est fixé à la surface. Un mélange de solvants volatil est ensuite nébulisé sur l'échantillon. Les microgouttelettes produites par l'électronébuliseur solubilisent une partie de l'échantillon et l'entraînent en phase gazeuse. Le bazooka aspire les métabolites chargés vers le spectromètre de masse.

Les sources d'ionisation sont diverses et elles peuvent être employées de différentes façons. Par exemple, la source ESI peut être utilisée pour la chromatographie liquide, l'infusion et l'injection directe. Seules les sources les plus pertinentes dans le cadre des travaux de recherche présentés ici ont été décrites. Il est important de considérer les particularités de la source d'ionisation employée dans une expérience. Dans certains cas, les molécules d'intérêt pourraient ne jamais entrer dans l'analyseur si la source d'ionisation n'est pas appropriée. On pourrait résoudre ce problème combinant les mesures obtenues avec deux sources d'ionisation différentes, chacune proposant une ionisation optimale pour certains types de composés. Ce choix aura un impact sur la vitesse d'acquisition et la qualité

des spectres obtenus. Ces caractéristiques sont aussi influencées par les analyseurs, la deuxième section d'un spectromètre de masse.

Analyseurs Il existe une multitude d'analyseurs, chacun ayant des caractéristiques favorisant son utilisation dans certains contextes. Ces analyseurs peuvent être classifiés en deux grandes catégories : les filtres et les séparateurs. Les premiers éliminent les ions ne correspondant pas à la fenêtre de masse désirée alors que les deuxièmes séparent les ions dans le temps et/ou l'espace en fonction de leur m/z. L'analyseur influence grandement la sensibilité, c'est-à-dire la capacité à détecter de petites quantités, ainsi que la résolution de masse, c'est-à-dire la capacité à distinguer deux masses très rapprochées. Le quadripôle et la chambre à temps de vol sont décrits ci-dessous afin d'exposer la différence entre les filtres et les séparateurs.

Le plus commun des analyseurs agissant comme filtre est le quadripôle. Il est composé de quatre tiges de métal équidistantes formant un losange. Les tiges les plus éloignées forment des paires sur lesquelles un potentiel électrique identique est appliqué. C'est ce potentiel, déterminé par les équations de Mathieu, qui permet de filtrer les ions naviguant à l'intérieur du quadripôle [32]. Les ions ne pouvant rester stables à l'intérieur du champ électrique produit par les tiges sortent du champ de stabilité et vont se coller sur une des tiges. Ce faisant, ils n'atteindront jamais le détecteur. Le seul paramètre affectant la vitesse d'acquisition est le temps requis pour changer le potentiel appliqué sur les pôles afin de sélectionner une autre plage de masses. Les quadripôles sont reconnus pour offrir une bonne sensibilité et une résolution permettant une séparation entre 1 et 0,5 unité de masse atomique (uma).

La chambre de temps de vol (ToF) à accélération orthogonale est, quant à elle, un analyseur de type séparateur. Contrairement au quadripôle qui filtre les ions, la ToF permet de voir l'ensemble des ions d'un seul coup. Il consiste en un *pusher* et une chambre de temps de vol sous forme de tube tenue à très basse pression (~$1x10^{-8}$ tor). Lors d'une acquisition, un paquet d'ions est poussé dans le tube par le *pusher*. Étant donné que les ions reçoivent tous une poussée identique au même moment, c'est l'énergie cinétique des ions, qui est directement liée au m/z, qui permet de les séparer. Les ions ayant différentes masses mais la même énergie cinétique, ils voyagent plus ou moins rapidement dans le tube et atteignent le détecteur à différents temps : les plus petits arrivent en premier et les plus lourds, en dernier. Dès que l'ion le plus lourd est arrivé au détecteur, une nouvelle série d'ions est

envoyée par le *pusher*. Étant donné le fonctionnement de l'analyseur, il est facile d'en accroître la résolution en augmentant la longueur du chemin parcouru par les ions. Il existe deux façons d'y arriver : soit ajouter un réflectron à l'extrémité du tube pour que les ions effectuent des allers et retours, soit augmenter la longueur du tube. La vitesse d'acquisition n'est limitée que par le temps de déplacement des ions dans le tube. La chambre de temps de vol peut donc acquérir jusqu'à 10 000 spectres par seconde. Afin d'obtenir une bonne sensibilité et une bonne précision, une unité de calcul incluse à même l'instrument fait la somme des spectres sur une unité de temps (ex. : 0,1 sec). La sensibilité de cet analyseur est inférieure à celle du quadripôle, mais sa résolution de masse est hautement supérieure. Certains instruments de type ToF sont même qualifiés de « haute résolution » puisqu'ils offrent une résolution supérieure à 50 000x.

Des analyseurs peuvent être couplés afin de gagner en spécificité et d'augmenter la polyvalence de l'instrument. Par exemple, on peut utiliser deux quadripôles pour produire un instrument de type triple-quadripôles (*triple-quad*). On peut aussi combiner un quadripôle et un ToF dans un même instrument afin d'obtenir un spectromètre de masse en tandem de type Q-ToF. Un des principaux avantages lors de la combinaison de deux analyseurs est l'ajout d'une cellule de collision sous la forme d'un quadripôle. Celle-ci permet la fragmentation des molécules préalablement sélectionnées par le premier analyseur. C'est d'ailleurs cette composante qui est le deuxième quadripôle d'un triple-quadripôle. Il devient possible de sélectionner un ion dans le premier analyseur, de fragmenter cet ion dans la cellule de collision, puis de mesurer le m/z des fragments grâce au deuxième analyseur. Le signal résultant est un spectre MS/MS. On améliore ainsi l'information qualitative obtenue sur une molécule par son patron de fragmentation et on augmente la spécificité des mesures en ciblant des fragments propres à une structure moléculaire (Figure 0.5).

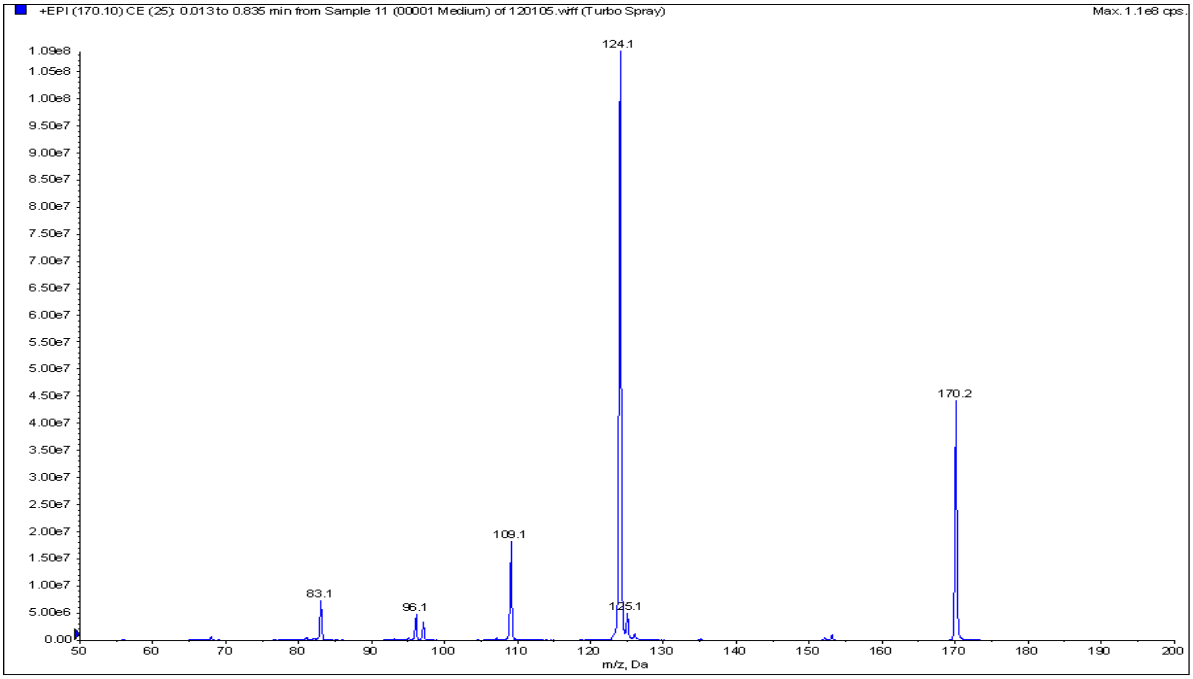


Figure 0.5 Spectre de fragmentation (MS/MS) de la 1-Methylhistidine (masse moléculaire 169,085 Da) en ionisation positive. L'ion moléculaire ($[M+H]^+$) a un m/z de 170,2 Da. Les autres pics de masses correspondent à des fragments moléculaires issus de la fragmentation de l'ion précurseur. Source : HMDB [38]

D'autres analyseurs filtrants et séparateurs tels que la cellule de résonance cyclotronique ionique (FT-ICR), l'orbitrap et le secteur magnétique peuvent être utilisés en métabolomique. Chacun possède un pouvoir de résolution, une exactitude de masse et une vitesse d'acquisition distincte. L'augmentation de la sensibilité se fait souvent au détriment de la résolution. Il est donc nécessaire d'établir les besoins analytiques avant de choisir un instrument.

Détecteurs Il existe quelques types de détecteurs pouvant être employés dans un spectromètre de masse. Le détecteur peut avoir un impact sur la résolution et la sensibilité. Les détecteurs les plus communs sont les multiplicateurs d'électrons et les plaques à microcanaux (*microchannel plates*). Lors d'un choix d'instrument, on aura tendance à sélectionner l'analyseur et à se contenter du détecteur fourni puisque ce dernier sera compatible avec l'analyseur.

Cellule de mobilité ionique En plus des composantes standards, un spectromètre de masse peut être équipé d'une cellule de mobilité ionique. Cette cellule sépare les ions sous l'influence d'un champ électrique dans un gaz inerte [39]. Elle agit donc comme une étape de

séparation supplémentaire se basant sur la forme des ions. Certains types de cellules de mobilité ionique comme le DTIMS (*Drift Tube Ion Mobility Spectrometry*), le TWIMS (*Travelling Wave Ion Mobility Spectrometry*) et le TIMS (*Trapped Ion Mobility Spectrometry*) permettent de mesurer la section efficace (C*ollisional Cross Section*, CCS) des molécules analysées. La CCS peut être grossièrement défini comme étant la moyenne de l'aire des surfaces d'une molécule. Cette propriété physico-chimique est stable et non-dépendante de l'instrument de mesure, ce qui la rend très intéressante pour caractériser une molécule, par exemple pour comparer la valeur mesurée à la valeur trouvée dans une base de données. Cette caractéristique des molécules s'ajoute donc à l'information qualitative mesurée par l'instrument lorsqu'une cellule de mobilité ionique est présente.

En résumé, les spectromètres de masse sont des instruments très polyvalents dont les performances sont définies par leurs composantes individuelles. Le type d'ionisation, la résolution sur la mesure de masse et la sensibilité sont des facteurs critiques dans une expérience de métabolomique. Aucun instrument n'offre une performance parfaite et, pour cette raison, il est important de connaître les limites de l'instrumentation lors de l'analyse de données.

Analyse de données en métabolomique

Un spectromètre de masse, peu importe ses composantes, produit des données. Comme mentionné précédemment, la spectrométrie de masse permet de mesurer le m/z et la quantité d'ions frappant le détecteur (c.-à-d. leur abondance). Ces informations sont habituellement représentées par un spectre de masse. Il s'agit souvent d'un graphique montrant l'intensité relative en fonction du m/z (Figure 0.6 A). Cette représentation peut être effectuée en mode continu, où l'on peut voir la distribution des masses mesurées pour un ion (Figure 0.6 B), ou en centroïde, où chaque pic de masse détecté est remplacé par une seule mesure (Figure 0.6 C). Une technique courante pour convertir un pic en centroïde est d'utiliser le m/z au point milieu à 50 ou 80% de la hauteur. Ce faisant, la transformation évite les biais qui pourraient être présents à l'apex, garde un maximum de précision sur la mesure du m/z et élimine les risques de contamination du signal à la base par un autre pic de masse mal résolu. La transformation en centroïde a certains avantages tels que la diminution de la taille des fichiers de données. On perd toutefois l'information quant à la distribution des mesures ce qui pourrait empêcher de diagnostiquer des problèmes tels que des composés non résolus.

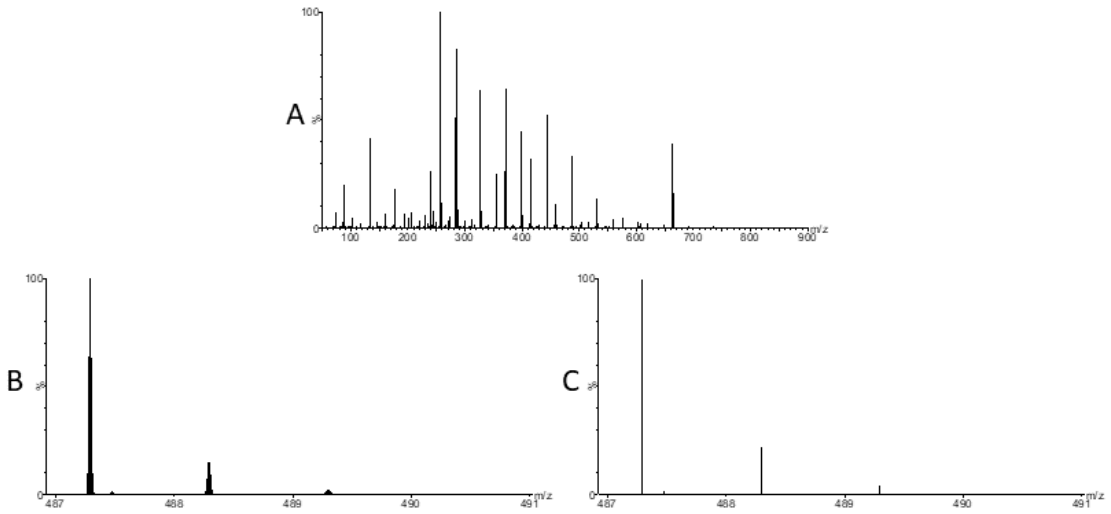


Figure 0.6 Représentations graphiques de spectres de masse. L'abscisse correspond au ratio m/z et l'ordonnée représente l'abondance relative normalisée au signal de l'ion le plus élevé. A- Spectre de masse complet. B- Section d'un spectre de masse en continu. C- Section d'un spectre de masse centroïde.

Un spectre de masse peut représenter un seul scan ou la somme des scans effectués par l'instrument sur une période donnée. Puisque l'acquisition des données s'effectue sur une période allant d'une fraction de seconde à plusieurs minutes, on peut représenter l'information sous forme de chromatogramme où l'abscisse représente le temps et l'ordonnée représente l'intensité du signal (Figure 0.7 A). Chaque point d'un chromatogramme peut correspondre à la somme des intensités d'un spectre de masse ou à l'intensité d'un sous-ensemble des masses mesurées. Dans le premier cas, on parle de chromatogramme des ions totaux (TIC) et dans le deuxième, de chromatogramme des ions extraits (XIC) (Figure 0.7). Finalement, certains spectromètres de masse spécialisés sont équipés d'une cellule de mobilité ionique permettant une séparation rapide des composés en fonction de leur structure et de leur charge. On comprend donc que les données de spectrométrie de masses sont minimalement bidimensionnelles (c.-à-d. m/z et intensité), mais peuvent aussi être tridimensionnelles (c.-à-d. temps, m/z et intensité) ou quadridimensionnelles (c.-à-d. temps, m/z, mobilité ionique et intensité).

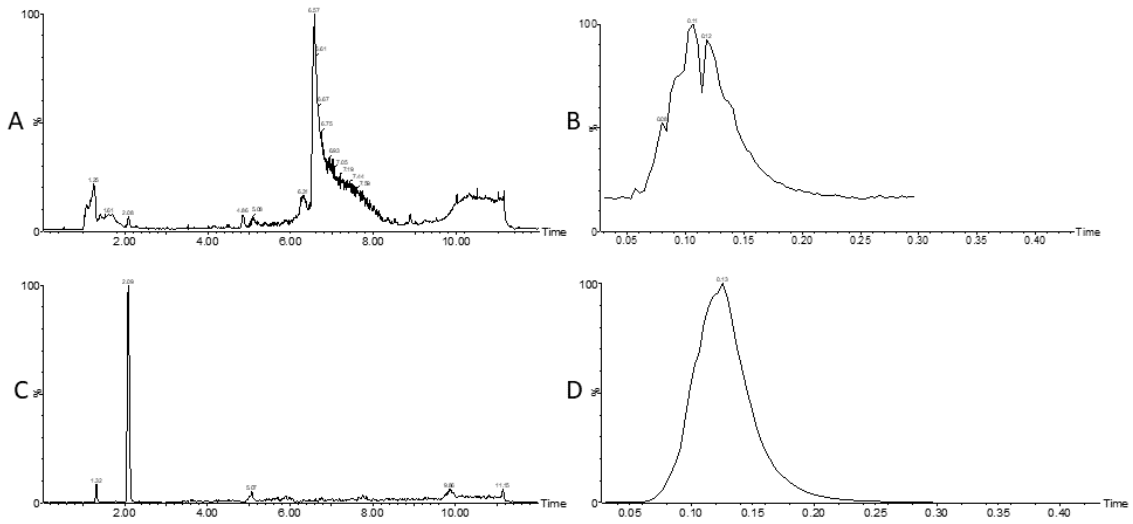


Figure 0.7 Chromatogrammes démontrant l'intensité du signal en fonction du temps. A- Chromatogramme des ions totaux (TIC) d'une acquisition par LC-MS de 12 minutes. B- TIC d'une acquisition par LDTD-MS de 0,3 minutes. C- Chromatogramme des ions extraits (XIC) de A pour le m/z 485,15. D- XIC de B pour le m/z 312,2.

Une fois l'acquisition des données effectuée, on doit aligner les signaux afin de synchroniser parfaitement les mesures dans les différentes dimensions qualitatives (m/z, temps de rétention et/ou mobilité ionique). Cette opération est nécessaire afin d'éliminer les variations pouvant être observées lors de l'acquisition et de rendre les données comparables. Selon l'instrumentation et les méthodes utilisées, les variations peuvent être plus ou moins importantes. Par exemple, pour une expérience de métabolomique non ciblée par LC-MS sur un système de chromatographie Waters Acquity I-Class couplé à un spectromètre de masse Waters Synapt G2-Si opéré en mode résolution, avec un correcteur de masse, on peut s'attendre à des variations de moins de 15 secondes sur le temps de rétention et de moins de 15 parties par million (ppm) sur la masse à l'intérieur de quelques heures. On remarque que la variation sur le temps de rétention est en unité absolue (secondes) alors que la variation sur la masse est en unité relative (ppm). Cette différence s'explique par le fait que la variation sur le temps de rétention est quasi constante sur la longueur de la chromatographie alors que la variation sur le m/z est corrélée avec le m/z. Cette distance entre deux points de masse, m_1 et m_2, peut être calculée selon l'Équation 0.1.

$$D = \frac{|m_1 - m_2|}{m_1 + m_2} * 2 * 10^6$$

Équation 0.1. Distance en parties par million entre deux points de masse

Deux types de variations peuvent être observées dans un jeu de données. Il y a d'abord les variations systémiques qui sont des variations relativement stables pour un échantillon. Un minime changement dans le pH ou la composition d'une phase mobile pourrait affecter les temps de rétention d'un groupe d'échantillons. De la même façon, un changement de température ambiante pourrait affecter la mesure des masses effectuée avec un ToF. Ces variations sont les plus faciles à corriger. Par exemple, l'utilisation d'un correcteur de masse (*lock-mass*) avec un spectromètre de masse à temps de vol permet de faire une très bonne correction de la mesure de masse pour la réduire à moins de 10 ppm de la masse réelle. L'autre type de variation, dite aléatoire, est lié à l'imperfection des mesures analytiques. Par exemple, l'apex d'un pic de masse peut être décalé pour un gain de 0,0002 Da dans un échantillon, puis décalé pour une perte de 0,0002 Da dans l'échantillon suivant. Ces petits décalages sont causés par l'imperfection de la distribution des mesures du m/z et sont imprévisibles. Les variations aléatoires sont beaucoup plus petites mais plus difficiles à corriger. Le *binning*, qui consiste à regrouper les signaux dans des boîtes de taille fixe, est une solution à ces variations [40].

L'alignement de chromatogrammes est un problème commun qui peut être résolu de différentes façons, bien que l'objectif final demeure le même : maximiser le critère d'alignement afin de rendre la dimension temporelle de deux chromatogrammes la plus semblable possible. Une première méthode consiste à compresser et à dilater certaines régions d'un chromatogramme afin de permettre un alignement parfait de certains pics marqueurs [41]. Ces marqueurs peuvent être prédéfinis et venir d'un échantillon contrôle ou ils peuvent être déterminés computationnellement post-acquisition. Une autre façon d'aligner des chromatogrammes est par l'insertion et la délétion de points de données. Contrairement à la première approche où l'on modifie les temps de mesure, on insère ou enlève plutôt des points afin de favoriser l'alignement des pics marqueurs. Ces deux méthodes sont particulièrement efficaces pour l'alignement d'un seul chromatogramme comme un XIC, mais sont difficilement applicables à la métabolomique non ciblée où les données ont au moins trois dimensions dont minimalement deux doivent être alignées. Pour ce genre de données, on peut tirer profit de la dimensionnalité des données acquises en combinant chromatographie et spectrométrie de masse pour permettre une décomposition trilinéaire [42]. D'autres ont proposé d'utiliser l'information spécifique des spectres MS/MS afin de détecter des marqueurs de référence permettant de calculer, sans ambiguïté, un facteur de correction pour l'alignement de l'information entre les marqueurs [43]. Ces méthodes

permettent de corriger dans toutes les dimensions. L'alignement étant une étape critique, elle est au cœur de plusieurs logiciels propriétaires pour l'analyse en métabolomique tels que *Progenesis QI* (Nonlinear Dynamics, Waters Corp.), *Compound Discoverer* (Thermo Scientific) de même que dans les logiciels ouverts comme *XCMS* [44] et *El-Maven* (Elucidata Inc).

Une fois l'alignement terminé, une étape de sélection de pics doit être effectuée. Dans un spectre de masse, la recherche de pics sert à trouver quels sont les m/z qui se démarquent du bruit de fond. Dans l'espace multidimensionnel d'un chromatogramme LC-MS, on cherche la combinaison m/z et temps de rétention qui se démarque du bruit de fond. Les pics sont ensuite caractérisés en fonction de différents critères tels que l'intensité du pic LC et MS, le patron isotopique, le profil d'élution et la largeur du pic chromatographique [45]. En calculant la différence entre la masse mesurée pour chaque pic, on peut effectuer une déconvolution. Cette dernière vise à regrouper ensemble les isotopes et les différents ions provenant d'une même molécule. Cette étape est particulièrement révélatrice dans le cas d'une chromatographie (par exemple, une chromatographie liquide) puisque les isotopes et les ions d'une même molécule auront tous le même temps de rétention. On réduit ainsi la quantité de pics transférés aux prochaines étapes d'analyse et par le fait même, la dimensionnalité des données et la corrélation entre différents éléments. De plus, s'il y a présence de plusieurs types d'ions, il est possible de calculer la masse neutre de la molécule, une caractéristique qui aide à l'identification. En absence de séparation, il devient difficile de distinguer le signal provenant des différentes molécules de celui provenant de différents types d'ions dans un spectre de masse. La déconvolution s'en trouve affectée et est souvent impossible.

La normalisation des intensités est la prochaine étape à effectuer dans un pipeline d'analyse métabolomique non ciblée. Elle est nécessaire afin de mettre en évidence les variations biologiques en absence des variations expérimentales. Cette transformation s'effectue au niveau de l'intensité des pics. Les variations peuvent venir d'un effet de matrice (c.-à-d. la variation du signal causée par la présence d'autres molécules), d'un pic co-éluant (c.-à-d. deux molécules ayant le même temps de rétention), d'une variation de l'efficacité de l'ionisation et même de la préparation de l'échantillon. Une fois la normalisation appliquée, la qualité de l'information quantitative se trouve grandement améliorée. Il existe deux grandes familles de méthodes de normalisation : à partir du signal total et à partir d'un

standard [46]. Dans le premier cas, on peut utiliser la somme des intensités afin de représenter l'abondance sous la forme d'un ratio. Une variante de cette méthode consiste à utiliser la somme des intensités des composés présents dans l'ensemble des échantillons au lieu de la somme des ions totaux. Cette méthode est plus puissante puisqu'elle permet d'utiliser l'homéostasie naturelle des échantillons afin de normaliser l'ensemble des métabolites en faisant abstraction des contaminants. Pour les méthodes utilisant un standard, on insère un ou plusieurs métabolites, à une concentration fixe, dans l'ensemble des échantillons. Puisque l'intensité de ces métabolites est censée être stable, on peut calculer un facteur de correction applicable aux métabolites de l'échantillon. L'utilisation d'un standard marqué avec du C^{13}, ^{2}H, ^{15}N et/ou ^{18}O permet d'éviter la contamination du signal des métabolites de l'échantillon. En employant un groupe de molécules marquées au lieu d'une seule, la qualité de la correction s'en trouve grandement améliorée. Des standards marqués provenant du métabolisme de la levure ou de l'algue peuvent être employés afin d'obtenir un standard complexe contenant une multitude de composés [46]. Ces derniers sont particulièrement intéressants pour des études métabolomiques chez d'autres organismes vivants avec qui ils partagent plusieurs sentiers métaboliques.

Une fois toutes ces étapes de préparation complétées, les données sont dites comparables, c'est-à-dire que l'intensité du pic m_1 de l'échantillon *i* peut être comparée à l'intensité du pic m_1 d'un autre échantillon. Cependant, en l'absence d'une courbe de quantification pour les molécules correspondant aux pics m_1 et m_2, il est impossible de comparer leur intensité puisque ces deux métabolites ont un taux d'ionisation différent. En comparant les intensités des signaux, on peut chercher à identifier un ou plusieurs ions pour lesquels l'intensité est différente entre deux ou plusieurs conditions. C'est ce qu'on appelle la recherche de biomarqueurs. Cette recherche est effectuée principalement par apprentissage automatique. Plusieurs précautions doivent être prises lors de cette étape, car on fait face à une matrice à haute dimension où le nombre d'exemples (les échantillons) est très inférieur au nombre d'éléments (les ions). Il devient donc statistiquement plus probable de tomber sur un élément montrant une différence significative entre les différents groupes d'échantillons. La prochaine section discutera plus en détail de l'utilisation de l'apprentissage automatique en métabolomique pour la recherche de biomarqueurs. Si un modèle interprétable est employé, les métabolites sélectionnés peuvent passer à la prochaine étape : l'identification des biomarqueurs.

L'identification des biomarqueurs consiste à chercher l'identité du métabolite ayant produit le signal observé en spectrométrie de masse. Cette étape requiert une bonne connaissance du contexte biologique de l'échantillon afin de limiter l'espace de recherche. Par exemple, un métabolite spécifique à la consommation de viande de baleine a peu de chance d'être retrouvé fréquemment dans le plasma d'une population nord-américaine. Pour la recherche, on se base sur l'information disponible : le m/z de l'ion moléculaire et le spectre MS/MS. Dans certains cas, un index de temps de rétention, la CCS ou d'autres informations peuvent être accessibles et favoriser l'identification. On compare ces informations à ce qui est offert dans différentes bases de données telles que la Human *Metabolome Database*[*], *FooDB*[†], *MzCloud*[‡], *MoNA*[§] et *GNPS*[**] [47,48]. Il existe en fait une multitude de bases de données, chacune étant plus ou moins spécialisée. En choisissant les bases de données les plus appropriées selon le contexte biologique, il est possible de réduire le champ de recherche ainsi que le taux de fausses identifications. Selon la quantité d'information utilisée, un indice de confiance peut être attribué à une identification. Cette échelle de confiance a été établie par la Metabolomics Society Initiative afin d'aider à la standardisation des expériences en métabolomique[49].

Une fois toutes ces étapes complétées, les métabolites et le modèle peuvent passer par différents processus de validation afin de s'assurer de leur qualité prédictive. Au même titre que l'instrumentation, chacune des étapes du traitement de données aura un impact sur la qualité des biomarqueurs identifiés. L'utilisation d'algorithmes d'apprentissage automatique à certaines étapes de l'analyse des données, tel que pour la recherche de biomarqueurs, est reconnue comme étant l'état de l'art. La prochaine section présente l'utilisation de tels algorithmes dans le contexte de l'analyse de données en métabolomique non ciblée.

Apprentissage automatique pour la métabolomique

L'apprentissage automatique est une branche de l'intelligence artificielle qui a pour objectif de donner la capacité à un ordinateur d'effectuer une tâche sans avoir été explicitement programmé pour l'effectuer [50]. On peut découper l'apprentissage automatique en quatre catégories : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-

[*] https://hmdb.ca/
[†] https://foodb.ca/
[‡] https://www.mzcloud.org/
[§] https://mona.fiehnlab.ucdavis.edu/
[**] https://gnps.ucsd.edu/

supervisé et l'apprentissage par renforcement [51]. Ceux-ci diffèrent dans l'approche utilisée pour apprendre et dans leurs objectifs d'apprentissage. Peu importe la méthode, un algorithme utilise un ensemble de N exemples ($X = \{x_i\}_{i=1}^{N}$) où chaque x_i est un vecteur de k éléments ($x_i = \{x_{i,j}\}_{j=1}^{k}$), représentant des descripteurs, à partir duquel l'algorithme doit apprendre.

Les apprentissages supervisé et non supervisé sont les deux types les plus couramment employés dans le domaine de la métabolomique. En apprentissage supervisé, on donne à l'algorithme les exemples, chacun associé à un objectif ou à une classe ($Y = \{y_i\}_{i=1}^{N}$). À partir de ces exemples, l'algorithme cherche à trouver, durant la phase d'entraînement, une règle permettant de déterminer y_i à partir de x_i pour tout i. À l'opposé, en apprentissage non supervisé, on donne seulement des exemples (X) à l'algorithme, sans Y. L'idée ici est d'avoir un algorithme qui convertit un exemple en une ou plusieurs valeurs. Les algorithmes de regroupement (*clustering*) et de réduction de dimensionnalité sont des exemples d'apprentissage non supervisé.

Dans tous les cas, un des objectifs d'un algorithme d'apprentissage automatique est de généraliser, c'est-à-dire de prendre la bonne décision à partir d'un exemple qui n'a jamais été vu par l'algorithme lors de l'entraînement. S'il y a absence de généralisation et que l'algorithme réussit seulement à travailler avec les exemples qu'il a déjà vus, on parle de surapprentissage. Ce problème affecte particulièrement l'apprentissage supervisé appliqué à la métabolomique non ciblée pour la recherche de biomarqueurs, car on est en présence de données à haute dimensionnalité où le nombre de descripteurs, dans ce cas des ions, est beaucoup plus élevé que le nombre d'exemples (les échantillons). Pour éviter le surapprentissage, on cherche à employer des algorithmes produisant des solutions compréhensibles, employant un petit nombre de descripteurs et on utilise des méthodes telles que la validation croisée pour la sélection des hyperparamètres.

Validation croisée

Deux types de validation croisée sont couramment employés en métabolomique : la validation croisée de Monte-Carlo et la validation croisée en *k* parties (Figure 0.8). La première a pour objectif d'évaluer l'effet de la constitution de l'ensemble d'entraînement et de l'ensemble de test sur les performances de l'algorithme tout en permétant d'évaluer la

généralisation. Pour ce faire, on sépare l'ensemble du jeu de données de *n* façons différentes. On obtient *n* ensembles d'entraînement et *n* ensembles de tests. On construit ensuite un modèle pour chaque ensemble d'entraînement et on évalue ses performances sur l'ensemble de test correspondant. En étudiant la distribution des métriques de performances sur les différents ensembles de test, on obtient une meilleure estimation des capacités de performance du modèle sur le jeu de données.

La validation en *k* parties a pour objectif de réduire les probabilités de surapprentissage. Pour l'effectuer, on sépare d'abord l'ensemble d'entraînement en *k* parties. Pour chaque combinaison d'hyperparamètres et pour chaque partie, on entraîne un modèle sur *k-1* parties et on évalue les performances sur la partie restante. La moyenne des scores obtenus sur les *k* parties d'évaluation correspond au score de validation croisée. La combinaison d'hyperparamètres donnant le meilleur score sera employée pour entraîner un modèle sur l'ensemble des *k* parties, l'ensemble d'entraînement complet.

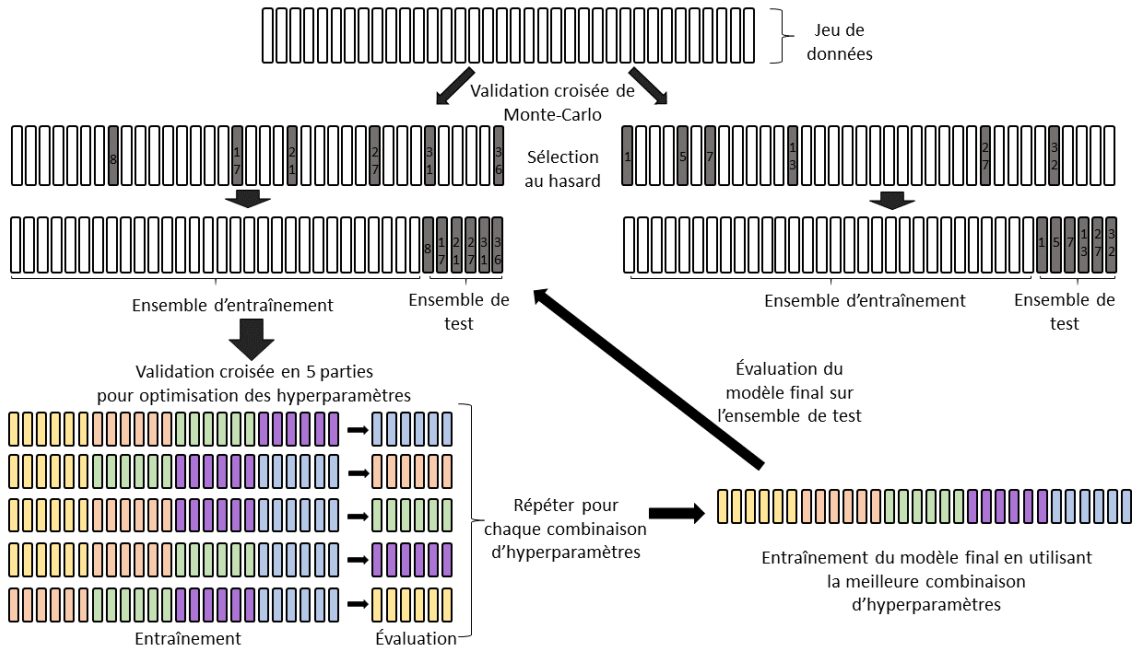


Figure 0.8 Processus de validation croisée employé en apprentissage automatique appliqué à la métabolomique. Une validation croisée de Monte-Carlo est d'abord employée pour évaluer l'effet de la constitution de l'ensemble d'entraînement et de test sur les performances. Pour chaque séparation du jeu de données, on procède à une validation croisée en k parties pour optimiser les hyperparamètres associés à chaque algorithme.

Algorithmes d'apprentissage automatique

Comme mentionné précédemment, l'apprentissage automatique est la principale méthode utilisée pour la recherche de biomarqueurs dans le contexte de la métabolomique non ciblée.

> « *Fundamentally, the goal of biomarker development in metabolomics is to create a predictive model from a collection of multiple compounds, which can be used to classify new samples/persons into specific groups (e.g. healthy vs. diseased) with optimal sensitivity and specificity.* » [52]

Trois algorithmes ont été utilisés à profusion dans le passé pour la construction de modèles supervisés en métabolomique : la PLS-DA (*Partial Least Square – Discriminant Analysis*), la forêt aléatoire et le SVM linéaire[12,53–58].

PLS-DA

La PLS-DA est une méthode de réduction de la dimensionnalité qui s'effectue de façon supervisée. L'idée ici est de construire un nouvel ensemble de n variables $Z_1, Z_2, \ldots Z_n$, où chacune est une combinaison linéaire des variables originales[59]. La classification s'effectue dans ce nouvel espace multidimensionnel. Dans un problème à deux classes, la nouvelle variable Z_1 est une composition linéaire des variables utiles à la classification. L'avantage principal de la PLS-DA est qu'elle offre la possibilité d'obtenir un graphique des scores qui illustre la séparation entre les groupes. De plus, il est possible de connaître l'ensemble des descripteurs les plus importants dans le processus de séparation en étudiant l'importance des variables sur la projection. Contrairement à d'autres algorithmes, il ne permet pas d'obtenir un modèle simple employant un nombre minimal de descripteurs discriminants. Il a aussi de fortes chances de montrer de bonnes performances seulement par la chance dans des données à haute dimension [54]. Il est couramment employé en métabolomique, car il est présent dans la majorité des logiciels d'analyses fournis par les manufacturiers d'instrument, ce qui le rend très facile d'accès pour les non-initiés à l'analyse de données.

Forêt aléatoire et arbre de décision

La forêt aléatoire est un algorithme qui utilise un ensemble d'arbres de décision qui effectuent un vote de majorité. Un arbre de décision est un graphe acyclique où chaque

nœud évalue la valeur d'un descripteur grâce à une règle [60]. En combinant ces règles, on arrive à une feuille correspondant à la prédiction. La Figure 0.9 est un exemple d'un arbre de décision produit à partir de données de métabolomique. Plus on descend profondément dans un arbre de décision, moins il y a d'exemples à classifier et plus on augmente le risque de surapprentissage où un descripteur est utilisé pour classifier correctement un échantillon sans être lié à la classe. Afin d'éviter ce genre de situation, on peut émonder un arbre de décision. Ce processus peut entraîner l'ajout des erreurs de classification sur les données d'entraînement, mais permet d'augmenter la généralisation sur les nouvelles données. À noter que malgré sa faible utilisation en métabolomique, l'arbre de décision peut aussi être employé avec des données à haute dimensionnalité afin de trouver un groupe minimal de descripteurs nécessaires à différencier deux groupes [61]. Dans une forêt aléatoire, chacun des arbres est construit à partir d'un sous-ensemble des exemples et des descripteurs, choisis au hasard, afin d'éviter une corrélation entre les différents arbres. La prédiction de la forêt aléatoire correspond alors au vote de majorité de la forêt d'arbres. Dans le but de déterminer les éléments importants dans le processus de décision, il est possible de calculer le coefficient de Gini correspondant à chaque élément. Cette métrique correspond à la moyenne de la diminution de l'impureté d'un nœud sur l'ensemble des arbres de la forêt. En d'autres mots, il indique à quelle fréquence un élément $x_{i,j}$ a été sélectionné pour un nœud et son pouvoir discriminant dans le problème de classification [57].

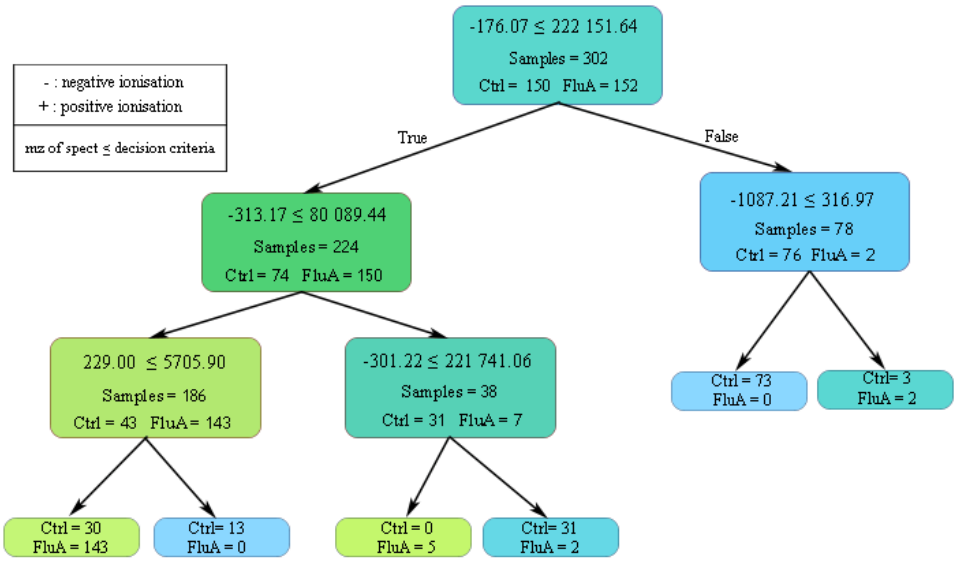


Figure 0.9 Exemple d'un arbre de décision appliqué au problème d'infection par le virus de l'influenza tel que présenté au chapitre 4. Chaque nœud représente une règle binaire jusqu'à atteindre une feuille qui prédit la classe d'un échantillon.

Machine à vecteurs de support linéaire

Le SVM linéaire est lui aussi parfois utilisé en métabolomique. L'algorithme place les *i* exemples d'entraînement dans l'espace multidimensionnel à *k* dimensions. Dans cet espace, il cherche à tracer un hyperplan permettant de séparer les différentes classes. L'hyperparamètre principal de l'algorithme, communément appelé C, permet de contrôler la marge entre les exemples et l'hyperplan de façon à contrôler le nombre d'erreurs de classification acceptées. En maximisant la marge, elle agit comme tampon qui permet de potentiellement réduire les erreurs de classification pour de nouvelles données pouvant tomber dans cette région. En étudiant l'hyperplan, il est possible de déterminer quels descripteurs sont les plus importants dans la classification. Il est possible d'utiliser le SVM pour effectuer des séparations non linéaires mais, ce faisant, le modèle devient plus complexe et l'interprétabilité s'en trouve affectée. Cette variante est donc peu utilisée en métabolomique pour la recherche de biomarqueurs, mais elle peut toutefois être employée pour la construction de modèles de classifications non interprétables.

Autres applications de l'apprentissage automatique en métabolomique

En plus d'être utilisé pour la construction de modèles de classification permettant la recherche de biomarqueurs, l'apprentissage automatique peut être utilisé pour l'identification des molécules. Comme mentionné précédemment, l'identification des molécules à partir des spectres MS et MS/MS peut se faire en comparant les données expérimentales à celles d'une base de données. Même si ces bases de données sont nombreuses et diversifiées, elles ne comprennent qu'une fraction de l'ensemble des métabolites pouvant être retrouvés en métabolomique. C'est en partie pour cette raison que seule une fraction des spectres observés dans une expérience de métabolomique non ciblée peuvent être liés à une molécule connue [62,63]. Une solution à ce problème est l'utilisation de l'apprentissage automatique. Par exemple, Dührkop *et al.* ont montré qu'en combinant des arbres de fragmentation à un régresseur par SVM pour la prédiction de propriétés moléculaires, il est possible d'identifier des molécules [64]. Van der Hooft *et al.*, quant à eux, utilisent des approches typiquement employées pour le traitement de la langue afin de convertir des spectres de masse pour des molécules connues en un dictionnaire de motifs basés sur les pics. Ce dictionnaire est ensuite utilisé sur les données expérimentales afin de proposer des structures moléculaires. Ces méthodes peuvent aussi proposer des fragments d'information, comme la présence d'un cycle aromatique, sur la structure d'une

molécule non identifiable. Malgré les avancées algorithmiques et la croissance de la taille des bases de données, ces méthodes sont limitées par l'information disponible pour l'entraînement et par l'ensemble des possibilités de prédictions. De ce fait, au lieu de présenter une seule identification potentielle, le résultat est souvent une liste d'identifications et il incombe à l'analyste d'identifier la molécule et de confirmer cette identification par des analyses supplémentaires.

Afin de diminuer la longueur de la liste des identifications potentielles, des prédicteurs de CCS peuvent être employés. LipidCCS et MetCCS sont deux prédicteurs de CCS basés sur des régresseurs à machine de support [65–68]. Ils utilisent une liste de descripteurs moléculaires pour une molécule pour faire leurs prédictions. D'autres emploient des réseaux de neurones profonds pour une tâche similaire [69,70]. Si la CCS prédit est significativement différent de la CCS mesuré pour une molécule inconnue, on peut supposer que l'identification hypothétique était erronée et éliminer cette possibilité de la liste des identifications potentielles. On estime que l'utilisation de la CCS permet de réduire de 60 % la taille de cet espace de recherche [71]. De plus, l'ajout de la CCS lors de l'identification permet d'augmenter le niveau de confiance tel que déterminé par les recommandations de la *Metabolomics Standards Initiative* [49,72].

On peut aussi noter l'utilisation croissante en métabolomique de modèles basés sur les réseaux de neurones profonds, un type d'algorithme d'apprentissage automatique fortement inspiré du fonctionnement du cerveau humain. Ils sont caractérisés par des modèles complexes et peu interprétables, mais souvent très performants. Contrairement aux algorithmes présentés précédemment, les réseaux de neurones profonds sont plus adaptés à l'utilisation de données brutes et moins structurées. Par exemple, les réseaux de neurones à convolution (CNN) sont des modèles qui sont très performants pour l'analyse d'image. De par leur fonctionnement, ils analysent l'exemple en entrée à la recherche d'éléments distinctifs. Dans le cas d'une analyse d'image, l'entrée serait les pixels et les éléments trouvés pourraient être des lignes, des courbes et des formes. Ces éléments sont ensuite compilés et servent à effectuer une classification. Cette recherche d'éléments est résistante aux transformations (c.-à-d. à la translation ou à la rotation) et s'effectue automatiquement par le réseau durant la phase d'apprentissage. Le réseau de neurones est donc responsable de trouver les éléments distinctifs et de comprendre comment ils s'assemblent pour former une image complexe. En appliquant les CNN en métabolomique, Seddiki *et al.* ont résolu

différents problèmes de classification à partir de spectres de masse [73]. Mollerup *et al.* quant à eux, ont utilisé un réseau de neurones pour la prédiction du temps de rétention en LC-MS [70]. Les réseaux de neurones sont très puissants et peuvent être employés pour différentes étapes de l'analyse de données, mais sont limités par leur manque d'interprétabilité.

Une des limitations majeures à l'utilisation d'algorithme d'apprentissage automatique en métabolomique non ciblée est le débalancement entre le nombre d'échantillons et le nombre de métabolites. Comme l'acquisition des données de haute qualité requiert habituellement l'utilisation d'une chromatographie, le nombre d'échantillons pouvant être analysés sur une période acceptable est très limité. Pour cette raison, la métabolomique non ciblée à haut débit, une alternative qui n'emploie pas de séparation, est très intéressante pour produire une grande quantité de données, sur une période limitée.

Métabolomique non ciblée par spectrométrie de masse à haut débit et apprentissage automatique

Le spectromètre de masse est un instrument pouvant acquérir de l'information à une très grande vitesse. Par exemple, un ToF peut acquérir un spectre de masse en moins de 0,1 seconde. Un triple-quadripôle peut acquérir assez de données pour obtenir de l'information quantitative de qualité (une dizaine de mesures de masse) pour deux molécules, accompagnées de leur standard interne, en moins d'une seconde. Étant donné leur vitesse d'acquisition, leur sensibilité et leur capacité à élucider des structures moléculaires, les spectromètres de masses sont tout indiqués pour effectuer des mesures métabolomique à haut débit [74]. En métabolomique par spectrométrie de masse à haut débit on vise l'optimisation des méthodes analytiques et d'analyse de données en tirant profit de la grande vitesse d'acquisition de ces instruments. Dans cette section, nous présenterons les différents défis et approches de cette nouvelle discipline.

D'abord, il est important de mentionner que le terme « haut débit » est utilisé pour caractériser une multitude d'analyses dans différents domaines « omiques ». Si on la compare à la génomique, on peut caractériser la métabolomique standard comme étant très rapide. Effectivement, le séquençage d'un génome prendra plusieurs heures alors que la mesure d'un profil métabolique en employant une approche par LC-MS s'effectue en moins

d'une heure. Cependant, la spectrométrie de masse est lente si on la compare aux lecteurs de plaques utilisés en criblage à haut débit. « Haut débit » est donc un terme très relatif:

> « *Typically in the field of high-throughput screening (HTS), high-throughput is considered 10 000 - 100 000 samples per day. In general, high-throughput in mass spectrometry based methods in metabolomics does not achieve this rate and hence the term 'high-throughput' in metabolomics is more a relative term to describe systems with an improved throughput compared to a standard of traditional liquid chromatography mass spectrometry (LC–MS) methods.* » [75]

Dans le cadre de cette thèse, la métabolomique à haut débit fera référence à l'utilisation de méthodes d'analyses, sans séparation chromatographique, qui permettent d'atteindre une cadence de 1 échantillon ou plus par minute. Cette distinction est nécessaire puisqu'il existe des méthodes chromatographiques à haut débit multiplexant les systèmes de chromatographie connectés à un seul spectromètre de masse afin d'atteindre cette vitesse d'analyse. Ces méthodes sont rapides, mais relèvent davantage de l'optimisation des procédés.

L'avantage principal de la chromatographie est qu'elle permet de séparer les composantes dans un mélange complexe de façon à les introduire séquentiellement dans le spectromètre de masse à une concentration ponctuelle très élevée. On peut donc observer le signal d'un petit nombre de molécules à la fois, séparer des composés isomériques et réduire la contamination du signal des molécules individuelles afin d'obtenir un spectre de haute qualité. Malheureusement, cette séparation se fait au détriment du temps requis pour analyser un échantillon. L'injection directe et l'infusion sont deux méthodes d'introduction d'échantillon qui peuvent être utilisées avec une source d'ionisation ESI et qui permettent une analyse à plus haut débit. En injection directe (ou injection en flux), une pompe envoie du solvant vers la source et un injecteur y envoie de façon régulière un échantillon. L'absence de chromatographie avec ces méthodes, crée un chromatogramme possédant un pic assez large et une queue prolongée. En infusion, l'analyte est d'abord dilué, puis envoyé directement dans la source. On infuse la solution durant une certaine période pendant laquelle on accumule le signal afin de pouvoir bien distinguer le signal provenant des différents métabolites. Pour ces deux méthodes, le facteur de dilution de l'échantillon est critique : on cherche un maximum de sensibilité sans saturer le signal. Ce facteur de dilution s'effectue au détriment des molécules peu abondantes qui se retrouvent près de la

limite de détection. Melo et al. ont montré le potentiel d'une approche par infusion combinée à l'apprentissage automatique pour identifier des marqueurs d'une infection au virus Zika et diagnostiquer les patients infectés [56]. Dans cette étude, du plasma était dilué par un facteur 10 000 avant d'être infusé dans un spectromètre de masse à haute résolution. Cinq spectres allant de 700 Da à 1700 Da étaient acquis par patient : deux réplicas biologiques et trois acquisitions par réplica pour 230 patients. En absence de séparation chromatographique, c'est le spectre de masse seul qui a été utilisé pour la recherche de biomarqueurs en utilisant l'algorithme de forêt aléatoire. Un ensemble de 42 marqueurs a été identifié parmi les 10 000 pics mesurés, dont 17 qui montraient une augmentation significative chez les patients infectés.

Phytronix Technologies, le développeur et manufacturier de la source LDTD, développe des méthodes de métabolomique ciblée à haut débit employant sa technologie. Appliquée principalement aux études toxicologiques, cliniques, environnementales et alimentaires, la technologie LDTD permet l'ionisation d'un échantillon en quelques secondes [76]. Que ce soit pour la quantification de différentes drogues, la détection d'hormones dans la viande de porc ou la quantification de métabolites humains comme le cholestérol, la technologie LDTD permet d'analyser plus de 5 000 échantillons par jour en considérant un débit de 15 secondes par échantillon. Toutefois, il n'a jamais été montré que la source LDTD pouvait être utilisée pour effectuer des analyses de métabolomique non ciblée. La vitesse d'acquisition proposée est donc hautement intéressante pour la production de données en métabolomique pour une analyse par apprentissage automatique puisqu'il serait possible de contrer le débalancement entre le nombre d'exemples et le nombre de descripteurs. Ce faisant, des cohortes de milliers de patients pourraient être analysées en quelques jours dans un laboratoire équipé d'un seul spectromètre de masse. Nous serions donc face à un changement de paradigme où la vitesse d'analyse et le nombre d'échantillons dans une étude métabolomique ne seraient plus des facteurs limitants pour l'investigation de phénotypes complexes et mal définis. Il serait possible de suivre le profil métabolique de milliers d'individus au cours de leur vie et en fonction de leurs problèmes de santé : l'apogée de la médecine personnalisée.

Contenu de la thèse

Dans cette thèse, j'explore le potentiel d'une approche d'analyse du métabolome, de manière non ciblée et à haut débit, utilisant la source LDTD (Figure 0.10). Je propose

différentes approches pour effectuer l'acquisition et l'analyse de données qui visent l'utilisation d'algorithmes d'apprentissage automatique pour la recherche et l'identification de biomarqueurs. Dans le premier, je présente le concept de *Virtual Lock-Mass*. Ces points de masses ont des caractéristiques bien définies et les algorithmes de correction les employant rendent les données produites par LDTD-MS comparables de façon à permettre l'utilisation d'algorithmes d'apprentissage automatique tout en conservant la précision des mesures de masse. Les chapitres 2 et 3 présentent *MetaboDashboard*, un outil d'exploration et de visualisation des résultats d'apprentissage automatique spécialement conçu pour la recherche de biomarqueurs en métabolomique non ciblée. Au chapitre 4, je présente une utilisation de *MetaboDashboard* pour la recherche de biomarqueurs dans le contexte de la détection d'infections virales des voies respiratoires en utilisant des données acquises par LDTD-MS. Cette étude de cas démontre le potentiel de *Metabodashboard* et de la métabolomique non ciblée à haut débit. Je poursuis avec les chapitres 5 et 6 où je présente *DeepCCS*, un outil aidant au processus d'identification des métabolites qui pourrait être particulièrement utile en absence de séparation chromatographique. Finalement, je conclus en discutant des contributions scientifiques présentées dans cette thèse, de leurs impacts sur le champ de recherche de la métabolomique non ciblée, des questions qui sont toujours sans réponse et des directions que pourraient prendre les recherches futures.

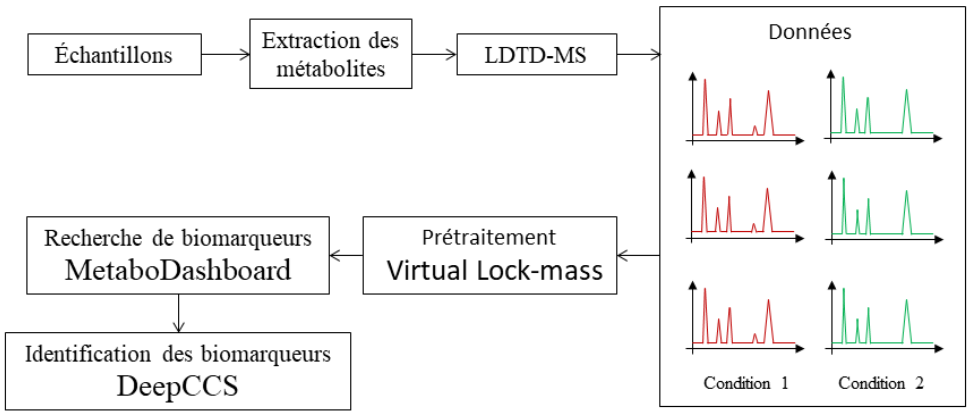


Figure 0.100 Représentation schématique d'une expérience de métabolomique non ciblée par LDTD-MS employant les outils présentés dans cette thèse.

Chapitre 1. Les correcteurs de masse virtuels

1.1 Référence

Ce chapitre est une discussion de la publication portant sur les correcteurs de masse virtuels tel que paru dans Scientific Reports :

Francis Brochu, Pier-Luc Plante, Alexandre Drouin, Dominic Gagnon, Dave Richard, Francine Durocher, Caroline Diorio, Mario Marchand, Jacques Corbeil, François Laviolette; *Mass spectra alignement using virtual lock-masses*; 2019, Scientific Reports; 9-1; DOI : 10.1038/s41598-019-44923-8

Étant donné l'importance de cette publication dans le contexte de travaux de cette thèse, elle peut être retrouvée à l'annexe A.

1.2 Introduction

Le Synapt G2-Si est un spectromètre de masse à haute résolution de type Q-ToF commercialisé par Waters Corporation et a été utilisé dans la majorité des travaux de recherches de cette thèse. Comme tout instrument utilisant une chambre de temps de vol, la mesure de masse de cet appareil est affectée par les changements de température ambiante puisque celle-ci a un impact sur les dimensions de la chambre de temps de vol. Afin d'assurer une mesure précise de la masse sur cet instrument, il est recommandé d'effectuer une calibration de façon hebdomadaire ou quotidienne selon les conditions de l'instrument. Cette calibration permet de corriger toute variation systémique affectant l'instrument. Entre les calibrations, le système d'acquisition de données utilise un correcteur de masse (*lock-mass)* afin de corriger les mesures de masses. La majorité des sources d'ionisations fournies par la compagnie sont équipées d'un électronébuliseur permettant d'infuser cette solution contrôle en alternance avec les analytes dans l'instrument. Comme la masse du composé contenu dans la solution contrôle est connue, un facteur de correction pour être calculé et appliqué aux analytes mesurés au même moment. Ce système de correction permet de réduire au maximum les variations entre le m/z réel et le m/z mesuré. La combinaison de la calibration hebdomadaire et du *lock-mass* en temps réel assure une performance optimale de l'instrument.

L'utilisation d'une source d'ionisation ne permettant pas l'infusion d'une solution de correcteur de masse durant l'acquisition des données empêche d'effectuer la correction du m/z. Cette limitation diminue l'exactitude des mesures de masse. Cette problématique affecte particulièrement les sources d'ionisation et méthodes d'introduction d'échantillons à haut débit comme le LDTD et le DESI lorsqu'employées pour des analyses non ciblées. Une solution alternative est d'ajouter, à même l'échantillon, un analyte contrôle pouvant être utilisé pour effectuer la correction de masse. C'est ce qu'on appelle un correcteur de masse interne. Comme la problématique d'instabilité de mesures du m/z affecte principalement les données acquises sans séparation, le composé contrôle est analysé durant toute l'acquisition au même titre que les analytes. Ce faisant, une correction peut être appliquée. Toutefois, l'ajout d'un composé dans une matrice complexe lors d'une analyse non ciblée sans séparation comporte des risques. Par exemple, le signal du nouveau composé peut être indistinguable du signal provenant des composés présents naturellement dans l'échantillon. L'ajout d'un composé qui doit avoir un signal suffisamment intense pour permettre une correction de qualité pourrait aussi affecter l'ionisation des autres molécules. Finalement, le composé utilisé doit aussi être compatible avec la source d'ionisation. En considérant ces problèmes et contraintes, il devient difficile de choisir un ou des composés pouvant servir de correcteur de masse interne en absence de chromatographie.

Comme expliqué précédemment, les spectres de masse doivent être alignés sur l'axe m/z afin d'effectuer une recherche de biomarqueurs. En absence de correcteur de masse, l'alignement des spectres est d'autant plus complexe, car la variation observée est plus grande. De cette problématique est venue la nécessité de développer une méthode de correction afin de réduire les variations systémiques et donc de permettre un alignement de qualité. Le développement de cet outil était nécessaire avant de poursuivre les expérimentations ayant comme objectif le développement d'une approche de métabolomique non ciblée par LDTD-MS.

Comme expliqué plus tôt, le principe du correcteur de masse interne est d'utiliser un composé au m/z connu ajouté à l'ensemble des échantillons afin de faire la correction. En calculant la déviation entre la masse théorique et mesurée du composé, on peut déterminer un facteur de correction. Lors d'une analyse de métabolomique non ciblée, les échantillons proviennent tous d'une même matrice biologique. En raison de l'homéostasie naturelle présente dans un système biologique tel que le plasma, certains métabolites sont présents

dans l'ensemble des échantillons d'une expérience. Des ions provenant du solvant ou du gaz porteur, encore ici identique d'une expérience à l'autre, sont aussi toujours présents dans le spectre mesuré pour chaque échantillon. Cette collection d'ions peut donc être utilisée afin de déterminer les variations observées dans un groupe de spectres donnés. Le premier défi est donc d'identifier un ou plusieurs pics produits par ces molécules communes. Lors des travaux initiaux, nous avons remarqué assez rapidement qu'il pouvait être complexe d'identifier ces pics dans plusieurs dizaines d'échantillons possédant chacun plus de 10 000 pics de masse. Après plusieurs heures de recherche manuelle, il est devenu clair que certains critères de recherche sont importants :

- Le signal du pic partagé doit être facilement distinguable du signal des pics voisins de façon à éliminer les erreurs d'identification.
- Le signal du pic doit être assez élevé de façon à observer une bonne précision sur la masse.

Du besoin de formaliser et d'automatiser le processus de sélection des pics sont nés les *Virtual Lock-Mass* (VLM). Je présente d'abord la définition des VLM en m'appuyant sur des concepts de spectrométrie de masse. Par la suite, je discute de l'optimisation de la taille de fenêtre de recherche de pics, le paramètre principal de l'algorithme. Finalement, je conclue en discutant de l'impact des VLM dans le processus d'analyse de données en métabolomique non ciblée par spectrométrie de masse et des améliorations futures.

1.3 Définition intuitive et formelle d'un VLM

L'objectif principal des VLM est de rendre les spectres le plus comparable possible avant d'effectuer des analyses employant des algorithmes d'apprentissage automatique. Ils ont aussi été créés dans un contexte très particulier: des analyses non-ciblées en absence de séparation chromatographique et employant un spectromètre à haute résolution de type QToF. Dans ce contexte, une première définition des VLM fut proposée : *un groupe de pics homologues facilement reconnaissables dans un type d'échantillon donné*. À partir de cette définition et du contexte analytique, on peut dériver un ensemble de contraintes s'appliquant aux VLM. D'abord, les pics homologues doivent être dans une fenêtre de masse assez restreinte puisque nous utilisons un spectromètre de masse à haute résolution. Ensuite, puisqu'ils proviennent d'un type d'échantillon particulier (ex. : du plasma), ces pics devraient être présents dans tous les échantillons d'une cohorte. Puisque les échantillons analysés sont de nature biologique, il est raisonnable de croire qu'un grand nombre de métabolites sont présents dans tous les échantillons de par l'homéostasie naturelle présente dans ce

type d'échantillons. Finalement, le critère de reconnaissabilité peut être lié à deux contraintes : les pics doivent avoir une intensité plus importante que les autres pics dans une fenêtre de masse et ils doivent pouvoir être identifiés sans ambiguïté. En plus de servir à la reconnaissabilité, la notion d'intensité augmente la précision des mesures de masses des pics qui seront employés dans le processus de correction. Effectivement, un pic plus intense signifie qu'une plus grande quantité d'ion correspondant à une molécule ont frappé le détecteur. Puisque chaque ion frappant le détecteur est une mesure expérimentale, plus il y a de points de mesure, plus la précision de la mesure est importante.

La définition intuitive des VLM est basée sur un processus d'exploration des données ainsi que sur les principes de fonctionnement d'un spectromètre de masse. Afin de faire passer les VLM d'un concept intuitif à un algorithme, il était nécessaire de formaliser la définition intuitive présentée précédemment. La définition formelle des VLM est donc la suivante :

Définition 1.1 Définition formelle d'un correcteur de masse virtuel (inspiré de Brochu et al. 2019)

Considérons $\mathcal{S} \stackrel{\text{def}}{=} \{S_1, \ldots, S_m\}$ comme un ensemble de spectres de masse. Chaque spectre S_i est une séquence de pics, où chaque pic est une paire (μ, ι) possédant une valeur de m/z μ et une intensité ι. Considérons deux seuils d'intensité t_a et t_b tels que $t_a < t_b$. Considérons une fenêtre de taille $2w$ centrée sur le pic (μ, ι) et définie comme un intervalle qui débute à $\mu \cdot (1 - w)$ et qui se termine à $\mu \cdot (1 + w)$. Notons que la taille de la fenêtre w est relative à la valeur de m/z μ.

Pour un ensemble de spectre $\mathcal{S}$ et une fenêtre de taille w, un correcteur de masse virtuel (VLM) par rapport à $(\mathcal{S}, w)$ est un point v sur l'axe m/z tel qu'il existe un ensemble de pics $\mathcal{P}$ provenant de $\mathcal{S}$ qui satisfait les propriétés suivantes :

1. $\mathcal{P}$ contient exactement un pic de chaque spectre de $\mathcal{S}$.
2. La moyenne des valeurs de m/z des pics de $\mathcal{P}$ est égale à v.
3. Chaque pic de $\mathcal{P}$ est situé dans l'intervalle $[v(1 - w), v(1 + w)]$.
4. Il n'existe aucun autre pic dans $\mathcal{S}$ qui fait partie de l'intervalle $[v(1 - w), v(1 + w)]$.
5. Tous les pics dans $\mathcal{P}$ ont une intensité dans l'intervalle $[t_a, t_b]$.

Si et seulement si tous ces critères sont satisfaits, nous pouvons statuer que $\mathcal{P}$ est un ensemble de pics associés au VLM v.

À partir de cette définition, un algorithme de recherche des points de VLM fut développé et une implémentation efficace fut produite tel que décrit dans l'article par Brochu *et al.* présent dans l'annexe A.

1.4 Optimisation de la taille de fenêtre

Le paramètre w , correspondant à la taille de fenêtre dans laquelle les pics de $\mathcal{P}$ doivent se trouver, est le principal paramètre de l'algorithme de recherche des VLM. Ce dernier a un impact important sur le nombre de points de VLM trouvés et donc, sur l'efficacité de la correction effectuée par l'algorithme.

Afin de permettre l'optimisation du paramètre w, nous avons développé une approche basée sur la maximisation du nombre de points de VLM. Pour une série de tailles de fenêtre (w) données, la fonction d'optimisation compte le nombre de points de VLM pouvant être possiblement obtenus pour chaque valeur. Avec l'augmentation de la taille de la fenêtre, le nombre de VLM trouvés augmente puisque nous étendons l'espace de recherche et donc la tolérance aux variations systémiques et aux erreurs de mesures de l'instrument. Cette augmentation se poursuit jusqu'à atteindre un maximum. Par la suite, une décroissance est observée alors que l'augmentation de la taille de fenêtre entraîne la contamination de $\mathcal{P}$ par des pics provenant de spectres de $\mathcal{S}$ ayant déjà un pic dans $\mathcal{P}$. La figure 1.1 montre le processus d'optimisation de w sur le jeu de données Malaria produit dans le cadre de l'article inspirant ce chapitre. Comme expliqué, le nombre de points VLM augmente avec la taille de la fenêtre jusqu'à atteindre un maximum à 30 ppm. Ensuite, le nombre de pics de VLM diminue puisque $\mathcal{P}$ ne contient plus un seul pic par spectre de $\mathcal{S}$.

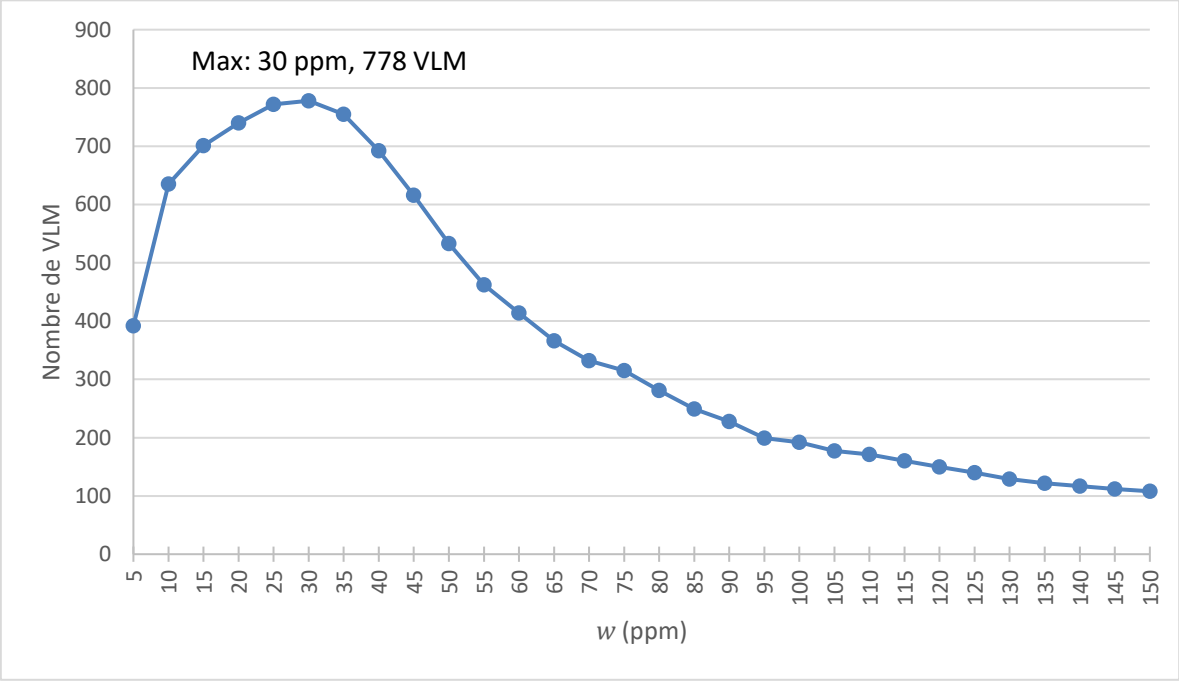


Figure 1.1 Variation du nombre de points de VLM en fonction de w. La valeur de w correspondant au maximum du nombre de points de VLM est la valeur optimale de w.

Une déviation globale de 30 ppm sur la mesure de masse peut sembler élevée pour un instrument à haute résolution effectuant une acquisition sur une très courte période de temps (une seule plaque d'analyse (Lazwell), 80 échantillons en environ 30 minutes au total). Toutefois, cette variation est en partie liée à la résolution de l'instrument. Le pouvoir

de résolution d'un spectromètre de masse est défini comme l'évaluation quantitative de sa capacité à distinguer deux molécules ayant un m/z rapproché et il se calcule selon l'équation 1.1 où R est la résolution, m est la masse et Δm est la différence de masse entre deux points rapprochés distincts.

$$R = \frac{m}{\Delta m}$$

Équation 1.1 Définition du pouvoir de résolution d'un spectromètre de masse

La valeur de Δm est typiquement remplacée par la largeur d'un pic à 50% de sa hauteur afin de calculer R à partir d'un seul pic de masse. Selon l'équation, on peut comprendre que pour une résolution donnée, Δm augmente linéairement avec la masse. La résolution est donc une métrique relative, au même titre que la distance entre deux pics (D) définie dans l'équation I.1. En combinant l'équation 0.1 qui définit la mesure de distance relative en ppm entre deux points de masse et l'équation 1.1 ci-haut, on peut démontrer la relation entre la résolution (R) et la distance minimale requise entre deux pics de masses similaires afin qu'ils soient distincts (D_{min}):

$$D_{min} = \frac{\Delta m}{m} * 10^6$$

$$D_{min} = \frac{m}{m * \ R} * 10^6$$

$$D_{min} = \frac{1}{R} * 10^6$$

Équation 1.2 Calcul du pouvoir de séparation d'un spectromètre de masse

Cette relation peut être validée *in silico* par la simulation présentée en figure 1.2 qui montre l'effet de la résolution sur la séparation de deux composés, de la vitamine D3 ($[C_{27}H_{44}O+H]^+$, m/z=385,34649) et du monoacylglycerol ($[C_{23}H_{44}O_4+H]^+$, m/z=385,33124), en proportion identiques. Ces deux ions sont séparés par 0,01525 Da ou 40 ppm. Au autant que la résolution reste inférieure à 25 000x, les deux pics sont indistinguables. Une fois cette valeur dépassée, la séparation est visible et les algorithmes de conversion du signal continu en pics centroïdes détectent bien les deux composés.

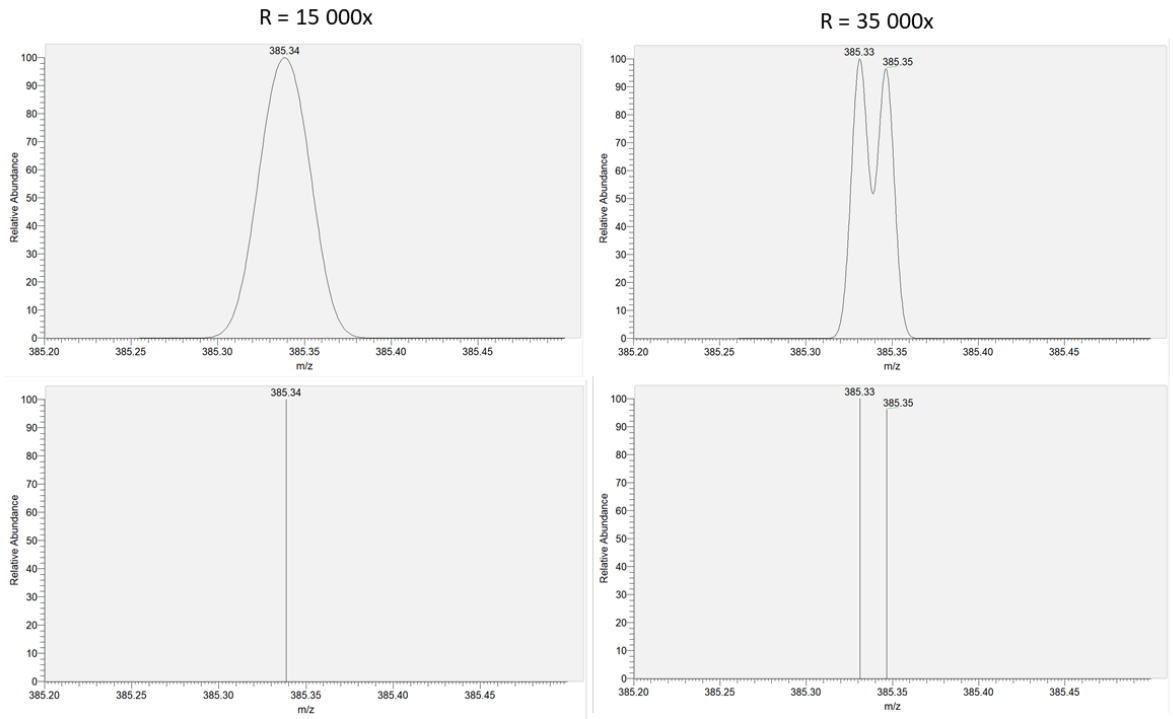


Figure 1.2 Simulation de spectres de masse montrant la séparation de la vitamine D3 ($[C_{27}H_{44}O+H]^+$, m/z=385,34649) et d'un monoacylglycerol ($[C_{23}H_{44}O_4+H]^+$, m/z=385.33124), en proportion identiques à une résolution de 15 000x et 35 000x. Le haut montre les pics en signal continu et le bas le résultat en centroïde. Les deux molécules, sont indistinguables à une résolution de 15 000x même s'ils n'ont pas le même m/z.

En fonction de cette analyse, on peut conclure que la valeur optimale de w de 30 ppm est directement influencée par la résolution de l'instrument. Effectivement, comme la résolution de l'instrument lors de l'acquisition se situait entre 30 000x et 35 000x, deux pics doivent être séparés par une distance approximative de 33 ppm et 28 ppm afin d'être séparés. Donc, l'ensemble des pics dans une fenêtre de cette taille ne peuvent être distingués par le spectromètre de masse et il est impossible de trouver deux pics provenant d'un même spectre de $\mathcal{S}$ dans cette fenêtre. Donc, tant que la limite de taille de fenêtre de w n'atteint pas la valeur maximale définie par la résolution, il est normal de regrouper des pics ensembles puisque l'algorithme et l'instrument sont incapable de les séparer, qu'ils proviennent de la même molécule ou non. En se basant sur les capacités de l'instrument, la valeur minimale de w serait donc calculable à partir de l'équation 1.2. On peut donc dire que $w_{optimal} \geq \frac{1}{R} * 10^6$. En absence de variations systémiques, w optimal peut être estimé à partir de la résolution de l'instrument alors qu'en présence de variation systémique, w optimal sera plus élevé que $\frac{1}{R} * 10^6$.

1.5 Discussion

L'algorithme développé ici est la première étape d'un long chemin menant au développement d'une approche de métabolomique à haut débit par spectrométrie de masse. Puisque nous étions les pionniers dans le développement d'une approche de métabolomique non ciblée par LDTD-MS, nous avons commencé par des essais simples. Par exemple, nous avons optimisé le processus d'extraction métabolique et les paramètres d'acquisition pour produire un jeu de données de plasma où nous avons manuellement ajouté de l'acétaminophène et du clomifène. La recherche de biomarqueurs pour un tel jeu de données était simple et il a rendu possible le développement d'approches d'apprentissage automatique. Dès nos premières tentatives à produire une preuve de concept, nous avons observé qu'il était impossible de tirer des conclusions tant que les données produites ne pouvaient être comparées. Bien que des approches de *binning*, consistant à séparer l'axe m/z en sections égales, aient déjà été suggérées et employées par d'autres groupes, elles allaient à l'encontre de nos objectifs [77]. Un spectromètre de masse à haute résolution permet une mesure de masse très précise au dix millième, mais le *binning* réduit cette précision au dixième ou centième d'unité de masse selon les paramètres employés. Cette approche vient donc potentiellement regrouper dans un seul *bin* les mesures appartenant à différentes molécules et entraîne une perte de la précision de masse associée aux mesures expérimentales à haute résolution.

Il est important de rappeler que la problématique à laquelle répondent les VLM est très spécifique. D'abord, les instruments à temps de vol sont parmi les plus susceptibles à souffrir de variations de masse assez importantes sur de courtes périodes. De ces instruments, celui employé pour l'ensemble de nos expériences (Waters Synapt G2-Si), utilise un correcteur de masse externe qui est infusé par le biais d'un deuxième électronébuliseur présent dans certaines sources d'ionisation. En changeant la source de l'instrument pour un LDTD, un DESI ou même un REIMS, il devient impossible d'acquérir le signal du correcteur de masse en temps réel sans affecter le signal des analytes. Avec d'autres instruments à haute résolution comme un Thermo Fusion, un correcteur de masse peut être employé, mais indépendamment de la source d'ionisation. Ce faisant, un changement de source d'ionisation n'empêche pas l'utilisation du correcteur de masse. La problématique à laquelle répond le VLM est aussi spécifique au contexte de métabolomique non ciblée effectuée sans séparation chromatographique. En absence de chromatographie, l'utilisation d'une ou de plusieurs molécules standards pourrait contaminer le signal mesuré.

Il est donc risqué d'employer cette méthode. Les VLM résolvent ces problèmes puisqu'ils proviennent des échantillons.

L'introduction des VLM a nécessité un certain changement de mentalité dans l'interprétation des résultats. Contrairement au correcteur de masse traditionnel, les VLM effectuent une correction virtuelle qui ne rapproche pas nécessairement les mesures expérimentales du m/z réel. Il n'y a donc aucune garantie que la correction augmentera l'exactitude des mesures de masse. Par contre, les VLM augmentent significativement la précision des mesures comparativement au *binning*. Cet effet croit avec la résolution de masse employée lors de l'acquisition des données. L'utilisation des VLM permet donc d'effectuer la recherche de biomarqueurs, même en l'absence d'une mesure de masse exacte, sur des données produites avec un spectromètre de masse à haute résolution. Afin d'effectuer une mesure de qualité permettant la recherche dans les bases de données, une seconde acquisition avec un correcteur de masse interne est nécessaire. Puisque la recherche de biomarqueurs est déjà complétée, l'ajout d'un correcteur de masse interne peut s'effectuer sans risque de contaminer le signal de la molécule d'intérêt dont la masse approximative sera connue.

Un aspect des VLM qui a très peu été exploré est leur utilisation pour la normalisation des intensités. Calculer un facteur de correction à partir du signal de composés qui sont partagés par l'ensemble des échantillons est une méthode de normalisation reconnue [46]. Il serait possible de calculer la médiane des variations d'intensité sur les pics de VLM et de trouver un facteur de correction générique pouvant être appliqué à l'ensemble des ions. Les jeux de données employés dans l'article seraient tout à fait pertinents pour explorer cette hypothèse. Par exemple, avec le jeu de données « Acétaminophène », on chercherait à étudier les variations sur le signal du composé pharmaceutique avant et après la normalisation, ce dernier étant en concentration égale dans l'ensemble des échantillons.

Finalement, les algorithmes développés pour détecter les VLM pourraient être utilisés dans d'autres domaines. La détection de VLM ou de points d'alignement est très similaire à un regroupement (*clustering*) dans un espace à une seule dimension. Toutefois, les algorithmes développés profitent des caractéristiques de l'espace unidimensionnel pour simplifier la complexité computationnelle à $O(n \log m)$ où m est le nombre de spectres et n est le nombre de pics [78]. Par exemple, on pourrait s'inspirer de l'approche développée ici pour la recherche de sites de liaison en ChIP-seq ou la recherche de regroupement

d’événements sur une échelle temporelle à condition d’avoir une mesure de distance appropriée.

Chapitre 2. Présentation du premier article

2.1 Référence

MetaboDashboard: simplified machine learning in metabolomics
Pier-Luc Plante, Francis Brière, Nancy Boucher, Élina Francovic-Fontaine, Didier Brassard, Benoit Lamarche, Jacques Corbeil
Soumis à Bioinformatics

2.2 Contexte

Le chapitre précédent présentait le concept des correcteurs de masse virtuels ainsi qu'un outil afin de rendre comparables un ensemble de spectres. L'objectif non exclusif de cette transformation était de permettre l'utilisation d'algorithmes d'apprentissage automatique pour la construction de modèles de classification et la recherche de biomarqueurs. Il a effectivement été montré que l'utilisation des correcteurs de masse virtuels permettait de réduire les erreurs de classification et diminuait la complexité des modèles de classifications construits par les algorithmes d'apprentissage automatique.

Les algorithmes d'apprentissage sont très utiles en métabolomique pour la recherche de biomarqueurs [52,53,79,80]. Malheureusement, leur utilisation est peu répandue puisqu'elle requiert des connaissances en programmation et en analyse de données qui ne sont pas acquises par une majorité de chercheurs dans le domaine. Le chapitre 4 présente MetaboDashboard, un outil visant à la démocratisation de l'apprentissage automatique pour la recherche de biomarqueurs en métabolomique par spectrométrie de masse ainsi que la simplification de la diffusion de résultats de ce domaine. Le chapitre 5, quant à lui, présente l'utilisation du MetaboDashboard pour la recherche de biomarqueurs en métabolomique à haut débit appliquée au contexte de l'influenza.

2.3 Contribution

Je suis responsable de la conception initiale du projet, de la construction de l'application et de l'écriture du manuscrit. Tous les auteurs ont participé à la conception de l'expérience de métabolomique, discuté des fonctionnalités de l'application et révisé le manuscrit.

2.4 Discussion

MetaboDashboard est le résultat de 5 ans de développement et d'évolution. Les premières itérations étaient constituées d'un ensemble de scripts permettant de simplifier l'analyse de données de façon à éviter les tâches répétitives. Ces scripts nécessitaient tout de même une certaine connaissance de la programmation pour en tirer le plein potentiel. Puis est venu un *Jupyter Notebook,* qui permettait d'effectuer rapidement et simplement des analyses exploratoires d'apprentissage automatique. À partir de ce moment, les chercheurs en spectrométrie de masse étaient un peu plus à l'aise d'utiliser ces approches d'apprentissage automatique. Grâce au *Notebook*, la majorité du code était préparé et il ne restait qu'à l'exécuter afin d'obtenir des statistiques de performances, des tableaux et des figures. Le désavantage du *Notebook* était qu'il n'était pas conçu pour permettre une reproduction exacte des analyses. De plus, il était facile de faire des erreurs, seules quelques visualisations de base étaient possibles et la configuration des expériences n'était pas centralisée. Finalement est venu le *MetaboDashboard*. Ce dernier a été développé en partenariat avec des chercheurs opérant les spectromètres de masse et effectuant des analyses de données ainsi que des spécialistes en nutrition. Ces derniers étaient intéressés beaucoup plus par les résultats et conclusions des analyses. Il était donc nécessaire de mettre une emphase particulière sur la transmission de l'information et non pas seulement sur la simplification de l'analyse des données. Cette équipe représente bien l'équipe multidisciplinaire typique : des spécialistes de plusieurs domaines qui essaient de parler le même langage. En évaluant les besoins et les demandes de chacun, nous en sommes venus à produire un outil polyvalent qui permet de produire et d'explorer les résultats d'algorithmes d'apprentissage automatique appliqués à la métabolomique.

Afin de permettre son utilisation, *MetaboDashboard* doit être installé. Cette étape peut être complexe pour un utilisateur ayant peu de connaissances des outils en ligne de commande. La mise en place d'un serveur permettant de tout effectuer par le biais d'une page web fournissant une interface graphique éliminerait ce problème. Par contre, le calcul des modèles pourrait devenir computationnellement lourd pour un serveur partagé par plusieurs dizaines d'utilisateurs. La construction d'un conteneur Docker, offrant une interface graphique pour la configuration des expériences, pourrait être un bon compromis pour la simplicité d'installation. Ce type d'installation pourrait être local et limiter les problèmes d'installation.

Actuellement, l'interprétabilité des modèles produits par *MetaboDashboard* est seulement possible pour les algorithmes offrant une solution relativement parcimonieuse et interprétable (arbre de décision, forêt aléatoire, SVM Linéaire, SCM). Ces algorithmes permettent donc d'effectuer la recherche de biomarqueurs. L'utilisation d'algorithmes produisant des modèles plus complexes est possible, mais ne permet pas la recherche de marqueurs. Des approches comme la permutation d'éléments et le calcul des valeurs de Shapley pourraient permettre d'ajouter une couche d'interprétabilité à ces modèles nativement non interprétables [81,82]. Toutefois, ces approches peuvent devenir computationnellement lourdes. Il serait plus pertinent d'ajouter ce type d'approche à une expérience seulement lorsque les performances de classification atteignent le niveau désiré. Il est donc souhaitable de s'interroger sur l'intérêt d'ajouter ce type d'analyse directement dans *MetaboDashboard* au lieu de les confier à un chercheur qualifié.

L'apprentissage automatique appliqué à la recherche de biomarqueur en métabolomique n'est pas très répandu. Contrairement à l'utilisation directe d'un langage de programmation tel que R ou Python, MetaboDashboard permet d'effectuer l'apprentissage avec très peu de code. De plus, il permet de garder une trace complète de ce qui a été effectué. MetaboAnalyst [83] est probablement l'outil le plus utilisé afin d'employer des approches d'apprentissage automatique en métabolomique. Il ne permet toutefois pas un contrôle complet sur la sélection des hyperparamètres et ne permet pas de publier l'ensemble des résultats sur une page web interactive afin de partager les résultats avec des collègues.

MetaboDashboard est donc unique en son genre et répond à plusieurs problèmes qui limitaient l'utilisation d'algorithmes d'apprentissage automatique appliqués au domaine de la métabolomique pour la recherche de biomarqueurs. Un cas d'utilisation est présenté au Chapitre 4.

Chapitre 3. MetaboDashboard: simplified machine learning for metabolomics

3.1 Résumé

L'objectif de la recherche de biomarqueurs en métabolomique est de construire un modèle de prédiction qui utilise un petit nombre de composés et qui peut ensuite être utilisé pour classifier de nouveaux échantillons dans leur groupe approprié (ex. : sain ou malade). Afin de produire de tel modèle de prédiction, l'apprentissage automatique supervisé est l'outil le plus approprié. Un des facteurs limitant l'utilisation de l'apprentissage automatique en métabolomique est la grande quantité de nouvelles connaissances requises. Nous proposons ici un nouvel outil, appelé MetaboDashboard, qui simplifie grandement l'utilisation d'algorithmes d'apprentissage automatique dans le contexte de la métabolomique non ciblée. De plus, l'outil favorise l'analyse de données reproductible et transparente, en concordance avec les principes FAIR pour l'analyse de données.

3.2 Introduction

One of the primary objective of biomarker discovery in metabolomics is to create a model, based upon a small number of compounds, that can predict or classify a specific phenotype reliably [52]. In order to produce such predictive or classifying model, supervised machine learning algorithms are the most appropriate instruments. Depending on the algorithm used, the predictive model can range from sparse and interpretable to complex and uninterpretable whilst still providing a statistically verifiable answer. In most context, such as health sciences, sparse and interpretable models should be privileged for acceptability since the process can be understood by field experts. Furthermore, biological interpretation of the model can be made if one knows which metabolites are used. Unfortunately, one of the limiting factors to the use of these powerful machine learning algorithms for biomarker discovery in metabolomics is the steep learning curve of machine learning and programing languages for researchers new to the field. Instrument manufacturers are also presently not favoring open-source analytical software for data analyses. Web applications such as Metaboanalyst offers to upload metabolomics results online to build machine learning models [83,84]. Even though these tools are useful because they require no local computing power, they only allow the use of limited and standard algorithms (i.e. Random Forest and

SVM). Machine learning approach in metabolomics, due to the fat-data dimension of the input (number of samples is a lot smaller than the number of variables), requires rigorous cross-validation for hyper-parameters selection and Monte-Carlo cross-validation for approach validation [52]. These parameters can be influenced by dataset and/or domain knowledge. Most metabolomics data analysis web servers offer little to no control over hyper-parameters and cross-validations, which is an important limitation. We propose a new tool, MetaboDashboard, that enables the use of a larger array of machine learning models in the context of metabolomics and addresses these limitations. Furthermore, it promotes reproducible and transparent data analyses, in line with FAIR Data Principles in that the one can find, access and reproduce the results [85]. Source code and installation instructions are available at https://github.com/plpla/MetaboDashboard.

3.3 Implementation and feature

MetaboDashboard is coded in Python3 and relies heavily on Scikit-learn and Dash. Basic usage requires using very simple command lines but deep customization and control over each step is available through appropriate configuration files and addition to the code base. MetaboDashboard separates the analysis in three independent steps: data preparation, machine learning model generation and models evaluation.

In the data preparation step, the user must provide an input file containing metabolomics data measurements. File reader for standard tabular format (xlsx, csv, tsv) are available. Files output from most metabolomics data analysis pipeline (i.e. Progenesis QI) can also be used with little to no modifications. This enables the use of MetaboDashboard in a very straight forward fashion. Using the user specified sample groups, MetaboDashboard creates different experimental designs that uses all or a subset of the samples. This can be highly interesting when an experiment contains more that two classes, some of which can be grouped together. Data cleaning and normalisation functions can be enabled in order to remove features with excessive variations or missing values. Finally, the dataset is randomly split between a training set and a test set. These splits are specific to an experimental design. Since Monte-Carlo cross validation (MMCV) requires evaluating model performances on different train-test sets, MetaboDashboard can create as many splits as required by the user for a specific experimental design. Each split is saved on disk to have reproducible results.

The next step is performing the machine learning model generation. The user can provide a list of algorithms for specific experimentation. These algorithms can be sparse and/or interpretable such as Decision Trees, Random Forest or LinearSVM with L1 regularization. Using algorithms that produce sparse and interpretable models enables the extraction of decision rules and/or features that are important in the decision process. The use of more complex algorithms such as non-linear SVM or neural network is also possible, at the expense of interpretability and sparsity. Custom algorithms can also be included if they follow Scikit-learn *BaseClassifier* class definition. Since it is considered a standard in the machine learning community, it makes MetaboDashboard easily upgradable. All the algorithms specified will be executed on the data splits generated at the previous step, making models performance highly comparable since they are trained using exactly the same data. Machine learning can become computing intensive with Monte-Carlo cross validation and *k*-fold cross-validation. For this reason, MetaboDashboard can perform model training in parallel to obtain results faster. Once again, in order to keep track of the analysis process, each model and cross-validation process is saved.

Finally, the models' evaluation step can be performed. It is at this step that the dashboard becomes highly attractive as it allows any user without any machine learning knowledge or coding skills to fully explore the experiments results. This is done through a web page that reads the configuration file, the data splits and machine learning results. It presents the data in a simple and efficient manner, allowing to evaluate algorithms globally (on all Monte-Carlo splits of an experiment) or individually on a single split. When using sparse models, a feature selection function extracts the important information from the models. This extraction can be performed using the feature selection metric on a single model or by averaging it over all these features. A box plot showing the abundance distribution across the different groups can also be displayed. Comparison of the models and algorithms is based on a list of metrics, that can be customized: AUC score, accuracy, f1-score and other metrics are computed automatically. Finally, PCA, t-SNE and UMAP representation of the data can be computed with or without filtering features. This allows a user to generate a 2D representation of the dataset, unveiling global tendencies.

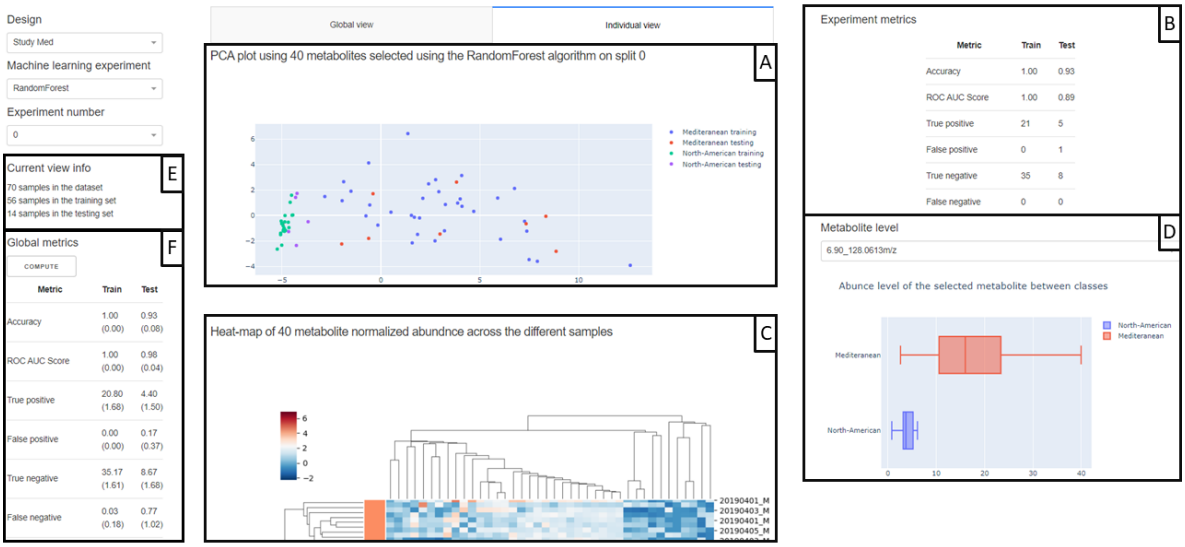


Figure 3.1 Screen capture of MetaboDashboard used in the context of a nutritional study. ***A****.The web application shows a PCA plot built using 40 metabolites selected by a Random Forest model.* ***B*** *The different prediction metrics for the selected experiment. Additionnal metrics can be enabled in the configuration file.* ***C*** *Heat-map with clustering showing the abundance of the selected metabolites across the samples.* ***D*** *Box-plot of the abundance of a single metabolite.* ***E*** *Design information* ***F*** *Average of the different metrics across all experiments for a single design and algorithm*

MetaboDashboard is a tool to leverage the power of machine learning in the field of metabolomics. It enables performing simple to complex experiments, with complete control over the whole process. Once completed, the web page can be hosted as a complement to a publication or simply put on a code sharing platform with the associated data. Source code, installation instructions and demo data are available at https://github.com/plpla/MetaboDashboard.

Chapitre 4. Détection d'infection virale des voies respiratoires par métabolomique à haut débit et apprentissage automatique

4.1 Introduction

Les infections respiratoires représentent une portion significative de la mortalité et de la morbidité de la population canadienne. De ces infections, le virus de l'influenza était la plus importante cause de consultation pour une infection respiratoire aiguë avant 2020 puisque la COVD-19 est survenu lors de cette année [86]. Santé Canada relève plus de 12 200 hospitalisations et plus de 3500 morts causées par l'influenza chaque année [87]. À l'échelle planétaire, le réseau FluNet de l'Organisation mondiale de la Santé a estimé en 2019 plus de 99 000 le nombre de cas d'influenza s'ayant soldés par la mort [86]. Ce virus à ARN, qui affecte une majorité de la planète, peut causer des complications chez les personnes vulnérables comme les personnes âgées de plus de 65 ans, les enfants de moins de 60 mois et les personnes souffrant de maladies chroniques [87]. Ses complications incluent des pneumonies bactériennes qui accroissent considérablement les risques de mortalité et représente à la fois un fardeau financier et humain pour le système de santé.

La métabolomique est l'étude des petites molécules contenues dans un échantillon biologique. Lors d'une infection virale, le corps humain effectue une réponse immunitaire. Une multitude de métabolites sont produits et transformés lors de cette réponse et à cela s'ajoutent ceux causés par le processus infectieux. Certains sont génériques au processus inflammatoire, comme l'acide arachidonique, alors que d'autres sont spécifiques à l'infection en cours[88–91]. De plus, chaque infection cause une panoplie de symptômes qui sont plus ou moins spécifiques : température, congestion, toux, etc. Ces symptômes font partie de la réponse de l'organisme et sont accompagnés de changements métaboliques aux niveaux tissulaire et systémique.

Les tests diagnostiques pour l'influenza recommandés par le CDC (*Centers for Disease Control and Prevention*) s'effectuent par un RT-PCR visant à amplifier une section de l'ARN viral. En cas d'amplification, la présence de ce dernier et l'infection sont confirmées avec une excellente spécificité et sensibilité. Ce test est relativement simple et rapide, les

résultats étant obtenus en moins de 90 minutes [92]. L'échantillon, un écouvillon nasopharyngé, est récolté par l'insertion d'une brosse d'environ 10 cm de long dans la cavité nasale des patients. Cet écouvillon récolte une partie de ce qui est présent dans les voies respiratoires : mucus, cellules humaines et virus. Le tout est inoculé dans un milieu de transport universel pour échantillon viral (*Universal Transport Media* - UTM). Les résultats de ce test, combinés à l'évaluation des symptômes du patient, permettent au médecin d'établir un diagnostic. La précision peut varier selon la souche du virus et le test employé, mais ils sont reconnus comme produisant très peu de faux positifs et de faux négatifs. De nouveaux tests diagnostiques très rapides et à faible coût basés sur des techniques d'immunochimie telles que l'immunochromatographie et les essais à enzyme immunoadsorbants ont été développés. Toutefois, même s'ils sont plus rapides, leur sensibilité et leur spécificité sont inférieures aux tests basés sur l'ARN puisque les anticorps employés sont spécifiques pour les souches circulantes. Ils pourraient donc être inaptes à détecter une nouvelle souche. Leur utilisation n'est donc pas recommandée par le CDC.

Lors de l'infection de souris par le virus de l'influenza, Chandler et al. ont montré que différents chemins métaboliques au niveau du tissu pulmonaire étaient significativement affectés comparativement à ceux de sujets non infectés [93]. Ils ont notamment relevé les voies liées au métabolisme de nombreux acides aminés ainsi que celles associées au métabolisme des vitamines B3, B6 et B2. Cui *et al.,* quant à eux, ont étudié la différence entre une infection par l'influenza et par la dengue par le biais du métabolome du sérum de sujets humains [94]. Des sujets en pleine santé ont aussi été inclus dans l'étude. Ils ont notamment observé une abondance différentielle pour des métabolites liés au métabolisme des purines et de certains acides aminés ainsi qu'à la synthèse des acides gras. Ces résultats montrent que le virus de l'influenza a un effet local au site d'infection, mais aussi au niveau systémique. Il est donc probable que la métabolomique pourrait être employée pour aider au diagnostic de l'influenza lorsque couplée à des données de nature clinique comme l'âge, le sexe et la présence de différents symptômes.

La métabolomique par spectrométrie de masse à haut débit permet l'analyse d'un échantillon en quelques secondes. Ce processus d'analyse élimine la traditionnelle séparation chromatographique. L'infusion directe et le DART sont deux technologies ayant déjà été employées afin de comparer la signature spectrale entre deux populations d'échantillons [56,95]. Ici, nous démontrons le potentiel de la technologie LDTD-MS en

métabolomique non ciblée à haut débit appliquée au diagnostic de l'influenza. En combinant la vitesse d'acquisition de cette approche et la puissance des algorithmes d'apprentissage automatique, nous montrons qu'il est possible, sous certaines conditions, de trouver des biomarqueurs liés à une infection virale des voies respiratoires. En plus de montrer le potentiel de l'approche analytique par LDTD-MS, nous démontrons l'utilisation de *MetaboDashboard*, un outil permettant de facilement employer l'apprentissage automatique pour la recherche de biomarqueurs en métabolomique non ciblée.

4.2 Méthode

4.2.1 Récolte des échantillons et extraction métabolique

Lors des saisons grippales 2016-2017 ainsi que 2018-2019, un volume d'au moins 150 µl par échantillons cliniques a été récupéré au laboratoire d'analyse virologique du Centre hospitalier de l'Université Laval où l'échantillon était gardé à -80°C. Cet aliquot était récupéré à même la partie inutilisée de l'échantillon et seulement une fois le test clinique effectué. À l'exception de 16 échantillons provenant de sujets en bonne santé, les échantillons proviennent tous de patients souffrant de symptômes d'une infection respiratoire, d'où la demande du test PCR. Chaque échantillon aliquoté était conservé à -80°C jusqu'à l'extraction. Quatre jeux de données ont été produits :

- 2017-02-10 : 93 échantillons contrôles et 94 échantillons confirmés positifs au virus Influenza A analysés en duplicata.
- 2017-03-02 : 95 échantillons contrôles, 95 échantillons confirmés positifs au virus Influenza A et 95 échantillons confirmés positifs au virus RSV analysés en duplicata.
- 2017-08-02 : 87 échantillons contrôles, 88 échantillons confirmés positifs au virus Influenza A et 16 échantillons de sujets en pleine santé, analysés en duplicata.
- 2019-02-14 : 327 échantillons contrôles, 172 échantillons confirmés positifs au virus influenza A, 2 échantillons confirmés positifs au virus Influenza B et 51 échantillons confirmés positifs au virus RSV.

L'extraction des métabolites a été effectuée par extraction liquide-liquide assistée par le sel (SALLE). 100 µl de l'échantillon clinique (UTM) ont été mélangés à 200 µl d'acétonitrile froid gardé sur glace et 200 µl d'eau saturée en sel (NaCl) ont été ajoutés au mélange afin de retirer les protéines et les autres molécules difficiles à désorber. Après 15 secondes de

vortex, l'échantillon reposait 5 minutes avant d'être centrifugé à 7200 rpm pendant 1 minute afin d'aider à la séparation des phases. 2 µl de la phase supérieure ont ensuite été déposés sur une Lazwell 96 puits (Phytronix) pour une analyse par LDTD-MS. Pour les acquisitions en ionisation négative, les puits des Lazwell ont été prétraités avec une solution de EDTA (100 µg/ml dans un mélange de méthanol/NH_4OH/H_2O 75:15 :10) afin de chélater la surface métallique. Pour chaque expérience, l'ensemble des échantillons ont été extraits et analysés par LDTD-MS à l'intérieur d'une même journée afin de réduire les variations expérimentales. Les échantillons ont été extraits et analysés par spectrométrie de masse dans un ordre aléatoire afin de limiter les biais.

4.2.2 Acquisition des données par LDTD-MS

L'acquisition des données a été effectuée avec une source LDTD (Phytronix inc.) couplée à un spectromètre de masse Synapt G2-Si (Waters Corp.). La méthode d'acquisition était une méthode indépendante des données (MS^e), permettant l'acquisition d'information sur l'ion moléculaire et les fragments. Dans la fonction de haute énergie, l'énergie de collision dans la cellule de transfert était une rampe de 10 V à 40 V. La charge de l'aiguille corona était de 4,0 µA en ionisation positive et 3,0 µA en ionisation négative. La source LDTD employait le patron laser suivant : 2 secondes à 0%, rampe jusqu'à 65 % en 6 secondes, plateau à 65 % durant 2 secondes, retour à 0% en 0,1 seconde.

4.2.3 Analyse des données

Le signal correspondant à chaque échantillon a été sommé et converti en centroïde avec le logiciel ProcessKernel (Waters Corp.) en utilisant seulement la première fonction de basse énergie. Les spectres de masse résultants de cette transformation ont été utilisés pour la suite des analyses.

Les spectres ont d'abord été corrigés avec les correcteurs de masse virtuels, puis alignés avec le même outil tel que décrit précédemment [78]. Brièvement, les données ont été séparées en ensembles d'entraînement et de test à 30 reprises. L'ensemble d'entraînement a été utilisé pour sélectionner les valeurs de taille de fenêtre et les pics de VLM. La transformation apprise a ensuite été appliquée sur l'ensemble d'entraînement et de test. Les données ainsi alignées ont été importées dans MetaboDashboard. Les ensembles d'entraînement et de test préalablement construits pour la correction ont été employés pour les expériences d'apprentissage automatique. Le problème de classification était de

différencier les sujets souffrant d'une infection virale de ceux ayant obtenu un résultat négatif par PCR, donc non infectés par un virus. En effectuant 30 séparations aléatoires par expérience, il a été possible de mesurer l'effet de la constitution de l'ensemble d'entraînement et de test sur les résultats. Les algorithmes d'arbre de décision (*Decision Tree*, DT), de forêt aléatoire (*Random Forest*, RF), de machine à vecteurs de support (*Support Vector Machine*, SVM) et de machine à couverture d'ensemble (*Set Covering Machine*, SCM) ont été utilisés. À l'exception du SCM dont l'implémentation est disponible à https://github.com/aldro61/pyscm, les implémentations fournies dans la librairie Scikit-Learn ont été employées [96]. Une validation croisée en 5 parties a été effectuée pour la sélection des hyperparamètres de chaque algorithme (Tableau 4.1).

Tableau 4.1 Hyperparamètres utilisés pour la validation croisée en 5 parties pour chaque algorithme.

Algorithme	Paramètre	Valeurs possibles
DecisionTreeClassifier	max_depth	[1, 2, 3, 4, 5, 10]
	min_samples_split	[2, 4, 6, 8, 10]
RandomForestClassifier	n_estimators	[1, 2, 4, 10, 30, 70, 100, 500]
LinearSVC	C	numpy.logspace(-5, 5, 30)
SetCoveringMachineClassifier	p	[0.001, 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]

4.3 Résultats et discussion

4.3.1 Développement de la méthode d'extraction et de désorption

La précipitation des protéines par l'ajout de solvant organique est une méthode d'extraction simple et rapide qui est hautement compatible avec la technologie LDTD. Elle a d'ailleurs été employée avec succès dans le passé pour la détection de composés pharmaceutiques dans le plasma et pour la détection d'infection par le parasite de la malaria [78]. Des essais préliminaires employant une précipitation des protéines dans l'acétonitrile avec le milieu de transport universel (UTM) vierge employé lors de la collecte des échantillons montraient la présence résiduelle de certaines molécules causant une marque brun-noir liée à leur carbonisation sur la plaque Lazwell. Cette observation indique qu'un ou des composés présents dans l'extrait sont incompatibles avec le processus de désorption et brûlent sur la surface métallique. La carbonisation de molécules sur la surface métallique peut diminuer la reproductibilité du processus de désorption pour d'autres molécules. Afin de résoudre ce problème, nous avons augmenté la sélectivité de la méthode de préparation d'échantillons. La séparation liquide-liquide par ajout d'une solution d'eau saturée en sel est une méthode

de préparation simple et assez rapide qui permet, encore ici, d'extraire les métabolites dans un solvant organique. Elle est donc compatible avec une approche de métabolomique à haut débit par LDTD-MS. En utilisant cette méthode, les marques carbonisées sur les Lazwell à la suite de la désorption ont disparu. La nature du composé carbonisé est toujours inconnue, mais il est possible de croire qu'il s'agit de sucrose. Ce saccharide est présent dans l'UTM et avec la chaleur, le sucrose peut caraméliser et coller à la surface métallique. De plus, il est possible de remplacer le sel par un sucre lors d'une séparation liquide-liquide assistée par le sel. On appelle cette variante séparation liquide-liquide assistée par le sucre. Nous croyons que lors de l'ajout de la solution aqueuse, le sucre se retrouve, tout comme le sel, dans la phase inférieure aqueuse et quitte la phase organique qui est déposée sur la Lazwell. Cette anecdote rappelle l'importance de la préparation d'échantillons dans le processus de développement d'une méthode d'analyse métabolomique à haut débit.

Une fois la méthode d'extraction des métabolites sélectionnée, la puissance du patron laser a été optimisée. La désorption reproductible d'une grande variété de composés peut être difficile. Puisqu'aucune séparation n'est employée, les métabolites cristallisés sur la Lazwell puis désorbés vont tous atteindre le spectromètre de masse en quelques secondes. Cette arrivée massive de molécules peut causer une compétition pour l'ionisation et/ou une saturation du détecteur. Ces phénomènes sont indésirables puisqu'ils affectent la qualité du signal produit lors de l'analyse par spectrométrie de masse. Dans le cas d'une analyse non ciblée avec la source LDTD, l'étude du chromatogramme de désorption permet de détecter la suppression ionique, qui se traduit souvent par un creux dans le pic de désorption (Figure 4.1). Malgré l'attention particulière donnée aux différents facteurs influençant la désorption comme l'extraction, la dilution et le patron laser, certains échantillons ont montré des signes de suppression ionique. Ces échantillons représentent moins de 15 % des échantillons d'une expérience. Malheureusement, l'information quantitative provenant de ceux-ci s'en trouve affectée. Par le fait même, les performances des modèles de classification qui se basent sur cette information pourraient être réduites. L'augmentation du facteur de dilution ou l'utilisation de plusieurs méthodes d'extractions différentes mais plus spécifiques pourraient aider à résoudre ce problème.

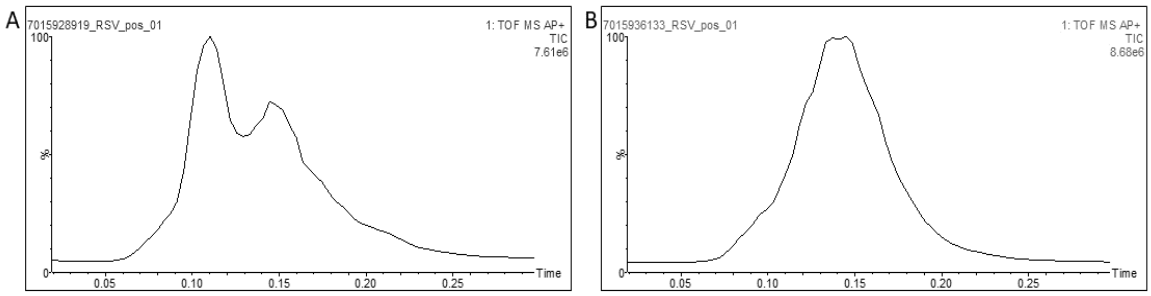


*Figure 4.1 Comparaison de chromatogrammes acquis par LDTD-MS en présence (**A**) et en absence (**B**) de suppression ionique.*

4.3.2 Recherche de biomarqueurs

Le MetaboDashboard a comme principale fonction la construction et la comparaison de modèles d'apprentissage automatique dans le but de trouver des biomarqueurs reliés aux phénotypes d'intérêts. Quatres algorithmes produisant des modèles relativement parcimonieux et interprétables ont été employés et comparés pour la tâche consistant à classifier les échantillons provenant de sujets positifs au test PCR pour une infection virale des voies respiratoires et ceux montrant des symptômes similaires, mais négatifs au test PCR (Tableau 4.2 et 4.3). Dans la grande majorité des cas, les algorithmes souffrent de surapprentissage puisqu'ils performent mieux sur l'ensemble d'entraînement que sur l'ensemble de tests, mais l'arbre de décision et la machine à couverture d'ensemble surapprennent moins que les autres algorithmes. Ce phénomène peut s'expliquer par la haute dimensionnalité des données ainsi que par la nature de ces algorithmes parcimonieux. On remarque aussi que sur les expériences 2017-02-10, 2017-03-02 et 2017-08-02, la plupart des algorithmes ont une exactitude moyenne supérieure à 75 % sur l'ensemble de test allant même au-dessus de 80 % dans certains cas. Toutefois, l'expérience 2019-02-14 donne de très mauvais résultats, l'exactitude moyenne étant inférieure à 60 % sur l'ensemble de test pour tous les algorithmes utilisés. Ces résultats suggèrent un problème de reproductibilité.

Tableau 4.2 Exactitudes moyennes des classifications sur 30 séparations de Monte-Carlo pour quatre algorithmes. La valeur entre parenthèses correspond à l'écart-type.

	DT		RF		SCM		SVM	
EXPÉRIENCE	Entrain.	Test	Entrain.	Test	Entrain.	Test	Entrain.	Test
2017-02-10	0,91 (0,05)	0,76 (0,06)	1,00 (0,00)	0,75 (0,06)	0,95 (0,01)	0,77 (0,07)	1,00 (0,00)	0,76 (0,06)
2017-03-02	0,86 (0,03)	0,80 (0,05)	1,00 (0,00)	0,81 (0,04)	0,91 (0,01)	0,73 (0,05)	1,00 (0,00)	0,80 (0,05)
2017-08-02	0,91 (0,04)	0,77 (0,04)	1,00 (0,00)	0,79 (0,06)	0,89 (0,02)	0,74 (0,05)	1,00 (0,00)	0,76 (0,06)
2019-02-14	0,75 (0,14)	0,56 (0,4)	0,99 (0,02)	0,57 (0,04)	0,64 (0,02)	0,58 (0,04)	0,99 (0,002)	0,56 (0,03)

Tableau 4.3 Taux de faux positifs (infection virale) et faux négatifs sur l'ensemble de test pour quatre algorithmes. La valeur entre parenthèses correspond à l'écart-type.

	DT		RF		SCM		SVM	
EXPÉRIENCE	**F.P.**	**F.N.**	**F.P.**	**F.N.**	**F.P.**	**F.N.**	**F.P.**	**F.N.**
2017-02-10	0.16 (0.06)	0.07 (0.04)	0.20 (0.06)	0.05 (0.03)	0.15 (0.09)	0.09 (0.06)	0.15 (0.05)	0.09 (0.04)
2017-03-02	0.15 (0.04)	0.05 (0.04)	0.17 (0.04)	0.02 (0.01)	0.18 (0.04)	0.01 (0.01)	0.12 (0.04)	0.07 (0.03)
2017-08-02	0.15 (0.05)	0.08 (0.05)	0.12 (0.05)	0.10 (0.06)	0.13 (0.08)	0.11 (0.08)	0.13 (0.05)	0.11 (0.05)
2019-02-14	0.15 (0.11)	0.30 (0.10)	0.08 (0.06)	0.35 (0.06)	0.02 (0.02)	0.39 (0.04)	0.21 (0.04)	0.23 (0.04)

La fonction d'efficacité du récepteur (courbe ROC) et l'aire sous la courbe de la fonction d'efficacité du récepteur (AUC) sont des métriques standards pour décrire et évaluer les performances d'un test diagnostic médical [97]. Ces métriques se basent sur la relation entre le ratio de vrai positif (infection virale confirmée par PCR) et taux de faux positifs produits par un modèle. La Figure 4.2 montre la courbe d'efficacité du récepteur pour l'algorithme de la forêt aléatoire à travers les quatre expériences. Pour toutes les expériences de 2017, l'algorithme fait beaucoup mieux qu'un modèle aléatoire alors que ce n'est pas le cas sur le jeu de données de 2019-02-14. Ces résultats corrèlent avec les observations obtenues par le score d'exactitude. L'AUC suit aussi la même tendance (Annexe B – Tableau 9.1). Avec un AUC moyen autour de 0,75, notre test de détection d'infection virale des voies respiratoires combinant LDTD-MS et apprentissage automatique est moins performant que les tests PCR qui offrent typiquement un AUC supérieur à 0,9, mais offre tout de même des performances acceptables.

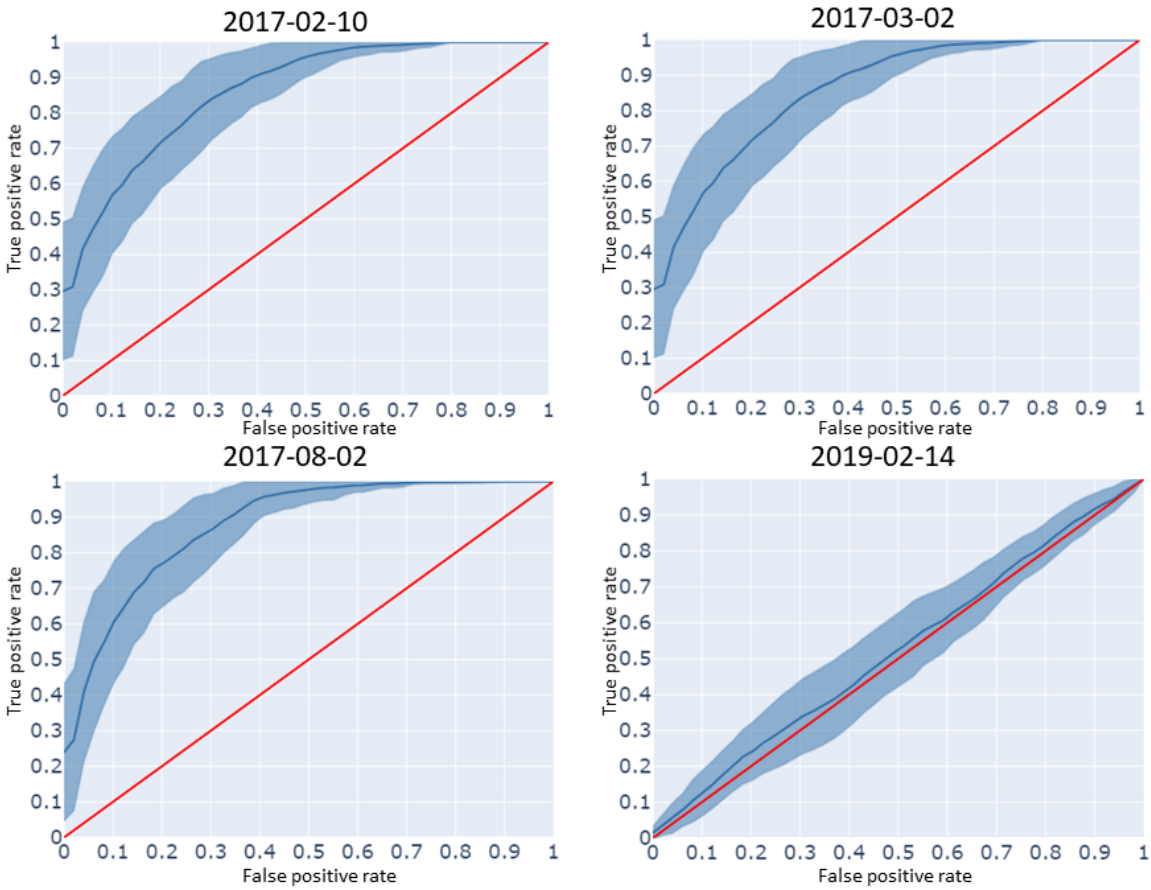


Figure 4.2 Fonctions d'efficacité du récepteur (Courbe ROC) moyennes pour l'algorithme de la forêt aléatoire sur les différents jeux de données. La ligne rouge correspond à un prédicteur aléatoire. La surface ombragée correspond à l'écart-type sur le taux de vrai positif.

Un de nos objectifs était la recherche de biomarqueurs. En étudiant les métabolites employés dans le processus de décision des différents modèles, on peut en trouver qui sont utilisés par plusieurs modèles pour un algorithme et même certaines qui sont utilisés par plusieurs algorithmes différents. Ces métabolites auraient donc un haut potentiel prédictif. Si l'on extrait l'ensemble des ions employés à plus de deux occasions sur les 30 modèles d'un algorithme pour une expérience donnée, on peut remarquer une certaine distribution dans l'utilisation des ions (Figure 4.3). Pour l'expérience 2017-03-02, les arbres de décision et les SCMs ont effectué l'ensemble de leurs classifications à partir de 12 et de 32 ions respectivement. Quant à elles, les forêts aléatoires ont employé le plus d'ions. On notera qu'il est possible qu'en restreignant les hyperparamètres affectant la construction des arbres constituant les forêts, le nombre d'ions employés par cet algorithme diminue, mais ce changement pourrait avoir un impact sur les performances.

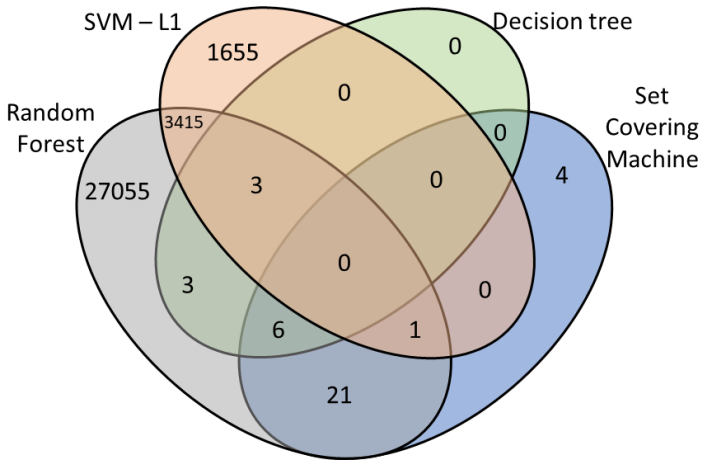


Figure 4.3 Diagramme de Venn des ions employés dans le processus de décision des différents algorithmes de classification pour l'expérience du 2017-03-02. Les ions employés par un seul modèle d'un algorithme ont été retirés de la figure.

En étudiant individuellement le patron d'abondance des ions employés à plusieurs reprises, on remarque que certains montrent effectivement une différence significative alors que d'autres montrent une abondance globale assez similaire entre les deux classes (Figure 4.4). On doit toutefois être prudent dans l'interprétation des abondances puisque la très grande majorité des modèles utilisent de multiples ions pour effectuer la prédiction et non pas un seul. Un ion pris individuellement peut sembler insignifiant, mais lorsque combiné à d'autres ions, son impact dans le processus décisionnel peut devenir très important.

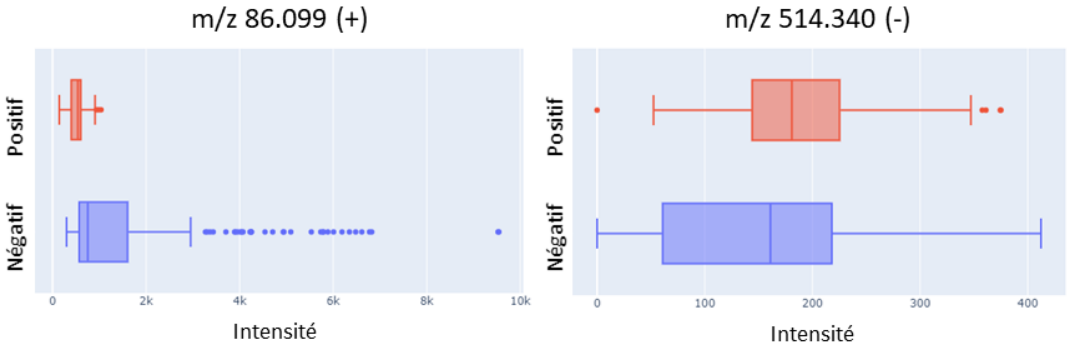


Figure 4.4 Diagramme en boîte comparant l'intensité de deux ions employés par plusieurs modèles entre les classes positives et négatives pour une infection virale.

Bien que nous ayons identifié des ions ayant le potentiel d'être des biomarqueurs, l'identification des métabolites correspondant à ces ions est très difficile avec les données à notre disposition. D'abord, en absence d'un vrai correcteur de masse, l'exactitude des

mesures de masse ne peut être garantie. Nous serions obligés d'étendre le champ de recherche à toutes les molécules à au moins ± 50 ppm des masses mesurées. Ensuite, l'absence d'un spectre de fragmentation, de la section efficace (CCS) et d'information sur le type d'ion limite grandement l'information à notre disposition pour effectuer la recherche dans les bases de données. Dans l'éventualité où nous aurions désiré développer un test diagnostique en métabolomique non ciblée à haut débit par LDTD, il aurait pu être nécessaire d'identifier les molécules correspondant aux ions. La validation du potentiel prédictif d'un modèle par une analyse ciblée et quantitative des métabolites pertinents aurait pu suivre. Cette étape pourrait potentiellement augmenter les performances du modèle en réduisant les variations expérimentales.

4.3.3 Reproductibilité

Pour finir, nous nous sommes penchés sur la diminution des performances dans les algorithmes sur le jeu de données de 2019. Nous émettons l'hypothèse qu'elle est possiblement liée à un problème reproductibilité lors de la collecte et/ou de la conservation des échantillons, entre autres. Nous avons effectivement remarqué une corrélation importante entre les pics présents dans un spectre et la couleur d'un échantillon (Figure 4.5). En présence d'un changement de pH, le rouge de phénol présent dans la solution de transport universelle change de couleur pour passer du rouge rosé au jaune orangé. Des essais ont été tentés afin d'augmenter la reproductibilité en changeant le pH d'un échantillon par ajout de tampon. Malheureusement, le changement de pH seul n'explique pas le changement d'allure du spectre de masse. On peut supposer qu'un autre facteur influence le pH et, par le fait même, les métabolites représentés dans le spectre. Nous n'avons pu confirmer si la suppression ionique observée dans un petit nombre d'échantillons des études de 2017 était aussi corrélée avec la couleur du milieu de transport des échantillons.

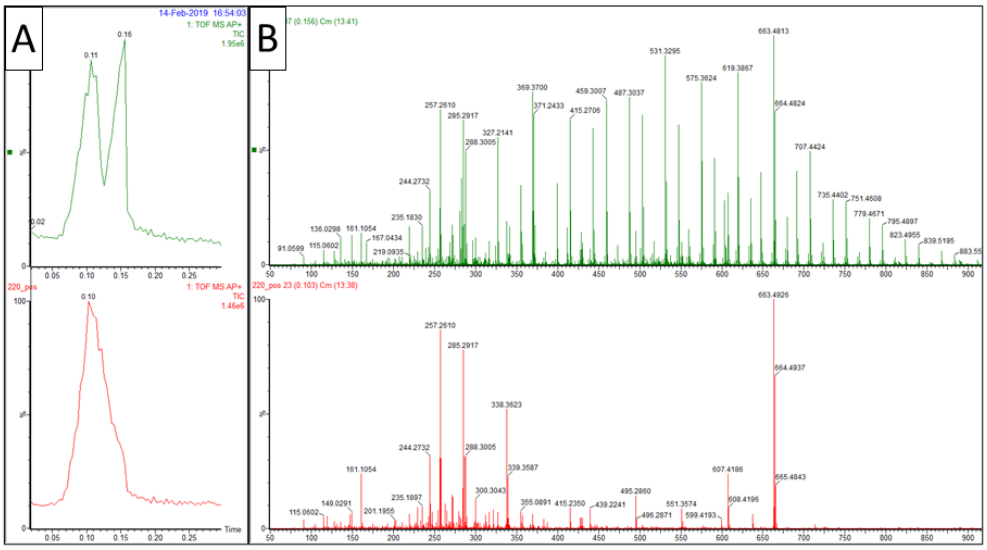


*Figure 4.5 Comparaison du chromatogramme de désorption (**A**) et du spectre de masse (m/z : 50-900) (**B**) d'un échantillon à pH basique (**haut**) et d'un échantillon à pH régulier (**bas**). Un facteur inconnu influence le pH et le contenu métabolique de certains échantillons du jeu de données 2019-02-14.*

4.4 Conclusion

Globalement, si l'on considère le peu d'optimisation effectuée sur les méthodes de préparation d'échantillons, ces résultats sont très encourageants et démontre le potentiel de l'approche proposée. On peut donc affirmer qu'il est possible d'effectuer des expériences de métabolomique à haut débit en utilisant la source LDTD et de chercher des biomarqueurs dans les données produites en employant des approches d'apprentissage automatique. L'utilisation du MetaboDashboard pour cette tâche a permis de construire rapidement différents modèles de classification à partir de nombreux algorithmes et d'identifier les ions les plus importants dans le processus de décision avec un minimum d'effort. Puisque cet outil sauvegarde l'ensemble des données et résultats, il est facile d'aller plus loin dans l'analyse des résultats que ce qui est offert directement dans l'outil.

L'utilisation d'une méthode basée sur la technologie LDTD-MS peut amener des problèmes pour l'identification des métabolites correspondant aux ions d'intérêt dans un modèle de classification. Toutefois, des expériences supplémentaires qui incluraient l'ajout d'un correcteur de masse réel ainsi que l'acquisition de spectres de fragmentation pourraient amoindrir ce désagrément. L'emploi de la mobilité ionique pourrait aussi permettre la mesure de la section efficace lors d'une première acquisition, sans nécessiter l'ajout d'un

standard internet ou d'un correcteur de masse. De plus, bien que la métabolomique à haut débit soit très intéressante pour effectuer l'analyse de grande cohorte d'échantillons, il est important de considérer les difficultés liées à l'obtention d'une quantité suffisante d'échantillons pour les classes à l'étude lors de la conception d'une étude. Par exemple, dans le cas ici présent, le nombre de patients présentant un résultat positif à une infection virale des voies respiratoires en dehors de la saison de la grippe est très faible. Il devient donc difficile de récolter un nombre d'échantillons suffisant pour permettre l'emploi d'une approche par LDTD-MS et apprentissage automatique.

Les résultats de cette étude démontrent tout de même le grand potentiel des études de métabolomique non ciblée à haut débit. L'emploi de différentes méthodes d'extraction ciblant des familles de composés d'intérêt tels que les acides aminés ou les lipides pourrait permettre d'observer des métabolites pertinents qui étaient masqués par les plus abondants dans notre étude. La collecte d'un plus grand nombre et d'un plus grand volume d'échantillons dans des conditions plus contrôlées aurait aussi pu aider à produire des modèles de classification plus performants. Finalement, l'utilisation de l'UTM comme échantillon de départ en métabolomique non ciblée pourrait aussi être reconsidérée. Quoique très utile pour un test PCR visant à détecter l'ARN viral, l'utilisation d'une brosse dans la cavité nasopharyngée ne garantit pas un échantillonnage équivalent d'un patient à l'autre. On fait donc face à des variations non contrôlées qui peuvent affecter les résultats d'études métabolomiques. L'emploi d'une matrice biologique tel que l'urine ou le sang, dont les constitutions sont régulées par des processus biologiques, pourrait être plus approprié à condition de chercher des biomarqueurs présents dans l'organisme entier et non seulement près du site d'infection.

Chapitre 5. Présentation du deuxième article

5.1 Référence

Pier-Luc Plante, Élina Francovic-Fontaine, Jody C. May, John A. McLean, Erin S. Baker, François Laviolette, Mario Marchand, and Jacques Corbeil. **Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS**; Analytical Chemistry (2019); 91 (8), 5191-5199; DOI: 10.1021/acs.analchem.8b05821

5.2 Contexte

Les précédents chapitres étaient consacrés à la préparation des données et à la recherche de biomarqueurs grâce à l'utilisation d'algorithmes d'apprentissage automatique. Une étape critique pour favoriser l'interprétation des modèles de classification et pour évaluer la pertinence des biomarqueurs utilisés est l'identification des biomarqueurs. Cette étape consiste à identifier la molécule correspondant au signal de chaque marqueur.

Cette identification peut être effectuée à l'aide de différents outils utilisant la masse moléculaire et le spectre de fragmentation. Cette recherche dans les bases de données permet d'obtenir une liste d'identifications hypothétiques. Afin de confirmer l'identification, il est requis d'acquérir un standard pur et de comparer différentes caractéristiques entre le biomarqueur et le standard. Afin de réduire le nombre d'identifications hypothétiques, il est possible d'utiliser la CCS (Collision Cross-Sections). Malheureusement, cette caractéristique est encore peu présente dans les bases de données. Nous proposons ici un outil de prédiction de la CCS basé sur un réseau de neurones à convolution afin d'aider à l'identification des biomarqueurs.

5.3 Contribution

Je suis responsable de la conception initiale du projet, de la coordination de la collaboration, de l'exécution du projet et de la construction de l'outil. Je suis aussi responsable de la rédaction du manuscrit. Élina Francovic-Fontaine a contribué à la réalisation des expériences et à la conception de l'interface de l'outil. Jody C. May, John A. McLean et Erin S. Baker ont contribué par leur expertise ainsi que par la récolte de données et à la révision du manuscrit. Ils ont également effectué des mesures analytiques supplémentaires pour

valider certaines hypothèses. François Laviolette, Mario Marchand et Jacques Corbeil ont participé à la conception du projet et à la révision du manuscrit.

5.4 Discussion

L'article présenté au chapitre 6 propose un nouvel outil pour la prédiction de la CCS mesuré par mobilité ionique en phase gazeuse. La structure d'un réseau de neurones à convolution dirigée par les données et permettant la construction d'une représentation interne à partir du SMILES (*simplified molecular-input line-entry system*) d'une molécule a été présentée. Les performances du modèle ont été évaluées sur une fraction de sept jeux de données, dont deux externes gardés entièrement pour l'évaluation des performances. Une comparaison à un autre outil, considéré comme étant l'état de l'art, a aussi été effectuée. La réutilisation de la structure du réseau pour d'autres tâches en chimio-informatique a aussi été évaluée.

Contrairement à l'étape de recherche, la simplicité et l'interprétabilité de modèles aidant à l'identification des biomarqueurs ne sont pas une nécessité. Pour cette raison, l'utilisation de SVM non linéaires et de réseaux de neurones peut permettre de construire des modèles plus complexes et plus performants. Une des raisons possibles de ce gain de performance est que le réseau peut lui-même déterminer l'information nécessaire pour la tâche de prédiction. DeepCCS utilise effectivement la représentation de la molécule (SMILES) comme entrée pour la prédiction au lieu d'un ensemble de descripteurs moléculaires. Cela lui permet de construire une représentation d'un SMILE optimisée pour la tâche de prédire la CCS. Au contraire, un modèle utilisant une série de descripteurs moléculaires est limité par cette information. L'interprétabilité des réseaux de neurones est un domaine de recherche en pleine effervescence. L'étude des facteurs influençant le processus décisionnel d'un réseau de neurones pourrait permettre de mieux comprendre la relation entre certains facteurs et la valeur prédite en plus de trouver un juste milieu entre interprétabilité, complexité et performance. Dernièrement, Ross et al. ont étudié en profondeur le lien entre différents descripteurs moléculaires et la valeur de CCS dans le but d'optimiser un algorithme de prédiction [68].

Une particularité intéressante des travaux effectués pour cet article inclut la correction de bases de données construites à partir de données expérimentales. Effectivement, les prédictions de l'algorithme divergeaient grandement des mesures expérimentales pour cinq

molécules utilisées pour nos expérimentations computationnelles. De nouvelles mesures à partir de standards certifiés ont permis de confirmer qu'il s'agissait d'erreurs expérimentales lors de la première mesure et non d'une erreur de prédiction du modèle. Une fois l'entraînement complété, la prédiction par un modèle d'apprentissage profond peut être extrêmement rapide. Comme montré dans nos travaux, ces algorithmes peuvent être résistants aux données erronées, à condition qu'elles soient peu nombreuses. Pour ces raisons, la validation de mesures expérimentales est un domaine où il pourrait être intéressant d'appliquer l'apprentissage profond. Dans ces cas, même un modèle offrant des performances non optimales pourrait suffire à détecter des erreurs et à augmenter la confiance par rapport aux valeurs expérimentales obtenues. Encore ici, le manque d'interprétabilité et la complexité du modèle ne nuiraient pas à son utilisation : peu importe si la prédiction est éloignée, on validera le résultat.

Depuis la parution de l'article, d'autres groupes se sont attaqués au problème de prédiction de CCS. Notamment, Ross et al. ont construit trois modèles différents qui peuvent être chacun utilisés pour prédire la CCS [68]. La sélection du modèle à utiliser s'effectue à partir d'un regroupement non supervisé qui est corrélé avec la classe chimique des molécules. La classe de la molécule est définie à partir de la même information utilisée pour effectuer la prédiction : un ensemble de descripteurs moléculaires. Les performances de CCSbase, ce nouveau modèle, seraient supérieures à DeepCCS. Comme le modèle de CCSbase a été entraîné avec un plus grand nombre d'exemples que DeepCCS, ces résultats semblent bien indiquer qu'avec l'augmentation de la quantité et de la qualité des données expérimentales les prédicteurs de CCS deviendront plus performants et polyvalents.

Chapitre 6. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS

6.1 Résumé

Les analyses de métabolomique non ciblée par spectrométrie de masse sont bien adaptées pour la recherche de biomarqueurs. Cependant, l'étape d'identification des petites molécules demeure encore un grand défi. L'utilisation de valeurs de section efficace en phase gazeuse (*collision cross section,* CCS) obtenue par mobilité ionique permet de réduire le nombre de faux positifs lors du processus d'identification des métabolites. Quoique prometteur, le nombre de valeurs de CCS empirique présentement disponible est limité. Il faut donc se tourner vers des méthodes de prédiction de valeur de section efficace. Nous proposons ici de résoudre ce problème en effectuant la prédiction à partir d'un algorithme à base de réseau de neurones profond. Nous avons développé et entrainé un modèle qui effectue cette prédiction à partir de la notation SMILES d'une molécule et du type d'ion. Considérant les performances, le temps et les ressources nécessaires ainsi que l'applicabilité à une grande variété de molécules, le modèle proposé surpasse l'ensemble des outils de prédiction de CCS précédemment disponibles.

6.2 Abstract

Untargeted metabolomic measurements using mass spectrometry are a powerful tool for uncovering new small molecules with environmental and biological importance. The small molecule identification step however still remains an enormous challenge due to fragmentation difficulties or unspecific fragment ion information. Current methods to address this challenge are often dependant on databases or require the use of nuclear magnetic resonance (NMR), which have their own difficulties. The use of the gas-phase collision cross section (CCS) values obtained from ion mobility spectrometry (IMS) measurements was recently demonstrated to reduce the number of false positive metabolite identification. While promising, the amount of empirical CCS information currently available is limited, thus predictive CCS methods need to be developed. In this manuscript, we expand on current experimental IMS capabilities by predicting the CCS values using a deep learning algorithm. We successfully developed and trained a prediction model for CCS values requiring only information about a compound's SMILES notation and ion type. The use of data from five

different laboratories using different instruments allowed the algorithm to be trained and tested on more than 2400 molecules. The resulting CCS predictions were found to achieve a coefficient of determination of 0.97 and median relative error of 2.7% for a wide range of molecules. Furthermore, the method requires only a small amount of processing power to predict CCS values. Considering the performance, time, and resources necessary, as well as its applicability to a variety of molecules, this model was able to outperform all currently available CCS prediction algorithms.

6.3 Introduction

Mass spectrometry (MS) is widely used for biomarker discovery and to explore prevailing metabolomic processes. Untargeted MS measurements coupled with high performance liquid chromatography (LC) allow the detection of thousands of ions in a matter of minutes. However, even though the resolution, mass accuracy, and sensitivity of mass spectrometers continue to improve, the identification of small molecules is still challenging due to their limited mass range and number of possible isomers [98–100]. To date, most methods for metabolite identification are based on mass spectra database comparison. In these comparisons, spectra obtained experimentally are matched to the database containing a list of known molecular masses and fragmentation patterns. However, the vast majority of features in an MS experiment cannot be identified due to limited entries in current databases and/or insufficient fragmentation coverage. Furthermore, even if the chemical formula of a particular species is convincingly identified, structural identification remains a challenge due to the number of isomer species which can exist for any given chemical formula [101].

To deal with these challenges, the Metabolomics Standards Initiative has published guidelines for metabolite identification [49,72], which give identification confidence levels depending on the amount of information discerned for the molecule. Based on these recommendations, the best way to confidently identify a metabolite is to use two independent and orthogonal data types for each authentic compound analyzed, such as GC or LC retention time, molecular mass, and tandem mass spectra. However, this step can be costly as numerous authenticated chemical standards are required to attain the highest confidence level (i.e., a level 1 identification). Additionally, chemical standards are oftentimes unavailable for true unknown molecules (novel compounds), thus requiring

several orthogonal analytical measurements and/or custom synthesis to support an identification.

Recently, the use of ion mobility spectrometry coupled with mass spectrometry (IM-MS) has become very promising for adding a structural dimension to MS analysis based on collision cross sections (CCSs) in support of metabolomic studies [102]. In contrast to other properties such as retention time, CCS is an ion parameter which can be measured with relative standard deviation (RSD) ranging from 0.29% to 6.2% when using different instrumental platforms [103] and 3% or lower on average when using a standardized method [104–106], making it a valuable property for metabolite identification that is reproducible between different laboratories. The use of CCS can reduce the number of possible identifications and the number of false positive identifications in untargeted metabolomic studies [66,107,108]. Additionally, when reference CCS values are not available in a database, it is possible to compute theoretical CCS values on structures obtained from molecular simulations. Currently this approach requires choosing the proper theory and approach for converting candidate structures to CCS, many of which are computationally-expensive without being as precise as an experimental measurement [66,109]. A more efficient process to produce CCS values for small molecules is through machine learning approaches [110]. By using a training set, the algorithm attempts to identify the relation between an input (usually a set of molecular descriptors) and the CCS values. If the learning step is successful, the function can be applied to new inputs to obtain accurate and efficient predictions. The performance of the model is verified using validation and testing sets that were not used during the training step.

Deep neural networks (DNN), a type of machine learning algorithm, are now commonly used in multiple domains such as self-driving cars, medical diagnosis, drug optimisation and speech recognition [111–113]. Compared with other machine learning algorithms such as support vector machines and random forests that learn models directly from a set of user-provided features, deep learning algorithms are composed of a cascade of layers which extract increasingly complex features (i.e., combinations of the original features) from the initial input (Figure 6.1). The DNN models are trained to build a representation which is then used to perform a prediction task. Convolutional neural networks (CNN), a subtype of DNN, are widely used in image recognition due to their capability to resist translation and transformation of features present in the input [114]. CNN structures can be separated in two

components (Figure 6.1 and Figure 6.2). The first is the feature-learning component, which is constituted of multiple successions of convolution filter layers and maximum pooling layers (Figure 6.2). The output of the feature-learning component is a hidden internal representation of the input constructed by the neural-network. The second component, known as the predictive section, performs classification or regression depending on the task at hand, through a series of fully connected layers using the internal representation as input. DNN and CNN have already been used successfully in chemoinformatics for predicting molecular properties and protein-ligand interactions [115].

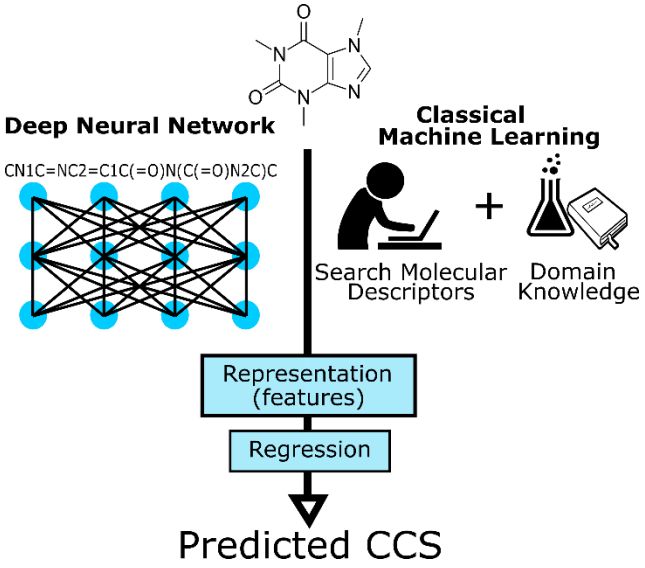


Figure 6.1 Comparison between Deep Neural Network and classical machine learning for CCS prediction. Blue sections are purely computational, making DNN an almost completely computational approach. Classical machine learning requires an input of well-defined and comprehensive molecular descriptors which can be adversely influenced by domain knowledge (e.g. CCS is correlated with the m/z value), reducing its accuracy.

CCS prediction using machine learning has been addressed on multiple occasions in prior studies [66,69,70]. Although results from these previous works were published, the prediction models and the code needed to reconstruct the models are unavailable. Moreover, previously published prediction models might not generalize well to new data from multiple laboratories because these models were mostly trained on datasets produced in a single laboratory and on a single instrument, making them highly specific to a certain context. Furthermore, most if not all prediction models to date use a set of molecular descriptors as the input for predictions. This transfers the problem of predicting CCS to the issue of finding or computing the values for a set of molecular descriptors (e.g., polar surface area, molar refractivity, etc.) which is not straightforward and thus is prone to user error. Simplified molecular-input line-entry system (SMILES) are structurally-descriptive notations which can

be readily assigned to any compound with a known structure, and are already used as input by different methods to compute molecular descriptors [116]. In this work, we utilized chemical SMILES, a chain of characters easily found in chemical compound databases, as the input of a CNN model to predict CCS for different types of molecules. We generated a neural network structure based on CNN for CCS prediction and measured the performances of the generated models on different testing sets. We also evaluated the reusability of the SMILES internal representation learned by the model on a multi-task learning problem. Finally, we offer a simple command line tool to use the generated model for CCS predictions.

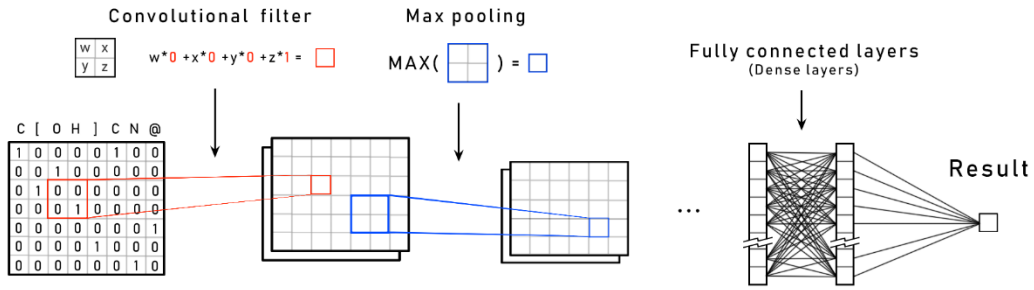


Figure 6.2 Schematic representation of the different operations performed by a convolutional neural network.

6.4 Experimental section

6.4.1 Datasets

Five datasets containing CCS and mass information were collected from multiple sources including measurements from drift tube ion mobility (DTIM) and traveling wave ion mobility (TWIM) instruments, in order to constitute a learning database that includes a large panel of molecules. These sources included:

- MetCCS [117]**:** 779 CCS values were measured on an Agilent 6560 DTIM-QTOF-MS instrument (Agilent Technologies, Santa Clara, CA). This dataset is already separated between a training (n=648) and testing (n=131) set.
- Astarita [106]**:** 205 CCS values were measured on a Waters Synapt G2 Q-TWIM-TOF-MS instrument (Waters Corporation, Manchester, UK). The positive and negative ion mode datasets were used as an independent testing sets for comparison with MetCCS web server predictions.

- Baker [118]**:** 857 CCS values were measured using the Agilent 6560 DTIM-QTOF-MS instrument customized for increased precision and reproducibility. This dataset contains multiple types of small molecules.
- McLean [119]: 211 CCS values were measured using the Agilent 6560 DTIM-QTOF-MS instrument. This diverse dataset contains CCS values for amino acids, lipids, metabolites and peptides.
- CBM2018 [70]: 357 CCS values were measured using a Waters Vion TWIM-QTOF-MS instrument. This dataset contains pharmaceuticals, drugs of abuse, and their metabolites.

6.4.2 Data preparation

For each molecular entry, the SMILES notation was retrieved from PubChem except for amino acids sequences for which the SMILES were generated using the Python RdKit module. The datasets were filtered to keep only SMILES with less than 250 characters/chemical symbols. This removed only a few entries and allowed the network input to be kept at a reasonable size. The only ions considered in this evaluation were (M+H)+, (M+Na)+, (M-H)- and (M-2H)2- in order to have at least 50 examples per ion type. The different datasets were split into a training, validation, and testing set following the schema in Figure 6.3. The Astarita datasets were all included in the DeepCCS testing set to allow a valid comparison to MetCCS predictors, while 20% of the Baker, McLean and CBM2018 datasets were included in the testing set to better evaluate the generalisation of the models.

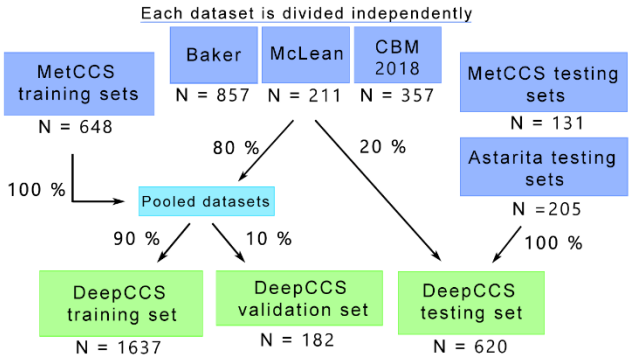


Figure 6.3 Partitioning of the different source datasets between the training, validation and testing set of DeepCCS.

To feed the neural network, SMILES and ions were encoded using one-hot vector encoding and padded to a length of 250 characters. This resulted in binary matrices of 250 x 36 for SMILES and binary vectors of length 4 for the ions.

6.4.3 Neural network structure optimisation and training

The decision to use a CNN was justified by the intuitive way this type of DNN learns. The CNN looks for features in the input, and in the case of image recognition, these features can be lines, points, or combination of these, leading to more complex objects (e.g, eyes or wheels). In our case, these features can be chemical groups and substructures. Since these groups of atoms can be anywhere in the SMILES representation, the translation resistance characteristic of CNN was a good fit. Furthermore, it was shown that using CNN with SMILES as input can give results equivalent or better than the state-of-the-art in many chemo-informatics applications[120].

A CNN structure is modulated by a set of hyperparameters that affect the number and width of layers, convolution filter size, and maximum pooling window size, among others. Different hyperparameters have a different impact on the capability of a model to learn and generalize. In this work, hyperparameter optimisation was performed by five-fold cross validation using a random search approach. Implementation and experiments were performed in Python. The CNN was built using the Keras library with the Tensorflow backend. After training, the model giving the best score on the validation set at the end of the different epochs was retained (Figure 6.4). The training of a model using the standard dataset partitioning (Figure 6.3) takes around 25 minutes on a Nvidia Tesla P100 GPU. Prediction of 100 CCS values using the DeepCCS command line tool takes approximately 3 seconds on a standard desktop computer without the use of a GPU. Additional information about model optimisation and construction are available in supplementary information. All the code needed to train the network and to reproduce the results on the different testing sets presented in this paper are available at github.com/plpla/DeepCCS/.

6.4.4 Evaluation of the internal representation reusability

In order to evaluate the reusability potential of the internal representation learned by a CNN using SMILES as input, a multi-task experiment was performed. The SMILES and molecular properties of every compound in the Human Metabolome Database (HMDB) were extracted. Only compounds with a valid SMILES and valid values for polar surface area, logS,

refractivity, polarizability, logP (ALOGPS) and logP (Chemaxon), were retained. This allowed the extraction of 71,232 compounds. This dataset was randomly separated between the training, validation and testing set using 72%, 8% and 20% of the complete dataset. The SMILES encoder previously learned was adjusted to include new chemical symbols not seen in the CCS datasets. A new CNN based on the DeepCCS structure (Figure 6.4) was built with the following changes: the second input (ion type) was removed, the concatenation layer was removed and six different dense sections, one per property, replaced the single dense section. The resulting multi-task CNN structure can be consulted in Annex C - Table 10.4. The task of the network was to predict the different properties using a common internal representation. This allowed the network to learn a general internal presentation of a SMILES. After the first training phase, a DeepCCS model was reconstructed using the convolution and maximum pooling layers that were trained on the multi-task problem. The weights of the feature learning part layers were locked to prevent further learning. The new half-trained network was retrained for 150 epochs using the CCS data after encoding using the updated SMILES encoder and the exact same dataset split.

6.5 Results and discussion

6.5.1 DeepCCS Network Structure

Convolutional neural networks are known to learn an internal representation of the input through a series of convolution and maximum pooling steps. This internal representation is then used as the input for a multi layer perceptron to perform predictions. The network structure of DeepCCS that was obtained after optimization uses the same principle: transformation of the SMILES provided in input to an internal representation and prediction using this representation and the ion type for which the CCS prediction must be made. The use of a second input allows DeepCCS to separate the internal representation of the input and the ion type to let the network focus only on the molecular structure in the feature-learning part. The final structure of the DeepCCS neural network is presented in Figure 6.4 and details can be found in Annex C - Table 10.3 and in the source code. It contains a series of 7 convolution and maximum pooling layers to study the molecular structure of the SMILES provided in input. It is worth mentioning that the downscaling factor (strides parameter) was increased to a value of 2 on the last maximum pooling layer to significantly reduce the number of trainable weights in the network without impacting the prediction accuracy.

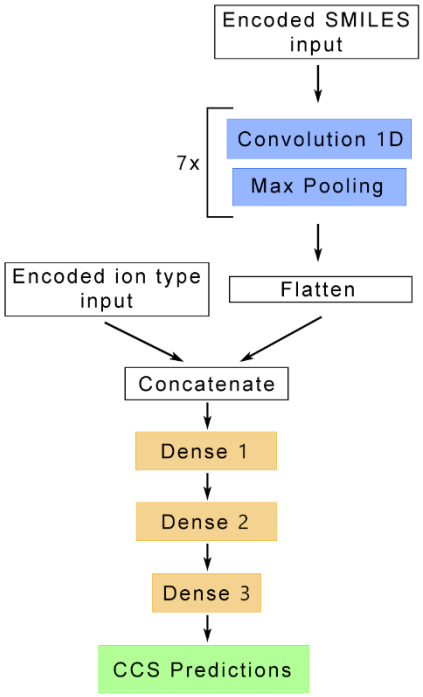


Figure 6.4 DeepCCS neural network structure. The upper part, in blue, performs a series of convolution and maximum pooling steps to learn an internal representation of the encoded SMILES input. This representation is flattened and concatenated with the ion type to be processed by the lower part of the network, in orange, to perform CCS prediction through a series of dense layers. The detailed network structure is available in Annex B

6.5.2 CCS prediction

To evaluate the robustness of the training step, two different experiments were performed (Table 6.1). First, ten different models were trained on a single dataset partition. Since the initialisation of the trainable weights are different for each model, the resulting model explored different path to finally converge to a final state that produced similar predictions. The second experiment consisted of training ten different models using ten different data splits generated randomly. This allowed to study the impact of the training and testing set composition on the performances of the model. For example, if a set of molecules exhibit a molecular substructure that the network could not interpret properly because of a lack of example was exclusively in the training set from the first experiment, it would have erroneously resulted in good results. As shown in Table 6.1, the results are very similar between the two experiments showing that dataset splitting, and the network initial weights values have a minimal impact on the performances of the various trained models.

Table 6.1 Average coefficient of determination (R2) and median relative error over ten different models trained using either a single dataset split or different dataset splits. Only data from the different testing set partitions were used. Standard deviation values are in parenthesis.

Dataset	Single split		Different splits	
	R^2	Median relative error (%)	R^2	Median relative error (%)
Global	**0.976 (0.001)**	**2.67 (0.18)**	**0.979 (0.004)**	**2.37 (0.27)**
MetCCS test pos.	0.960 (0.005)	2.02 (0.24)	0.964 (0.007)	1.93 (0.35)
MetCCS test neg.	0.969 (0.005)	3.11 (0.49)	0.967 (0.007)	3.15 (0.63)
Astarita pos.	0.901 (0.013)	4.86 (0.30)	0.897 (0.011)	4.77 (0.44)
Astarita neg.	0.955 (0.006)	3.13 (0.48)	0.949 (0.008)	3.36 (0.37)
Baker	0.954 (0.006)	2.43 (0.11)	0.967 (0.010)	2.02 (0.17)
McLean	0.995 (0.001)	1.49 (0.14)	0.996 (0.001)	1.15 (0.28)
CBM 2018	0.930 (0.010)	2.26 (0.28)	0.969 (0.010)	1.26 (0.47)

When all testing set were merged into a single, global testing set, the coefficient of determination (R^2) was greater than 0.97 and the absolute median relative error (MRE) was below 2.6%, indicating an excellent accuracy of prediction when compared to experimentally measured values. Considering that this global testing set was not used during the training step and that it contains data originating from five different laboratory and multiple instruments, one can conclude that the model achieved a state of generalization where it can be applied to new molecules. Furthermore, since the reported deviation for ion mobility CCS measurement can be as high as 6.2% [121], these results appear acceptable. Similar results were obtained when removing from the test sets the SMILES-ion combinations that are present in the training set with different CCS values, thus making sure no similar examples were already seen by the model before generating predictions (Annex C - Table 10.). Using the compounds classification provided in the Baker dataset, Figure 6.5 shows that the model performs well for different types of compounds such as amino acids, fatty acids and lipids, hence the model correctly discriminates the differences between multiple types of chemical structures and can predict the CCS value properly. With the exception of a few outliers, the concordance between measured and predicted CCS values is close to the reference line which indicates overall good predictions (Figure 6.5).

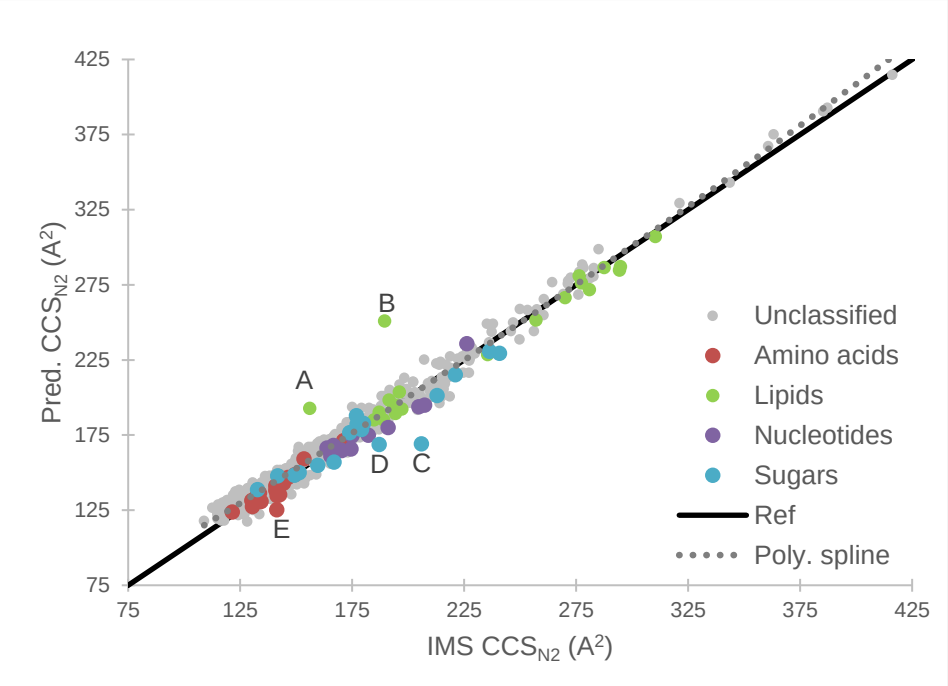


Figure 6.5 Comparison of IMS measured and predicted CCS values for all compound from the testing set. Compound classes are from the Baker dataset, others are unclassified. The solid line (Ref) represents a reference line of perfect fit (a slope of 1). The dotted line indicates a 2nd order polynomial spline fit to the data to show overall tendency. Letters A to E indicate outliers, respectively methyl behenate, 1,2-diacyl-sn-glycero 3-phosphocholine, D-maltose, sophorose and L-threonine (see Annex B).

Although global performances of the algorithm were satisfactory, the Astarita datasets showed poorer correlation compared to the other datasets, with an average R^2 lower than 0.9 and a MRE close to 5% for positive ions. Since the performances on the other testing datasets were better, we hypothesized that either the measurement accuracy of the Astarita datasets were lower than the other datasets or that a bias in measurement between datasets is present. To investigate this further, CCS values for identical SMILES and ion type were compared, which allows the variation between datasets to be evaluated (Table 6.2). The average difference between the Astarita positive dataset and Baker (n=57) or McLean (n=14) datasets was approximately 5% for overlapping measurements, which is significantly higher than the differences observed between other datasets. This seems to indicate that a CCS measurement bias is present in these datasets, which serves to decrease the performance of the model during the testing step. In fact, previous studies have demonstrated that large (>5%) differences in CCS can exist when comparing measurements obtained from drift tube (e.g., Baker and McLean datasets) and traveling wave instruments (e.g., Astarita datasets) [121]. Since both Astarita datasets were not included in the training step, this bias does not affect the model directly. The comparison of CCS measurement in

multiple laboratories using different experimental conditions also allows us to get a better idea of the real variation that can be expected when comparing experimental CCS values from different studies.

Table 6.2 Comparison of CCS measurement for identical molecules and ion type between the different datasets. The value is the non-absolute mean percent difference relative to the average CCS measured by two datasets.

	MetCCS train pos.	MetCCS train neg.	MetCCS test pos.	MetCCS test neg.	Astarita pos	Astarita neg	Baker	McLean	CBM 2018
MetCCS train pos.	0.00	--	0.37	--	**4.36**	--	-1.77	-0.03	1.14
MetCCS train neg.	--	0.00	--	--	--	--	-3.74	-1.92	--
MetCCS test pos.	-0.37	--	0.00	--	**2.91**	--	-2.40	-1.20	--
MetCCS test neg.	--	--	--	0.00	--	-0.23	-4.02	-3.21	--
Astarita pos.	-4.36	--	-2.91	--	0.00	--	-5.54	-4.88	--
Astarita neg.	--	--	--	0.23	--	0.00	-4.99	-2.82	--
Baker	1.77	3.74	2.40	4.02	**5.54**	**4.99**	0.00	0.26	3.62
McLean	0.03	1.92	1.20	3.21	**4.88**	**2.82**	-0.26	0.00	0.38
CBM 2018	-1.14	--	--	--	--	--	-3.62	-0.38	0.00

Based on our results, CCS measurements can be reproducible well below the reference 2% values (Baker vs McLean, n=54) but it can also be as high as 5% (Astarita vs Baker, n=57). These results corroborate what was observed by Schmitz et al. when comparing CCS values obtained from different IM instrumentation and techniques [121]. The difference between TWIMS (i.e. Astarita, CBM2018) and DTIMS (i.e. MetCCS, Baker, McLean) are not systematic for most molecule types but some show appreciable differences [121]. Once again, an appropriate calibration for TWIMS is critical to obtain high quality measurement using this technique. These results also put emphasis on the importance of using data from different contexts, such as instruments (DTIMS and TWIMS) and laboratories, to obtain a predictor that can generalize to every context when insufficient training data from a specific context are available. As more datasets are published and included in the training step of DeepCCS, the real variation of CCS measurements will become more precise and the DeepCCS model will improve such that it should be able to predict values closer to the average CCS, therefore increasing its performance. The addition of the IMS measurement technique could also be

included in the model as more data become available, making the model adaptable to the different contexts.

6.5.3 Outliers detection for database validation

Outliers in Figure 6.5 (points A to E) were further investigated. These data points respectively correspond to 1,2-diacyl-sn-glycero 3-phosphocholine, methyl behenate, D-maltose, sophorose and L-threonine. All outliers except sophorose were confirmed as measurement error by remeasuring the CCS value of the compounds (Annex C - Figure 10.). For sophorose, we hypothesise that the error is similar to the one for D-maltose. Carbohydrates are prone to aggregation, and these multimers readily dissociate between the IM and the MS stage, resulting in ion signals at higher CCS values (Annex C - Figure 10.-C). Since the measured value is, like for D-maltose, higher than the predicted value, we hypothesis that it might also be a case of aggregation-dissociation of sugar molecules leading to an erroneous measurement.

Outliers investigation allowed to detect four confirmed and one unconfirmed but highly probable erroneous measurements. These results highlight on another potential utility of CCS prediction tools: database validation. By comparing predicted and measured CCS values, one can easily detect suspect measurements and further investigate their validity. The ease of use and good performances of DeepCCS makes it ideal for this task.

6.5.4 Comparison to existing tools

The DeepCCS model uses SMILES notation as input which is easy to obtain for most small molecules. When performing metabolite identification using MS data, a popular approach is to compare the empirically-measured spectra to reference spectra from a database. This reference database necessarily contains the structure of the compounds and therefore, the SMILES notation is either already present or easily computed. In contrast, other CCS predictors use molecular descriptors as input, which are not always available in databases and can require licensing commercial software to compute them.

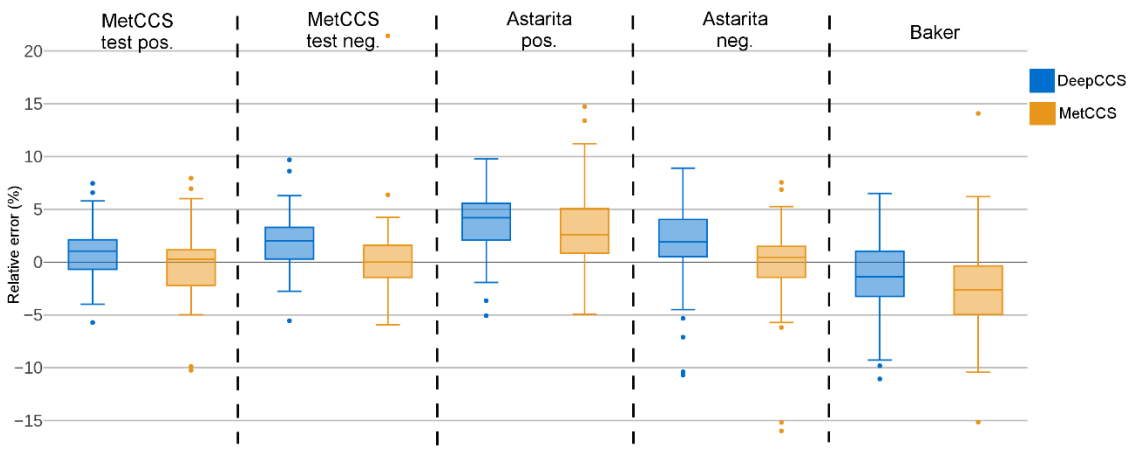


Figure 6.6 Comparison of the error distribution on five different testing sets between DeepCCS and MetCCS. Previously detected database errors were removed from the comparison.

The MetCCS web server is a CCS prediction tool based on Support Vector Regression and uses 14 common molecular descriptors available in the HMDB database. MetCCS has been used to generate over 176,000 CCS values for over 35,000 small molecule metabolites from the HMDB [66]. Although MetCCS does not allow CCS prediction for molecules other than metabolites, a separate tool, LipidCCS, has been developed by the authors for lipids and fatty acids [65]. In contrast, CCS prediction for all molecule types can be done directly in DeepCCS. Results from the DeepCCS model were compared with those obtained using the MetCCS server in order to evaluate the performance of each machine learning approach (Figure 6.6). The MetCCS testing datasets and the Baker dataset were used, as MetCCS does not work for all molecule types and requires HMDB identifiers to collect the associated molecular descriptors. The MetCCS prediction server produce the most extreme values on most datasets. This could be explained by the sub representation of certain molecules types in MetCCS training set that is much smaller. Overall, both predictors perform similarly with most predictions within a 5% window, but we can discern the impact of the different training set of the predictions. Both models were trained almost exclusively with data from DTIMS (DeepCCS used a fraction of the CBM2018), but MetCCS was trained with data from a single laboratory. The Baker and McLean CCS values are far more distant than MetCCS training set value to Astarita values.

When considering the MRE and R^2 (Table 6.3), the performances were found to be similar when using the MetCCS testing set, but MetCCS performed better when the Astarita datasets were used. This might be explained by the close proximity of MetCCS and Astarita

CCS values (Table 6.2). Although, when evaluating the performances using the Baker dataset, DeepCCS performed better and could predict the CCS for more molecules (171 instead of 134). This shows that the model improves on current methods through better generalisation and by providing accurate predictions on multiple compound types.

Table 6.3 Comparison of DeepCCS and MetCCS predictive performances using different CCS testing sets.

	R^2		Median relative error (%)	
	DeepCCS	MetCCS	DeepCCS	MetCCS
MetCCS Test Positive	0.97	0.95	1.63	1.74
MetCCS Test Negative	0.98	0.97	2.30	1.54
Astarita Positive	0.93	0.93	4.22	2.96
Astarita Negative	0.97	0.97	2.21	1.47
Baker Testing Set (n=171 & 134)	0.95	0.9	2.50	3.00

Even with these advances, DeepCCS has its own limitations. It can only perform predictions using features that have already been observed. For example, only chemical symbols of atoms contained within the dataset are available for encoding, therefore only SMILES with these symbols can be used for predictions. This limitation is in place to ensure predictions are based on features that are recognized by the CNN: that is, the model cannot perform predictions using items it has never seen. Similar limitations apply to molecule types, thus while DeepCCS can perform CCS prediction on any compound, the predictions might be less accurate for molecule types and substructures not seen during the initial training stage. Figure 6.7 shows molecule distribution at the superclass level from the ClassyFire taxonomy [122]. Even though the training set contains multiple examples, it clearly does not cover all possible molecule. In all cases, these limitations can be solved by generating sufficiently large and diverse CCS datasets to train a new model.

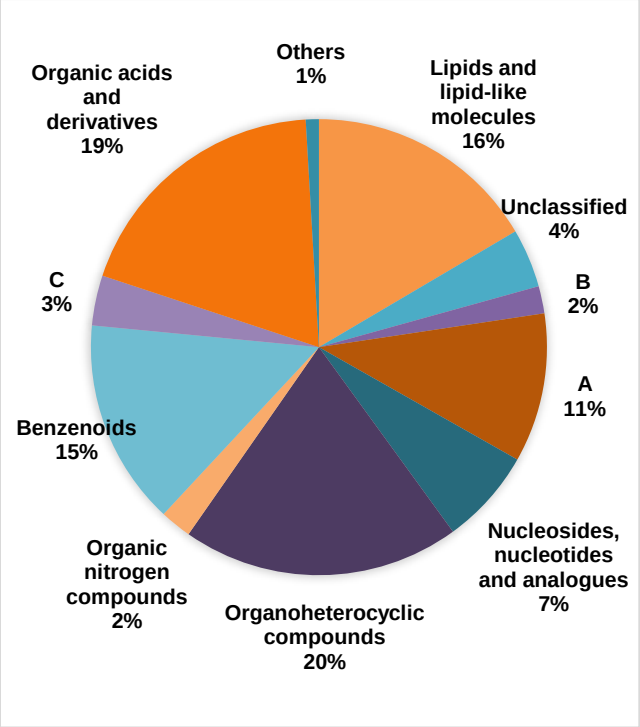


Figure 6.7. Classification at the superclass level of the molecules from DeepCCS datasets using the ClassyFire taxonomy. The Others group contains the following classes: Organohalogen compounds, Organic Polymers, Organosulfur compounds and Homogeneous non-metal compounds. A. Organic oxygen compounds. B. Alkaloids and derivatives. C. Phenylpropanoids and polyketides. Classification tables at the subclass and class levels are available in Annex C.

6.5.5 Generalisation of the internal representation

The SMILES input contains all of the structural information of the molecule, and as such, the CNN internal representation should be generalizable to predicting other molecular properties. This assumption was evaluated by performing molecular properties prediction on a multi-task problem. Six different chemical properties available in the Human Metabolome database (polar surface area, logS, refractivity, polarizability, logP (Alogps), and logP (Chemaxon)) were predicted with the objective to learn an internal representation. The resulting model predicted these chemical properties with very good accuracy ($R^2 > 0.98$) and a median relative error below 0.7% (Annex C).

This generalized internal representation incorporating the dataset from the single split experiment was subsequently used for CCS prediction. Performances similar to what was obtained in previous experiments were also obtained (Table 6.4). When predicting CCS, this new model performed with a global R^2 of 0.968 and a median relative error of 2.6%. The results of this experiment showed that the internal representation learned by a CNN using

SMILES can be reused to predict different molecular properties beyond CCS and that significantly increasing the number of SMILES used to learn the internal representation does not have an impact of the accuracy of CCS predictions. Therefore, we hypothesise that to further increase CCS prediction accuracy, the predictive part of the network would need additional data for a better understanding of the link between the network internal representation and the CCS value. The other possibility, and probably the best way to increase prediction accuracy, would be to decrease the variations between CCS measurement in the different datasets used for training the model. A large CCS database (n>3,800) exhibiting high measurement precision has recently been developed by the authors and will include further CCS measurement as they will be published. It will be used in future work to further improve the predictive capabilities of DeepCCS [108].

Table 6.4 Model performances on CCS prediction after training the feature learning section of the network on a multi-output problem. The dataset split is identical to the single split previously used.

Dataset	R^2	Median relative error (%)
Global	**0.968**	**2.55**
MetCCS test pos.	0.928	2.43
MetCCS test neg.	0.941	2.37
Astarita pos.	0.878	4.27
Astarita neg.	0.945	2.89
Baker	0.950	2.15
McLean	0.986	1.50
CBM 2018	0.912	2.76

6.6 Conclusion

CCS prediction using machine learning is necessary to populate the numerous possible small molecule CCS values with high speed and accuracy. The DeepCCS prediction algorithm uses SMILES notation as an input instead of a more traditional set of molecular descriptors, which allows DeepCCS to be fast (100 predictions in ~3 seconds) and, due to the CNN structure used, is also generalizable to a large number of different molecule types. Additionally, the DeepCCS command line tool provides an easy way to train a new model using newly generated data or to simply predict CCS values using the provided model. The precision of empirical CCS measurements used as a training set was found to have a significant impact on the overall prediction accuracy of the model. In this case, the wide

variations observed (2% to more than 5%) in measured CCS is certainly a limiting factor on the capability of the model to predict CCS with less than 3% error. The performance of machine learning models such as DeepCCS will improve as more high-quality CCS measurements are made available.

6.7 Acknowledgments

The authors acknowledge Alexandre Drouin and Prudencio Tossou for valuable comments and discussions. This work was supported in part by a Fond de recherche québécois en santé (FRQS) Doctoral award to PLP and the Canada Research Chair in Medical Genomics (JC). JC, PLP and EFF also acknowledge funding from Mitacs (MC00006). J.C.M and J.A.M acknowledge funding from the National Institutes of Health (NIGMS R01GM092218 and NCI 1R03CA222452-01). E.B. would also like to acknowledge funding support from the National Institute of Environmental Health Sciences of the NIH (R01 ES022190 and P42 ES027704). We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPU used for this research. Part of the computations were performed on the Compute Canada supercomputer infrastructure.

6.8 Supplementary information

Method supplementary information is available in Annex C. It contains: Outliers description; IM spectra and CCS values for repeated measurement; Neural network structures and hyper-parameters values used during cross-validation; HMDB chemical properties prediction results; Effect of repeated SMILES-ion combination on model performances; ClassyFire classification of the compounds.

Discussion et conclusion

Dans cette thèse, nous avons développé des outils pour l'analyse de données en métabolomique non ciblée à haut débit dans le but d'effectuer la recherche et l'identification de biomarqueurs grâce à l'utilisation de méthodes d'apprentissage automatique. Les contributions scientifiques produites dans le cadre de ce doctorat touchent trois étapes critiques de l'analyse de données en métabolomique non ciblée (Figure 7.1).

D'abord, le concept de correcteur de masse virtuel a été développé et différents outils pour leur recherche ainsi que la correction de spectres ont été présentés. Des algorithmes ont été créés afin de permettre la correction de jeux de données contenant des milliers de spectres acquis par un instrument utilisant une chambre de temps de vol. Entre autres, les algorithmes permettent de minimiser la distance entre les pics de masses d'un ensemble de spectres afin de rendre ces derniers hautement comparables. L'avantage principal de cette approche est qu'elle permet de conserver la haute précision des mesures de masses ce qui n'est pas le cas du *binning*. Il a été montré que l'utilisation du correcteur de masse virtuel et de l'aligneur de spectres dérivé des VLM peut augmenter la précision des algorithmes d'apprentissage automatique sur ces données en plus de réduire le nombre de pics utilisés dans le processus de classification de différents modèles.

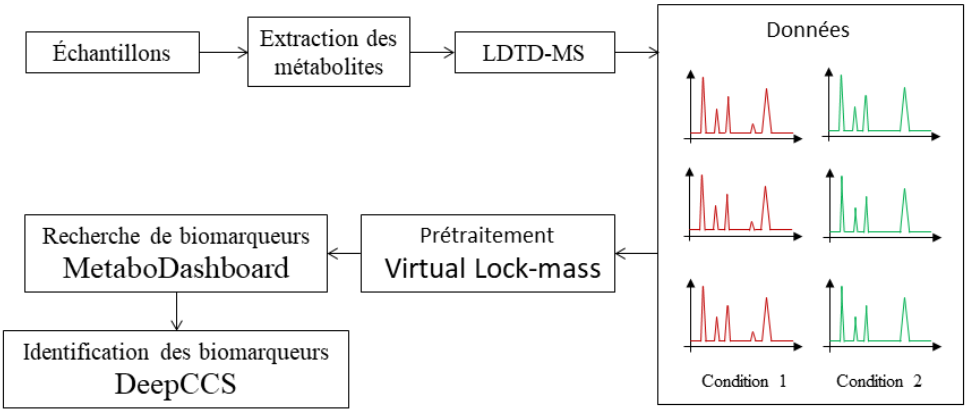


Figure 7.1. Représentation des étapes d'une expérience de métabolomique non ciblée par LDTD-MS.

Nous avons ensuite développé un outil pour simplifier la recherche de biomarqueurs par apprentissage automatique. Le *MetaboDashboard* est un outil permettant de construire des

modèles de classification sans connaissances préalables de l'apprentissage automatique et de la programmation. Une attention particulière a été mise sur la présentation des résultats, la simplification du processus d'entraînement et l'emploi de bonnes pratiques pour l'utilisation de l'apprentissage automatique dans le contexte de la métabolomique. Le potentiel du *MetaboDashboard* a été mis de l'avant dans une expérience visant la recherche de biomarqueurs par LDTD-MS dans le contexte de l'influenza. Dans cette étude, nous avons cherché la présence de biomarqueurs permettant de distinguer des patients en proie à une infection virale des voies respiratoires et ceux présentant des symptômes similaires, mais sans infection virale. Cette étude a permis de montrer le potentiel de l'approche de métabolomique non ciblée par LDTD-MS et cette dernière ouvre la porte à l'étude de larges populations d'échantillons à une vitesse inégalée.

Finalement, nous avons développé un outil pour la prédiction de la valeur de section efficace (CCS) à partir du SMILES d'une molécule. L'utilisation du CCS en métabolomique non ciblée lors du processus d'identification de biomarqueurs permet de réduire le nombre de mauvaises identifications. Le modèle développé est basé sur un réseau de neurones à convolution, effectue des prédictions avec une erreur médiane relative de 2,7% et permet d'effectuer des prédictions pour une grande variété de molécules. Nos résultats ont aussi montré le potentiel de la structure du réseau de neurones développée pour la prédiction de différentes propriétés physicochimiques de molécules à partir du SMILES.

Ensemble, ces outils forment la première étape du développement d'une approche de métabolomique non ciblée à haut débit à partir de données de spectrométrie de masse et d'apprentissage automatique. Afin d'en arriver à un processus performant et reproductible, plusieurs étapes du pipeline doivent maintenant être optimisées. En utilisant les outils développés dans la cadre de cette thèse, il est possible de considérer l'optimisation des paramètres expérimentaux au niveau de la préparation des échantillons et de l'instrumentation, de la préparation des données et de la recherche de biomarqueurs par apprentissage automatique.

Extraction des métabolites et analyse instrumentale

La spectrométrie de masse à haut débit en est encore à ses balbutiements. La source LDTD, une des technologies les plus rapides pour l'introduction d'échantillons dans un spectromètre de masse, est couramment utilisée pour la métabolomique ciblée, mais les

analyses non ciblées l'employant sont très peu répandues, voire inexistantes, en dehors des travaux présentés dans cette thèse. Maintenant qu'une série d'outils aidant à la recherche et à l'identification de biomarqueurs dans le contexte de la métabolomique non ciblée à haut débit sont disponibles, une attention particulière devra être portée sur les étapes affectant la production des données. Les approches employées jusqu'à maintenant sont relativement simples et devront être optimisées pour produire des données de qualité. Nos résultats mitigés à propos de nombreux problèmes de classification comme le lupus, le cancer du sein, les accidents cardiovasculaires, la dégradation du plasma dans le temps et la différenciation homme/femme démontrent que les méthodes employées ne permettent pas nécessairement de trouver des biomarqueurs alors qu'un phénotype est bien présent. Puisque des biomarqueurs ont déjà été identifiés pour la majorité de ces phénotypes en employant des méthodes plus classiques comme le LC-MS, on peut croire que nos difficultés à trouver des biomarqueurs par LDTD-MS proviennent des types de métabolites ionisés et de la qualité des données produites.

Nos résultats préliminaires visant à évaluer la reproductibilité en LDTD-MS montrent que pour un large éventail de métabolites de fromage cheddar, la moyenne des déviations standards relatives à travers 12 réplicats techniques se situe autour de 22 % [123]. En employant une source DESI, cette métrique descend à une valeur beaucoup plus acceptable de 9 %. Récemment, Phytronix Technologies a montré que pour plusieurs familles de composés, l'ajout sur la Lazwell d'une couche protectrice à partir de composés possédant des fonctions chimiques similaires, permet d'améliorer la désorption et d'obtenir un signal linéaire en fonction de la concentration. Pour ces raisons, je crois qu'afin de produire des résultats de qualité offrant les meilleures chances de trouver des biomarqueurs par LDTD-MS, un éventail de méthodes de préparation d'échantillons devrait être développé. Chaque méthode pourrait cibler des groupes de métabolites qui possèdent les mêmes groupements chimiques dans leur structure moléculaire. Ce faisant, une couche protectrice appropriée pourrait être appliquée sur la Lazwell et un patron laser offrant des performances maximales pourrait être utilisé. De plus, contrairement à l'approche employée dans cette thèse où nous essayions d'extraire le maximum de composés dans une seule extraction, la méthode proposée pourrait permettre l'utilisation de standards internes puisque les spectres produits seraient beaucoup plus simples. La séparation de l'extraction métabolique en différentes fractions pourrait aussi simplifier le processus d'identification des métabolites correspondant aux biomarqueurs en réduisant le champ de recherche. Le

désavantage d'une telle approche est qu'elle nécessiterait d'effectuer plusieurs extractions et plusieurs désorptions sur le même échantillon, augmentant le temps et le coût d'analyse pour un seul échantillon. Malgré cela, l'acquisition en moins de 10 secondes garantit un gain de temps considérable lors de l'analyse instrumentale comparativement au LC-MS. De plus, l'automatisation du processus d'extraction pourrait augmenter la cadence d'analyse des échantillons en plus de réduire la variabilité relative aux manipulations.

D'autres sources d'ionisation à haut débit comme le DESI, le REIMS (iKnife) et le DART ont montré un certain potentiel pour les analyses de métabolomique non ciblée. Comme mentionné précédemment, différentes sources d'ionisation sont plus ou moins performantes pour certains types de molécules. L'utilisation d'une multitude de sources d'ionisation pourrait augmenter les chances de trouver des biomarqueurs liés à un phénotype. On pourrait potentiellement obtenir une couverture des métabolites d'un échantillon supérieure à celle obtenue par LC-MS en une fraction du temps d'analyse. Des résultats préliminaires ont permis de montrer que seuls 3 à 10 % des correcteurs de masse virtuels étaient partagés entre les données produites par DESI, LDTD et REIMS. Ces résultats préliminaires suggèrent donc une grande variabilité au niveau des différents ions obtenus par ces trois sources d'ionisation.

Pour en arriver à optimiser les méthodes d'extraction des métabolites et l'acquisition des données, un processus d'évaluation et de comparaison des résultats devra être élaboré. Je crois que deux métriques pourraient être employées à cette fin. D'abord, la distribution des écarts-types sur l'intensité des points de masse employés comme VLM permettrait d'avoir un aperçu de la variation expérimentale. On émettrait l'hypothèse qu'en réduisant au maximum les écarts sur l'intensité des pics partagés par l'ensemble des échantillons, on diminuerait les variations dans les spectres en entier, y compris les potentiels biomarqueurs. La deuxième métrique de comparaison pourrait être le nombre de métabolites trouvés dans un mélange dont la composition est connue. Par exemple, un extrait métabolique consistant en des métabolites marqués au ^{13}C serait approprié. On chercherait à maximiser le nombre de métabolites trouvés dans cet extrait à travers les différentes extractions. Ceci permettrait d'obtenir la plus grande couverture métabolique possible.

Préparation des données

Une fois que les méthodes d'acquisition de données seront optimisées, il pourrait être pertinent de travailler sur l'optimisation de la préparation des données. La recherche de pics, la normalisation des intensités et la déconvolution sont des transformations qui pourraient avoir un impact significatif sur les résultats de la recherche de biomarqueurs.

Pour la recherche de pics, une sélection des ions dont le chromatogramme extrait montre une corrélation avec le patron laser pourrait aider à diminuer significativement le débalancement entre le nombre d'exemples et le nombre de variables pour la recherche de biomarqueurs (Figure 7.2). Effectivement, plusieurs pics de masse pourraient être éliminés en sélectionnant seulement ceux provenant de métabolites subissant une désorption. Les ions provenant du gaz porteur et le bruit produit par l'instrument seraient ainsi éliminés. J'estime que cette transformation pourrait réduire de moitié le nombre d'ions total.

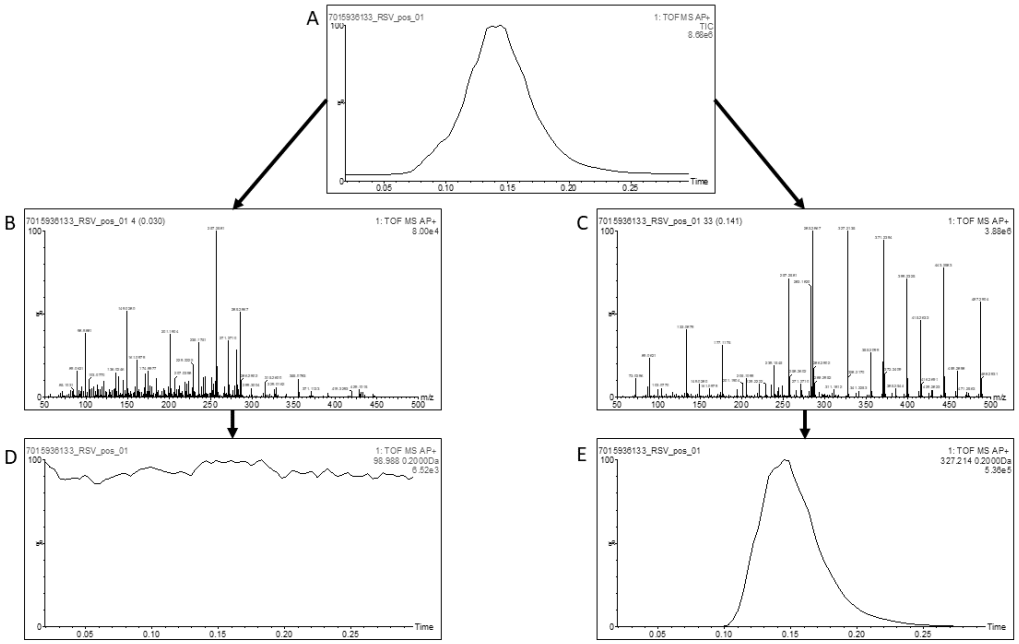


*Figure 7.2 Décomposition des spectres lors d'une acquisition par LDTD-MS. **A**. Chromatogramme des ions totaux (TIC) acquis par LDTD-MS. **B**. Spectre de masse correspondant à la section 0-0,05 minutes du chromatogramme. **C**. Spectre de masse correspondant à l'ensemble du chromatogramme. **D**. Chromatogramme d'un ion (m/z 98,988) ne subissant pas de désorption. Cet ion est non pertinent puisqu'il ne provient pas de l'échantillon. **E**. Chromatogramme d'un ion (m/z 327,214) subissant une désorption.*

Pour la normalisation, deux options sont à envisager : l'insertion de standards internes marqués et/ou une normalisation sur la somme de l'intensité des pics de VLM. Pour

permettre l'ajout de standards internes, il faudrait fractionner l'extrait métabolique de façon à regrouper certains types de molécules ensemble. En sélectionnant un ou quelques standards internes marqués, il serait possible d'effectuer une bonne première correction des intensités en transformant le signal sous la forme d'un ratio métabolite/standard interne. L'autre option, qui pourrait être combinée à la première, est de corriger les intensités en normalisant grâce à l'intensité des pics de VLM, tel que suggéré au chapitre 1.

Finalement, le développement d'une méthode de déconvolution pourrait réduire le nombre d'ions lors de la recherche de biomarqueurs. La déconvolution vise à regrouper ensemble les ions provenant du même métabolite. Par exemple, de par la distribution des isotopes naturels, on peut souvent observer la masse de l'ion constitué uniquement de ^{12}C ainsi qu'un certain nombre d'ions ayant un nombre croissant de ^{13}C. Ces isotopes sont séparés par un dalton pour chaque ^{13}C présent dans la structure de la molécule. La recherche de patrons isotopiques pourrait être assez simple malgré la complexité du spectre étant donné l'utilisation d'un spectromètre de masse à haute résolution. Cette étape permettrait donc d'éliminer un certain nombre d'ions dont l'intensité à travers les échantillons est parfaitement corrélée avec celle de son parent. Lors de la recherche de biomarqueurs, un filtre visant à garder seulement les ions pour lesquels une déconvolution a pu être appliquée permettrait d'éliminer le bruit présent dans les spectres. Il y a tout de même un risque de perdre certains métabolites dont l'intensité est trop faible pour produire le signal de l'isotope de ^{13}C.

En optimisant les méthodes de préparation de données, on pourra augmenter la qualité des données employées pour la recherche de biomarqueurs et ainsi accroître les chances de succès et les performances des modèles construits. Comme il s'agit de transformations *in silico*, aucune analyse instrumentale supplémentaire n'est nécessaire et leurs impacts peuvent être mesurés directement sur les performances des modèles lors de la recherche de biomarqueurs.

Apprentissage automatique en métabolomique

L'apprentissage automatique en métabolomique est majoritairement employé pour la recherche de biomarqueurs. Le développement d'algorithmes spécifiquement conçus dans cet objectif, la combinaison de plusieurs sources de données et l'utilisation de réseaux de neurones profonds pour la construction de modèles de classification sont des champs

d'études qui pourraient avoir un impact significatif sur la recherche en métabolomique à haut débit.

Au niveau de l'apprentissage automatique, la machine à vecteur de support linéaire et la forêt aléatoire sont les algorithmes les plus couramment utilisés en métabolomique non ciblée. La détection d'un grand nombre de potentiels biomarqueurs et les performances de classification de ces algorithmes expliquent leur usage dans ce contexte. D'un autre côté, l'arbre de décision et la machine à couverture d'ensemble (SCM) offrent de très bonnes performances, mais ils limitent grandement le nombre de métabolites employés dans le processus de classification. Ils sont donc très appropriés pour la recherche d'un petit ensemble de marqueurs permettant la classification. Toutefois, lorsqu'on cherche à comprendre les processus biologiques liés à un phénotype, ce petit ensemble de marqueurs peut être insuffisant. Les données de métabolomique non ciblée sont caractérisées par un petit nombre d'exemples, un grand nombre de variables et une corrélation élevée entre plusieurs de ces variables. Le développement de nouveaux algorithmes d'apprentissage automatique spécifiquement conçus pour bien performer sur ce type de données pourrait améliorer les retombées d'études de métabolomique non ciblée, peu importe le débit d'analyse. Notamment, nous avons dernièrement débuté le développement d'un nouvel algorithme, baptisé Random SCM, qui vise à regrouper les points forts de la machine à couverture d'ensemble (SCM) et de la forêt aléatoire (Random Forest) dans un seul algorithme. Le SCM permet de trouver un très petit nombre de métabolites caractérisant un phénotype alors que le Random Forest permet de trouver plusieurs métabolites permettant de caractériser un phénotype tout en minimisant le nombre d'erreurs de classification. Grossièrement, nous croyons qu'en remplaçant l'apprenant faible du Random Forest par des SCM, nous allons obtenir un vote de majorité de SCM, chacun optimisé sur une partie des données. En étudiant les règles de classification des nombreux SCM produits, il serait possible d'identifier les variables les plus importantes dans le processus décisionnel. Nos résultats préliminaires avec le Random SCM montrent qu'il propose une précision de prédiction similaire à celle du Random Forest sur le problème de classification de l'influenza présenté au chapitre 5 tout en réduisant le surapprentissage comparativement à ce dernier. On peut donc croire que les métabolites sélectionnés par le Random SCM sont plus pertinents que ceux utilisés par le Random Forest. L'étude du comportement de l'algorithme en fonction des valeurs d'hyperparamètres (c.-à-d. le paramètre de compromis (p), le nombre d'estimateurs, la fraction d'exemples et de variables employés pour entraîner

chaque sous-modèle) ainsi que l'analyse des règles de décision et des propriétés théoriques de l'algorithme sont encore à faire, mais les résultats préliminaires sont prometteurs.
Une autre approche méritant d'être étudiée serait la réduction de dimensionnalité des données avant la construction de modèles par apprentissage automatique. Les travaux de Tološi *et al.* sur l'emploi d'algorithme de classification en présence de variables corrélées, une situation courante dans les contextes biologiques, mériteraient d'être étudiés. [124] La capacité du Random SCM et d'autres approches à résister au *biais de corrélation* serait nécessaire afin de ne pas laisser de côté prématurément des métabolites lors la recherche de biomarqueurs.

L'ajout d'information biologique dans les algorithmes de recherche de biomarqueurs pourrait aussi aider à construire, dès la première itération, un modèle simple avec une haute pertinence biologique. En augmentant la quantité d'information fournie à un algorithme, il pourrait être possible de favoriser la construction de modèles employant des métabolites provenant des mêmes chemins métaboliques, des mêmes fractions d'extraction ou d'un seul mode d'ionisation. Le même principe s'appliquerait si l'on voulait combiner les résultats d'analyses obtenus par plusieurs sources d'ionisation. L'emploi d'algorithmes multivues pour la recherche de biomarqueurs permettrait de combiner des sources d'information et de séparer les ions en vues liées aux chemins métaboliques ou aux caractéristiques physicochimiques. Cela permettrait notamment de se rapprocher de la réalité biologique où les métabolites d'un même chemin métabolique ont des structures similaires.

Les algorithmes pour la recherche de biomarqueurs employés dans cette thèse permettent de construire des modèles de classification liés aux phénotypes tout en permettant l'étude des métabolites employés dans le processus de classification. Contrairement aux algorithmes d'apprentissage automatique classiques, les réseaux de neurones profonds sont capables d'apprendre leur propre représentation des données, mais sont difficilement interprétables. Dans l'objectif de construire un modèle de classification non interprétable, on pourrait éliminer l'ensemble des étapes de préparation des données et utiliser simplement un réseau de neurones qui apprendrait à partir des données brutes. Un réseau de neurones à convolution serait tout indiqué pour cette tâche puisqu'ils performent extrêmement bien sur les images et les signaux. Les filtres à convolution effectueraient la recherche de pics et l'interprétation des données multidimensionnelles (m/z, temps de rétention, mobilité ionique) afin de permettre la classification. Ce faisant, le réseau de neurones serait le seul outil

nécessaire pour l'analyse des données et le réseau pourrait apprendre une transformation optimale des données dans le but d'effectuer la classification, sans biais humain. L'utilisation d'une telle approche se ferait au détriment de l'interprétabilité du modèle. L'identification des biomarqueurs serait difficile, mais le modèle pourrait tout de même être employé tel quel afin de classifier de nouveaux échantillons dans des domaines où l'usage d'un modèle « boîte noire » est possible et accepté. Les travaux de Seddiki *et al.* ont montré que l'utilisation d'un réseau de neurones à convolution sur des données de métabolomique était possible et offrait de très bonnes performances. [73] Malheureusement, les données fournies à l'algorithme étaient des spectres transformés par *binning*. Ce faisant, une partie de l'information pouvant être employée par le réseau de neurones dans son processus décisionnel (par exemple la forme des pics et la précision des mesures de masse) ne lui avait pas été fournie, limitant la puissance de l'approche. Date *et al.,* quant à eux, ont utilisé un réseau de neurones pour la sélection de biomarqueurs en calculant la baisse de précision moyenne dans une expérience de métabolomique à partir de spectres de résonnance magnétique nucléaire. Une approche similaire pourrait probablement permettre la sélection de variables importantes à partir de spectres de masse[125].

L'apprentissage automatique par les réseaux de neurones profonds offre aussi d'autres possibilités dans le domaine de la métabolomique. Par exemple, Wei *et al.* ont produit un modèle permettant la prédiction du spectre de masse d'une molécule afin d'augmenter synthétiquement le contenu des bases de données. [126] La combinaison de telles approches avec DeepCCS pourrait considérablement aider à l'identification des biomarqueurs.

Le développement de nouveaux algorithmes et de nouvelles approches d'analyse pour la construction de modèles de classification ainsi que pour la recherche et l'identification de biomarqueurs est très prometteur. La conception de nouveaux outils d'apprentissage automatique qui seront optimisés pour les données de métabolomique pourrait augmenter les retombées d'études à grande échelle. De plus, des outils généralistes comme Weka ou plus spécialisés tels que MetaboDashboard, qui offrent des interfaces graphiques et des solutions simples, aideront à propager l'utilisation de l'apprentissage automatique en métabolomique. [127]

L’avenir de la métabolomique à haut débit

Beaucoup reste à faire pour que la métabolomique non ciblée à haut débit devienne une part intégrale de l’arsenal quotidien du diagnostic et du pronostic au service de la biologie. L’objectif est clair et simple : permettre la détection de biomarqueurs associés à différents phénotypes avec un maximum de précision et à une vitesse encore jamais vue. L’analyse de l’incroyable masse de données ainsi produite pose un nouveau défi. Alors que l’acquisition de données s’effectue à un rythme effréné, les logiciels et outils pour analyser ces données n’ont pas suivi le pas. Dans cette thèse, nous avons proposé des solutions pour aider à résoudre ce problème.

Les domaines de la santé et de la nutrition seront particulièrement impactés par la métabolomique non ciblée à haut débit. Dans ces domaines, le nombre d’échantillons est souvent un facteur limitant. Avec l’approche proposée dans cette thèse, le suivi du profil métabolique de chaque patient deviendrait maintenant possible.

> « *We envision a time in the future when all patients will be surrounded by a virtual cloud of billions of data points, and when we will have the analytical tools to reduce this enormous data dimensionality to simple hypotheses to optimize wellness and minimize disease for each individual.* » [128]

La métabolomique non ciblée à haut débit pourrait faire de ce futur une réalité. En combinant le profil génétique à la métabolomique, les équipes de soins auraient accès à un portrait beaucoup plus complet de l’état d’un individu. Des changements métaboliques pourraient être détectés avant même l’apparition de symptômes. Ainsi, nous pourrions nous rapprocher de l’objectif initial des sciences « omiques » : comprendre et guérir toutes les maladies.

Bibliographie

1. Green, E. D., Watson, J. D. & Collins, F. S. Human Genome Project: Twenty-five years of big biology. *Nature* vol. 526 29–31 (2015).

2. Collins, F. Has the revolution arrived? *Nature* vol. 464 674–675 (2010).

3. Collins, F. S. & McKusick, V. A. Implications of the human genome project for medical science. *J. Am. Med. Assoc.* **285**, 540–544 (2001).

4. Vervoort, R. *et al.* Mutational hot spot within a new RPGR exon in X-linked retinitis pigmentosa. *Nat. Genet.* **25**, 462–466 (2000).

5. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. & Ciccodicola, A. Transcriptome profiling in human diseases: New advances and perspectives. *International Journal of Molecular Sciences* vol. 18 (2017).

6. Payne, S. H. The utility of protein and mRNA correlation. *Trends in Biochemical Sciences* vol. 40 1–3 (2015).

7. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Letters* vol. 583 3966–3973 (2009).

8. Bauernfeind, A. L. & Babbitt, C. C. The predictive nature of transcript expression levels on protein expression in adult human brain. *BMC Genomics* **18**, 322 (2017).

9. Ahmed, A. M. History of diabetes mellitus. *Saudi Medical Journal* vol. 23 373–378 (2002).

10. Rinschen, M. M., Ivanisevic, J., Giera, M. & Siuzdak, G. Identification of bioactive metabolites using activity metabolomics. *Nat. Rev. Mol. Cell Biol.* (2019) doi:10.1038/s41580-019-0108-4.

11. Mishra, P. & Ambs, S. Metabolic Signatures of Human Breast Cancer. *Mol. Cell. Oncol.* **2**, 37–41 (2015).

12. Hori, S. *et al.* A metabolomic approach to lung cancer. *Lung Cancer* **74**, 284–292 (2011).

13. Armitage, E. G. & Southam, A. D. Monitoring cancer prognosis, diagnosis and treatment efficacy using metabolomics and lipidomics. *Metabolomics* **12**, 1–15 (2016).

14. Banoei, M. M. *et al.* Plasma metabolomics for the diagnosis and prognosis of H1N1 influenza pneumonia. *Crit. Care* **21**, 97 (2017).

15. Campos, A. I. & Zampieri, M. Metabolomics-Driven Exploration of the Chemical Drug Space to Predict Combination Antimicrobial Therapies. *Mol. Cell* **74**, 1291-1303.e6 (2019).

16. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine.

Clinical Pharmacology and Therapeutics vol. 92 414–417 (2012).

17. Jin, Q. *et al.* Metabolomics and microbiomes as potential tools to evaluate the effects of the mediterranean diet. *Nutrients* vol. 11 (2019).

18. Tebani, A. & Bekri, S. Paving the way to precision nutrition through metabolomics. *Frontiers in Nutrition* vol. 6 (2019).

19. Le Boucher, C. *et al.* LC–HRMS fingerprinting as an efficient approach to highlight fine differences in cheese metabolome during ripening. *Metabolomics* **11**, 1117–1130 (2015).

20. Consonni, R. & Cagliani, L. R. The potentiality of NMR-based metabolomics in food science and food authentication assessment. *Magn. Reson. Chem.* **57**, 558–578 (2019).

21. Verplanken, K. *et al.* Rapid evaporative ionization mass spectrometry for high-throughput screening in food analysis: The case of boar taint. (2017) doi:10.1016/j.talanta.2017.03.056.

22. Villiers, F. *et al.* Investigating the plant response to cadmium exposure by proteomic and metabolomic approaches. *Proteomics* **11**, 1650–1663 (2011).

23. Viant, M. R., Pincetich, C. A. & Tjeerdema, R. S. Metabolic effects of dinoseb, diazinon and esfenvalerate in eyed eggs and alevins of Chinook salmon (Oncorhynchus tshawytscha) determined by 1H NMR metabolomics. *Aquat. Toxicol.* **77**, 359–371 (2006).

24. Guy, C., Kaplan, F., Kopka, J., Selbig, J. & Hincha, D. K. Metabolomics of temperature stress. *Physiologia Plantarum* vol. 132 220–235 (2008).

25. Beale, D. J. *et al.* Application of metabolomics to understanding biofilms in water distribution systems: a pilot study. *Biofouling* **29**, 283–294 (2013).

26. van Ravenzwaay, B. *et al.* The use of metabolomics for the discovery of new biomarkers of effect. *Toxicol. Lett.* **172**, 21–28 (2007).

27. Lankadurai, B. P., Nagato, E. G. & Simpson, M. J. Environmental metabolomics: An emerging approach to study organism responses to environmental stressors. *Environmental Reviews* vol. 21 180–205 (2013).

28. Zayed, M. A. & Taha, M. M. Spectrophotometric Determination of Glucose in Pure Form and in Human Embryos' Culture Medium Using Selective Reagent via Studying Their Reaction product. *J. Pharm. Appl. Chem* **4**, 1 (2018).

29. Arneson, W. L. & Arneson, D. L. Current Methods for Routine Clinical Laboratory Testing of Vitamin D Levels. *Lab. Med.* **44**, e38–e42 (2013).

30. Food and Drug Administration (FDA). *Guidance for Industry Analytical Procedures and Methods Validation*. *Federal Register* vol. 513 (2015).

31. Broadhurst, D. *et al.* Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* vol. 14 (2018).

32. Gross H., J. *Mass Spectrometry Second Edition*. (Springer, 2011). doi:10.1007/978-3-642-10711-5.

33. Ho, C. S. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin. Biochem. Rev.* **24**, 3–12 (2003).

34. Wang, J. N., Zhou, Y., Zhu, T. Y., Wang, X. & Guo, Y. L. Prediction of acute cellular renal allograft rejection by urinary metabolomics using MALDI-FTMS. *J. Proteome Res.* **7**, 3597–3601 (2008).

35. Jaskolla, T. W. & Karas, M. Compelling evidence for lucky survivor and gas phase protonation: The unified MALDI analyte protonation mechanism. *J. Am. Soc. Mass Spectrom.* **22**, 976–988 (2011).

36. Picard, P., Letarte, S., Lacoursière, J. & Auger, S. Label free high throughput screening of amino acid based assays: old tricks, new speed using LDTD-MS/MS. in *ASMS 2017* 1 (2017).

37. Takáts, Z., Wiseman, J. M., Gologan, B. & Cooks, R. G. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science (80-.).* **306**, 471–473 (2004).

38. Wishart, D. S. *et al.* HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **37**, (2009).

39. Dodds, J. N. & Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **30**, 2185–2195 (2019).

40. Krishnan, S. *et al.* Instrument and process independent binning and baseline correction methods for liquid chromatography-high resolution-mass spectrometry deconvolution. *Anal. Chim. Acta* **740**, 12–19 (2012).

41. Jeffries, N. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics* **21**, 3066–3073 (2005).

42. Gu, H. W. *et al.* Solving signal instability to maintain the second-order advantage in the resolution and determination of multi-analytes in complex systems by modeling liquid chromatography-mass spectrometry data using alternating trilinear decomposition method assisted w. *J. Chromatogr. A* **1407**, 157–168 (2015).

43. Li, L. *et al.* An alignment algorithm for LC-MS-based metabolomics dataset assisted by MS/MS information. (2017) doi:10.1016/j.aca.2017.07.058.

44. Nordström, A., O'Maille, G., Qin, C. & Siuzdak, G. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: Quantitative analysis of endogenous and exogenous metabolites in human serum. *Anal. Chem.* **78**, 3289–3295 (2006).

45. Zhang, J., Gonzalez, E., Hestilow, T., Haskins, W. & Huang, Y. Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry. *Curr. Genomics* **10**, 388–401 (2009).

46. Mizuno, H. *et al.* The great importance of normalization of LC–MS data for highly-accurate non-targeted metabolomics. *Biomed. Chromatogr.* **31**, e3864 (2017).

47. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, (2018).

48. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* vol. 34 828–837 (2016).

49. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).

50. Samuel, A. L. Eight-move opening utilizing generalization learning. *IBM J.* **3**, 210–229 (1959).

51. Burkov, A. *The Hundred-Page Machine Learning Book*. (Andriy Burkov, 2019).

52. Xia, J. *et al.* Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* **9**, 280–299 (2013).

53. Mendez, K. M., Reinke, S. N. & Broadhurst, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **15**, 1–15 (2019).

54. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta* vol. 879 10–23 (2015).

55. Ghaffari, M. H. *et al.* Metabolomics meets machine learning: Longitudinal metabolite profiling in serum of normal versus overconditioned cows and pathway analysis. *J. Dairy Sci.* **102**, 11561–11585 (2019).

56. Melo, C. F. O. R. *et al.* A Machine Learning Application Based in Random Forest for Integrating Mass Spectrometry-Based Metabolomic Data: A Simple Screening Method for Patients With Zika Virus. *Front. Bioeng. Biotechnol.* **6**, (2018).

57. Menze, B. H. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, 213 (2009).

58. Gaul, D. A. *et al.* Highly-accurate metabolomic detection of early-stage ovarian cancer. *Sci. Rep.* **5**, 16351 (2015).

59. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning*. *Springer Texts in Statistics* (2013).

60. Hastie, T., Tibshirani, R. J. & Friedman, J. *The Elements of Statistical Learning*. *Elements* vol. 1 (2009).

61. Drouin, A. *et al.* Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* **9**, 1–13 (2019).

62. Ernst, M. *et al.* MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *bioRxiv* 654459 (2019) doi:10.1101/654459.

63. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738–13743 (2016).

64. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. **112**, 12580–12585 (2015).

65. Zhou, Z., Tu, J., Xiong, X., Shen, X. & Zhu, Z.-J. LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility–Mass Spectrometry-Based Lipidomics. *Anal. Chem.* **89**, 9559–9566 (2017).

66. Zhou, Z., Shen, X., Tu, J. & Zhu, Z.-J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal. Chem.* **88**, 11084–11091 (2016).

67. Zhou, Z., Shen, X., Tu, J. & Zhu, Z.-J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal. Chem.* **88**, 11084–11091 (2016).

68. Ross, D. H., Cho, J. H. & Xu, L. Breaking down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Anal. Chem.* **92**, 4548–4557 (2020).

69. Bijlsma, L. *et al.* Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Anal. Chem.* **89**, 6583–6589 (2017).

70. Mollerup, C. B., Mardal, M., Dalsgaard, P. W., Linnet, K. & Barron, L. P. Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. *J. Chromatogr. A* **1542**, 82–88 (2018).

71. Picache, J. A. A. *et al.* Collision Cross Section Compendium to Annotate and Predict Multi-omic Compound Identities. *Chem. Sci.* (2019) doi:10.1039/C8SC04396E.

72. Rochat, B. Proposed Confidence Scale and ID Score in the Identification of Known-Unknown Compounds Using High Resolution MS Data. *J. Am. Soc. Mass Spectrom.* **28**, 709–723 (2017).

73. Seddiki, K. *et al.* Towards CNN Representations for Small Mass Spectrometry Data Classification: From Transfer Learning to Cumulative Learning. *bioRxiv* (2020) doi:10.1101/2020.03.24.005975.

74. Li, L. P. *et al.* Applications of ambient mass spectrometry in high-throughput screening. *Analyst* **138**, 3097–3103 (2013).

75. de Raad, M., Fischer, C. R. & Northen, T. R. High-throughput platforms for metabolomics. *Curr. Opin. Chem. Biol.* **30**, 7–13 (2016).

76. Phytronix. Luxon Ion Source® - Fastest process for mass spectrometry | Phytronix. https://phytronix.com/luxon-learn-more/.

77. Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D. & Lu, L. J. Computational and statistical analysis of metabolomics data. *Metabolomics* **11**, 1492–1513 (2015).

78. Brochu, F. *et al.* Mass spectra alignment using virtual lock-masses. *Sci. Rep.* **9**, (2019).

79. Heinemann, J. Machine learning in untargeted metabolomics experiments. *Methods Mol. Biol.* **1859**, 287–299 (2019).

80. Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K. & Blank, L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **10**, 243 (2020).

81. Altmann, A., Tolo¸si, L., Tolo¸si, T., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. **26**, 1340–1347 (2010).

82. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* vols 2017-Decem 4766–4775 (2017).

83. Chong, J. *et al.* MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **46**, W486–W494 (2018).

84. Xia, J. & Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.* **6**, 743–760 (2011).

85. Wilkinson, M. D. Comment: The FAIR Guiding Principles for scientific data management and stewardship. (2016) doi:10.1038/sdata.2016.18.

86. Paget, J. *et al.* Global mortality associated with seasonal influenza epidemics: New burden estimates and predictors from the GLaMOR Project. *J. Glob. Health* **9**, (2019).

87. Santé Canada. Grippe (influenza) : Pour les professionnels de la santé - Canada.ca. https://www.canada.ca/fr/sante-publique/services/maladies/grippe-influenza/professionnels-sante.html (2020).

88. Eisenreich, W., Rudel, T., Heesemann, J. & Goebel, W. How viral and intracellular bacterial pathogens reprogram the metabolism of host cells to allow their intracellular replication. *Frontiers in Cellular and Infection Microbiology* vol. 9 42 (2019).

89. Samuelsson, B. Arachidonic acid metabolism: Role in inflammation. in *Zeitschrift fur Rheumatologie* vol. 50 3–6 (1991).

90. HIGGINS, A. J. & LEES, P. The acute inflammatory process, arachidonic acid metabolism and the mode of action of anti-inflammatory drugs. *Equine Vet. J.* **16**, 163–175 (1984).

91. Chandrasekharan, J. A. & Sharma-Walia, N. Arachidonic acid derived lipid mediators influence Kaposi's sarcoma-associated herpesvirus infection and pathogenesis. *Frontiers in Microbiology* vol. 10 358 (2019).

92. Centers for Diseases Control and Prevention. Information on Rapid Molecular Assays , RT-PCR , and other Molecular Assays for Diagnosis of Influenza Virus Infection Influenza Testing of Hospitalized Patients. 1–8 https://www.cdc.gov/flu/professionals/diagnosis/molecular-assays.htm (2018).

93. Chandler, J. D. *et al.* Metabolic pathways of lung inflammation revealed by high-resolution metabolomics (HRM) of H1N1 influenza virus infection in mice. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* **311**, R906–R916 (2016).

94. Cui, L., Fang, J., Eng, ‡, Ooi, E. & Lee, Y. H. Serial Metabolome Changes in a Prospective Cohort of Subjects with Influenza Viral Infection and Comparison with Dengue Fever. (2017) doi:10.1021/acs.jproteome.7b00173.

95. Cajka, T. *et al.* Evaluation of direct analysis in real time ionization-mass spectrometry (DART-MS) in fish metabolomics aimed to assess the response to dietary supplementation. *Talanta* **115**, 263–270 (2013).

96. Pedregosa, F. *et al. Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

97. Xia, J., Broadhurst, D. I., Wilson, M. & Wishart, D. S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **9**, 280–299 (2013).

98. Lynn, K.-S. *et al.* Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Anal. Chem.* **87**, 2143–51 (2015).

99. Nguyen, D. H., Nguyen, C. H. & Mamitsuka, H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief. Bioinform.* (2018) doi:10.1093/bib/bby066.

100. Blaženović, I. *et al.* Increasing Compound Identification Rates in Untargeted Lipidomics Research with Liquid Chromatography Drift Time-Ion Mobility Mass Spectrometry. *Anal. Chem.* **90**, 10758–10764 (2018).

101. May, J. C. & McLean, J. A. Advanced Multidimensional Separations in Mass Spectrometry: Navigating the Big Data Deluge. *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **9**, 387–409 (2016).

102. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **27**, 1897–1905 (2016).

103. Hinnenkamp, V. *et al.* Comparison of CCS Values Determined by Traveling Wave Ion Mobility Mass Spectrometry and Drift Tube Ion Mobility Mass Spectrometry. *Anal. Chem.* acs.analchem.8b02711 (2018) doi:10.1021/acs.analchem.8b02711.

104. Stow, S. M. *et al.* An Interlaboratory Evaluation of Drift Tube Ion Mobility-Mass Spectrometry Collision Cross Section Measurements. *Anal. Chem.* **89**, 9048–9055 (2017).

105. Paglia, G. *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Anal. Chem.* **87**, 1137–44 (2015).

106. Paglia, G. *et al.* Ion mobility derived collision cross sections to support metabolomics applications. *Anal. Chem.* **86**, 3985–93 (2014).

107. Nichols, C. M. *et al.* Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility-Mass Spectrometry. *Anal.*

Chem. acs.analchem.8b04322 (2018) doi:10.1021/acs.analchem.8b04322.

108. Picache, J. A. *et al.* Collision cross section compendium to annotate and predict multi-omic compound identities. *Chem. Sci.* (2019) doi:10.1039/c8sc04396e.

109. Colby, S. M. *et al.* ISiCLE: A molecular collision cross section calculation pipeline for establishing large in silico reference libraries for compound identification. (2018).

110. Zhou, Z., Tu, J. & Zhu, Z.-J. Advancing the large-scale CCS database for metabolomics and lipidomics at the machine-learning era. *Curr. Opin. Chem. Biol.* **42**, 34–41 (2018).

111. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

112. Ehteshami Bejnordi, B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318**, 2199 (2017).

113. Zhao, K. & So, H.-C. Using Drug Expression Profiles and Machine Learning Approach for Drug Repurposing. in *Methods in molecular biology (Clifton, N.J.)* vol. 1903 219–237 (2019).

114. LeCun, Y., Kavukcuoglu, K. & Farabet, C. Convolutional networks and applications in vision. in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* 253–256 (IEEE, 2010). doi:10.1109/ISCAS.2010.5537907.

115. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).

116. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).

117. Zhou, Z., Xiong, X. & Zhu, Z.-J. MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics* **33**, 2235–2237 (2017).

118. Zheng, X. *et al.* A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry. *Chem. Sci.* **8**, 7724–7736 (2017).

119. May, J. C. *et al.* Conformational ordering of biomolecules in the gas phase: nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer. *Anal. Chem.* **86**, 2107–16 (2014).

120. Kwon, S. & Yoon, S. End-to-end representation learning for chemical-chemical interaction prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **16**, 1436–1447 (2019).

121. Hinnenkamp, V. *et al.* Comparison of CCS Values Determined by Traveling Wave Ion Mobility Mass Spectrometry and Drift Tube Ion Mobility Mass Spectrometry. *Anal. Chem.* **90**, 12042–12050 (2018).

122. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).

123. Plante, P. L. *et al.* High-Throughput Cheese Metabolomics Using Diverse Ion Sources. in *Proceeding of the 65th ASMS Conference on Mass Spectrometry* (2017).

124. Toloşi, L. & Lengauer, T. Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics* **27**, 1986–1994 (2011).

125. Date, Y. & Kikuchi, J. Application of a Deep Neural Network to Metabolomics Studies and Its Performance in Determining Important Variables. *Anal. Chem.* **90**, 1805–1810 (2018).

126. Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks. *ACS Cent. Sci.* **5**, 700–708 (2019).

127. Heinemann, J. Machine learning in untargeted metabolomics experiments. in *Methods in Molecular Biology* vol. 1859 287–299 (Humana Press Inc., 2019).

128. Hood, L. & Rowen, L. The human genome project: Big science transforms biology and medicine. *Genome Med.* **5**, 79 (2013).

129. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass Spectrometry-Based Metabolomics. *Mass Spectrom. Rev.* **26**, 51–78 (2007).

130. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology* vol. 12 483–490 (2008).

131. Fenselau, C. & Demirev, F. A. Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrometry Reviews* vol. 20 157–171 (2001).

132. Caprioli, R. M., Farmer, T. B. & Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.* **69**, 4751–4760 (1997).

133. Cox, J. & Mann, M. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. *Annu. Rev. Biochem.* **80**, 273–299 (2011).

134. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 6233–6237 (1986).

135. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. in *Electrophoresis* vol. 20 3551–3567 (Wiley-VCH Verlag, 1999).

136. Alonso, A., Marsal, S. & Julià, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology* vol. 3 (2015).

137. Dunn, W. B. *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9**, 44–66 (2013).

138. Römpp, A. & Karst, U. Current trends in mass spectrometry imaging mass spectrometry imaging. *Anal. Bioanal. Chem.* **407**, 2023–2025 (2015).

139. Huang, M.-Z., Yuan, C.-H., Cheng, S.-C., Cho, Y.-T. & Shiea, J. Ambient Ionization

Mass Spectrometry. *Annu. Rev. Anal. Chem.* **3**, 43–65 (2010).

140. Semmes, O. J. *et al.* Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin. Chem.* **51**, 102–112 (2005).

141. Tibshirani, R. *et al.* Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* **20**, 3034–3044 (2004).

142. Tracy, M. B. *et al.* Precision enhancement of MALDI-TOF MS using high resolution peak detection and label-free alignment. *Proteomics* **8**, 1530–1538 (2008).

143. Barry, J. A., Robichaud, G. & Muddiman, D. C. Mass recalibration of FT-ICR mass spectrometry imaging data using the average frequency shift of ambient ions. *J. Am. Soc. Mass Spectrom.* **24**, 1137–1145 (2013).

144. Olsen, J. V *et al.* Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010–2021 (2005).

145. Psychogios, N. *et al.* The human serum metabolome. *PLoS One* **6**, (2011).

146. Bouatra, S. *et al.* The Human Urine Metabolome. *PLoS One* **8**, (2013).

147. Freund, Y. & Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).

148. Cortes, C., Vapnik, V. & Saitta, L. *Support-Vector Networks Editor*. *Machine Leaming* vol. 20 (1995).

149. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees*. vol. 8 (2017).

150. Marchand, M. & Shawe-Taylor, J. The Set Covering Machine. *J. Mach. Learn. Res.* **3**, 723–746 (2002).

151. Gammerman, A., Vovk, V. & Vapnik, V. Learning by Transduction. (2013).

Annexe A- Mass spectra alignment using virtual lock-masses

8.1 Résumé

La spectrométrie de masse est une méthode très appréciée pour évaluer le contenu métabolique d'un échantillon biologique. Les avancées récentes des technologies d'ionisation rapides telles que la Désorption Thermique par Diode Laser (LDTD) et l'Analyse Directe en Temps Réel (DART) permettent aujourd'hui d'effectuer des acquisitions par spectrométrie de masse à haut débit. Elles peuvent donc être utilisées pour des analyses d'échantillons à grande échelle visant la comparaison de population. En pratique, plusieurs facteurs tels que l'environnement, le protocole et même l'instrument peuvent insérer des variations mineures entre les spectres mesurés, rendant les analyses comparatives automatisées difficiles. Dans ces travaux, nous proposons une séquence d'algorithmes pour corriger les variations entre les spectres. Les algorithmes corrigent l'ensemble des spectres en identifiant des pics qui sont communs à tous et desquels une correction spécifique à chaque spectre est calculée. Nous montrons que ces algorithmes augmentent la comparabilité de grands ensembles de spectres, facilitant les analyses comparatives telles que l'apprentissage automatique.

8.2 Abstract

Mass spectrometry is a valued method to evaluate the metabolomics content of a biological sample. The recent advent of rapid ionization technologies such as Laser Diode Thermal Desorption (LDTD) and Direct Analysis in Real Time (DART) has rendered high-throughput mass spectrometry possible. It is used for large-scale comparative analysis of populations of samples. In practice, many factors resulting from the environment, the protocol, and even the instrument itself, can lead to minor discrepancies between spectra, rendering automated comparative analysis difficult. In this work, a sequence/pipeline of algorithms to correct variations between spectra is proposed. The algorithms correct multiple spectra by identifying peaks that are common to all and, from those, computes a spectrum-specific correction. We show that these algorithms increase comparability within large datasets of spectra, facilitating comparative analysis, such as machine learning.

8.3 Introduction

Mass spectrometry (MS) is a widely used technique for acquiring data on the metabolome or the proteome of individuals [129,130]. Proteomics applications can consist, among others, of typing of microbial organisms, imaging MS, quantitative comparisons, and peptide sequencing [131–135]. For metabolomics applications, the two main approaches fall into the categories of targeted and untargeted studies. In comparison with targeted studies, untargeted studies acquire data using a shotgun approach. Therefore, this type of study is a good option for novel biomarker discovery and hypothesis generation [136,137].

Through recent years, novel ionization technologies have emerged, facilitating the high-throughput acquisition of mass spectra [138]. Technologies such as Laser Diode Thermal Desorption (LDTD) or Direct Analysis in Real Time (DART), allow for the rapid acquisition of large datasets. These methods often preclude or bypass the time separation process used in Liquid Chromatography (LC) or Gas Chromatography (GC) [139]. Thus, without any time separation, a single mass spectrum will often be represented as lists of peaks, composed of the mass-to-charge ratio of the ion (m/z value) and its intensity. Although, in this case, a single mass spectrum will be more complex and composed of more peaks and compounds than a spectrum with a time-separation method.

With the rise of larger datasets, multiple problems of comparability between spectra have emerged. Datasets are acquired in multiple batches over numerous days, on different instruments in multiple locations, with recalibrations of the instruments occurring between batches [140]. These factors induce variations in the spectra that hinders their comparison.

In the past, three algorithms have been proposed to address this problem, mainly affecting Time-of-Flight mass spectrometers. These include the work of Tibshirani et al., Jeffries and Tracy et al.[41,141,142]. Tibshirani's algorithm relies on a clustering algorithm to align peaks that are present in multiple spectra and picks them for further statistical analyses. However, unlike the algorithms proposed in this article, it does not address the problem of inter-batch variations. Jeffries' algorithm is more appropriate for this problem. This method uses cubic splines to recalibrate spectra, based on the shifts between observed peaks and known reference masses. A similar algorithm has been proposed by Barry et al. for Fourier-Transform Mass Spectrometry [143]. This approach uses ambient ions in order to correct the spectra using known reference masses. One limitation of these algorithms is that they

require known reference masses. The algorithm presented in this work alleviates this constraint, by automatically detecting such reference points. Another algorithm of interest for MALDI-ToF spectra has been proposed by Tracy et al. [142]. In this case, commonly occurring peaks within the dataset are used to correct the spectra and determine the binning distance used. However, this method computes a single constant correction factor for the entire spectrum, while the method proposed in this work computes correction factors that vary across the m/z axis of the spectra in order to obtain a more accurate correction.

The algorithms proposed in this article aim to render spectra more comparable prior to peak selection and statistical analyses. We draw inspiration from the internal lock mass approach and exploit the fact that spectra of samples of the same nature (i.e., blood plasma samples, urine samples, etc.) are very likely to share common peaks (i.e., compounds that are present in each sample). The internal lock mass approach consists of introducing a standard compound along with the sample into the ion source [144]. This known compound can then be used to correct instrumental drift in m/z values, potentially in real-time. For example, human blood plasma contains compounds, such as glucose and amino acids [145]. Similarly, urine contains urea, creatinine, citric acid, and many more [146]. Hence, we propose to correct the spectra based on the position of peaks that are detected to be consistently present in samples of the same nature. We call these peaks “virtual lock masses" (VLM) and propose an algorithm to detect them automatically. This idea is similar to the one proposed by Barry et al., but the peaks are not limited to ambient ions [143]. In this work, we show that our algorithm allows the detection of tens to hundreds of peaks that can be used as reference points to re-align the spectra. These points will serve to reduce inter-batch and intra-batch variations in the spectra but will not correct the spectra to the true m/z values of the ions. However, our approach is fully compatible with the classical lock mass approach, which can be used complementarily. Moreover, we show that a slight modification to the VLM detection algorithm can produce an alignment algorithm that can be used to further correct the spectra.

Hence, our key contributions are: an algorithm that automatically detects reference points in mass spectra, an algorithm that corrects the spectra based on these points, and an alignment algorithm to align large sets of spectra. The next section describes the details of the algorithms and their implementation and we then present results supporting the accuracy of our reference point detection algorithm. Moreover, we show that the proposed algorithms

yield classifiers with increased accuracy in the context of machine learning analysis performed on ToF mass spectra. Finally, we discuss these results and their implications.

8.4 Methods

In this section, we present the mathematical basis of the proposed methodology. First, the problem of virtual lock-mass identification is addressed. A formal definition of VLM peaks is introduced, along with an highly efficient algorithm capable of identifying such peaks in a set of mass spectra. Second, a methodology for correcting mass spectra based on a set of identified virtual lock masses is described. Third, an algorithm for mass spectra alignment based on the previous algorithm is proposed. Finally, the datasets used and the experimental methodologies are presented. Note that the algorithms are designed to be applied to partially pre-processed spectra, specifically processed to centroid format.

8.4.1 Definitions

Let us first recall that a set is an un-ordered collection of elements whereas a sequence is an ordered collection of elements. Hence, in a sequence we have a first element, a second element, and so on. If A is a sequence or a set, $|A|$ denotes the number of elements in A.

Let $\mathcal{S} \stackrel{\text{def}}{=} \{S_1, \ldots, S_m\}$ be a set of mass spectra. Each spectrum S_i is a sequence of peaks, where each peak is a pair (μ, ι) with an m/z value μ and a peak intensity ι. Let a window of size $2w$ centered on the peak (μ, ι) be an interval that starts at $\mu \cdot (1 - w)$ and ends at $\mu \cdot (1 + w)$. Notice that the size of the window w is relative to μ. The reason for using window sizes in relative units is that the mass measurement uncertainty of ToF mass spectrometers increases linearly with the m/z value of a peak.

Given a set $\mathcal{S}$ of mass spectra and a window size parameter w, a virtual lock mass (VLM) with respect to $(\mathcal{S}, w)$ is a point v on the m/z axis such that there exists a set $\mathcal{P}$ of peaks from $\mathcal{S}$ that satisfies the following properties

1. $\mathcal{P}$ contains exactly one peak from each spectrum in $\mathcal{S}$.
2. The average of the m/z values of the peaks in $\mathcal{P}$ is equal to v.
3. Every peak in $\mathcal{P}$ has a m/z value located in the interval $[v(1 - w), v(1 + w)]$.
4. No other peak in $\mathcal{S}$ has an m/z value that belongs to $[v(1 - w), v(1 + w)]$.
5. Every peak in $\mathcal{P}$ has an intensity in the interval $[t_a, t_b]$.

If and only if all these criteria are satisfied, we say that $\mathcal{P}$ is the set of peaks associated with the VLM v.

Note that we impose a lower intensity threshold t_a, since peaks with a lower intensity will tend to have a lower mass accuracy, and can even be confused with noise. In addition, there is also accuracy issues when the intensity of a peak is higher than the machine specifications. Consequently, we also impose an upper intensity threshold t_b. Hence, in principle, a VLM is defined only with respect to (w.r.t.) $(\mathcal{S}, w, [t_a, t_b])$. However, we will drop the reference to t_a and t_b to simplify the notation.

A crucial aspect of the definition of a VLM is the fact that it holds only w.r.t. a given window size w. Indeed, consider Figure 8.1 which represents peaks coming from three different spectra. We can observe that a first window size w_1 will correctly detect four VLM points. If the window size is too large however, we observe the case of w_2 : peaks that are further apart can be erroneously grouped into a VLM group. Moreover, w_2 can detect the first grouping of peaks within the figure as a VLM, and then the shown grouping as a second one. Thus, the same peaks would be part of two distinct VLM points. This would create ambiguity in the correction and is nonsensical. The last possible case is that of a window size that is too small. In this situation, the window would be unable to detect groups of peaks coming from each spectra of $\mathcal{S}$.

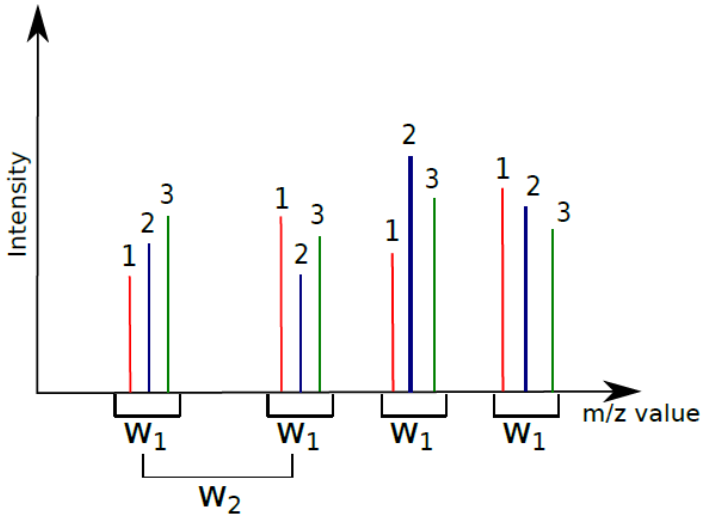


Figure 8.1 Definition of window size for the detection of VLM peaks. The peaks identified as 1, 2, and 3 are presumed to originate from three different spectra. Window size w_1 correctly detects four VLM groups. Window size w_2 however is too wide and will detect ambiguous and erroneous groups. Moreover w_2 will detect several overlapping VLM groups.

Hence, this motivates the following definition of overlapping VLM points. Given $(\mathcal{S}, w)$, a VLM v_i w.r.t. $(\mathcal{S}, w)$ is said to *overlap* with another VLM v_j (with respect to $(\mathcal{S}, w)$) if and only if there exists an intersection between the m/z interval $[v_i(1-w), v_i(1+w)]$ and the m/z interval $[v_j(1-w), v_j(1+w)]$. Moreover, we say that a VLM v w.r.t. $(\mathcal{S}, w)$ is isolated from all all other VLM with w.r.t. $(\mathcal{S}, w)$ if and only if there does not exists any other VLM v' w.r.t. $(\mathcal{S}, w)$ that overlaps with v. For a given window size w, the algorithm that we present in the next subsection identifies all isolated VLM points w.r.t. $(\mathcal{S}, w)$. Consequently, the best value for w is one for which the number of isolated VLM points is the largest.

8.4.2 An Algorithm for Virtual Lock Mass Identification

Given a set $\mathcal{S} = \{S_1, \ldots, S_m\}$ of m spectra, each peak is identified by a pair (σ, ρ) where $\sigma \in \{1, \ldots, m\}$ is the index of its spectrum of origin and $\rho \in \{1, \ldots, n_\sigma\}$ is the index of the peak in spectrum S_σ containing n_σ peaks. Given that we have a total of n peaks in $\mathcal{S}$, we have that $\sum_{\sigma=1}^{m} n_\sigma = n$. For the description of the algorithm, $\mu(\sigma, \rho)$ denotes the m/z value of peak (σ, ρ). Finally, we assume that the peaks in each spectra S_i are listed in increasing order of their m/z values.

The proposed algorithm uses two data structures: a *binary heap* and a so-called *active sequence*. A binary heap is a classical data structure used for priority queues which are useful when one wants to efficiently remove the element of highest priority in a queue. In our case, the heap will maintain, at any time, the next peak of each spectra to be processed by the algorithm. Hence, given a set $\mathcal{S}$ of m spectra, the heap generally contains a set of m peaks, where each peak belongs to a different spectrum of $\mathcal{S}$. The "priority value" for each peak (σ, ρ) in the heap is given by its m/z value $\mu(\sigma, \rho)$; a peak with the smaller mass is always on top of the heap. A heap H containing the first peak of each spectrum can thus be constructed in $O(m)$ time. Moreover, we can read the m/z value at the top of the heap in constant time; we can remove the peak (σ, ρ) of the top of the heap and replace it with the next available peak in the spectrum S_σ in $O(\log m)$ time.

The second data structure is, what we call, the *active sequence* A. At any time, A contains a sequence of peaks, listed in increasing order of their m/z values, which are currently being considered to become a VLM sequence. That data structure uses a doubly linked list L and a boolean-valued vector B of dimension m. The linked list L is actually containing the

sequence of peaks to be considered for the next VLM and the vector B is such that, at any time, $B[\sigma] = True$ if and only if a peak from spectrum S_σ is present in L. The active sequence A also maintains the m/z value μ_l of the last peak that was removed from L, the average m/z value μ_A of the peaks in L, and a copy w_A of the window size w chosen by the user. Since L is a linked list, we can read the front (first) and back (last) values of L in constant time, as well as obtaining its size (number of peaks). Removing the value at the front of L is also performed in constant time.

We now present a short description of the algorithm for virtual lock mass identification. The fully detailed description is provided in Supplementary Information.

8.4.2.1 Validation of an active sequence

For this step, we use a method, call $A.isValid()$, that returns $True$ if and only if the peaks in the active sequence A satisfies all the criteria enumerated in the definition of a VLM. A precondition for the validity of this method is that L contains only peaks that belong to distinct spectra of $\mathcal{S}$. This precondition holds initially for an empty list L and will always be maintained for each new peak inserted in A (see the next paragraph for details). Thus, this step of the algorithm checks first that the active sequence contains exactly $|\mathcal{S}|$ peaks, thus one peak from each spectrum in the set. Then, if there are still peaks in the heap, we verify that the peak at the top of H (thus, the peak immediately following the active sequence) has a m/z value that is out of the interval $[\mu_A(1-w), \mu_A(1+w)]$. Similarly, it is verified that the peak whose m/z value immediately precedes the active sequence also has an m/z value outside of $[\mu_A(1-w), \mu_A(1+w)]$. If either peak lies inside this window, then the property (4) of a VLM is violated, as the window contains more than $|\mathcal{S}|$ peaks. Finally, we ensure that the first and last peaks in the active sequence A are both within the window $[\mu_A(1-w), \mu_A(1+w)]$. If all checks pass, then the current sequence is considered a potential virtual lock mass.

8.4.2.2 Advancing the active sequence

This step tries to insert at the end of the list L of A the peak (σ, ρ) located on top of H. The insertion succeeds if the resulting A still have some probability that the peak sequence can become a VLM after zero or more future insertions. Thus, we first verify if another peak from spectrum S_σ is present in A. If that is the case, then the insertion fails. Otherwise, we compute the new value μ'_A that μ_A will have after the insertion. If the peak at the front of L

(the peak in A having the smallest m/z value) and the new peak (σ, ρ) have masses that are within the window $[\mu'_A(1-w), \mu'_A(1+w)]$, then the insertion succeeds. The peak is inserted, and H is updated by removing the peak (σ, ρ) and adding the next peak from the spectrum S_σ. Thus, this step ensures that we can insert a new peak in A and still have some probability that the sequence can become a VLM after zero or more future insertions.

Whenever we have an insertion failure, it means that the active sequence cannot become a valid VLM and that we must remove from A the peak having the smallest m/z value (which is located in the front of L) in order to have a chance that the sequence of peaks in A becomes a valid VLM.

8.4.2.3 Advancing the lower bound

This step is used to remove the peak (σ, ρ) at the front of L until a valid insertion can be made. First, it updates $B[\sigma]$ to $False$, as peak (σ, ρ) is about to be removed and no peak from S_σ will be in the active sequence A anymore. The m/z value of peak (σ, ρ) is copied in μ_l, and the peak is then removed from L. If L is empty at this point, its average m/z value μ_A is set to 0. Otherwise, μ_A is set to the average value of the peaks remaining in the active sequence.

8.4.2.4 Removing overlapping virtual lock masses

The final step of the algorithm removes all overlapping VLMs. As described in Supplementary Information, a Boolean vector (with a number of components equal to the number of VLMs found) is initialized to $False$. Then, we simply iterate over all the VLM points found and assign the corresponding vector entry to $True$ whenever a VLM point (with m/z value μ) is found such that its window $[\mu(1-w), \mu(1+w)]$ overlaps with that of its neighboring VLMs. Only the VLMs whose entry in the vector is $False$ are kept.

8.4.2.5 Main algorithm

Having described the data structures used and their methods, we are now in position to present the main algorithm for virtual lock mass detection, which is described by Algorithm 8.1. The task of this algorithm is to find all the isolated VLM points w.r.t. $(\mathcal{S}, w)$. To achieve this, the central part of the algorithm is to find the sequence

$\mathcal{U} = \langle \mu_1, \ldots, \mu_{|\mathcal{U}|} \rangle$ of all possible VLM points w.r.t. $(\mathcal{S}, w)$. This sequence may contain several pairs of overlapping VLMs. The strategy to achieve this central task is to use $A.insert(H, \mathcal{S})$ to try to insert in A (consequently in L) the next unprocessed peak of $\mathcal{S}$, which is always located on the top of the heap H.

Initially, the first peak of $\mathcal{S}$, a peak having the smallest m/z value among those in $\mathcal{S}$, gets eventually inserted in an empty A by $A.insert(H, \mathcal{S})$. Next, after verifying with $A.isValid(H)$ if the content of A satisfies the criteria to be a valid VLM sequence, we try to insert again in A the next available peak. On each insertion failure, we test if, before this insertion, the content of A was a valid VLM sequence. This is done with the Boolean variable found (which is set to $True$ as soon as the content of A is a valid VLM sequence and which is set to $False$ immediately after the average m/z value μ_A of A's content is appended to $\mathcal{U}$). Hence, for each considered peak in $L.front()$, we try to insert one more peak in L and test after the insertion if L's content is a valid VLM sequence. If we cannot insert an extra peak in L with the current peak in $L.front()$ this means that there is no possibility of finding one more VLM sequence with the current peak in $L.front()$. In that case we remove that peak from L with $A.advanceLowerBound()$ and, consequently, $L.front()$ now becomes the peak that was next to $L.front()$ in L.

Hence, with this strategy, the algorithm attempts to find the largest consecutive sub-sequence of peaks from $\mathcal{S}$ that starts with any given peak in $\mathcal{S}$ and that forms a valid VLM sequence. In addition, note that in the $else$ branch of Algorithm 8.1, we verify if H becomes empty after a successful insertion. In that case, we need to check if we can find a valid VLM sequence by incrementing sequentially the lower bound $L.front()$ and then append to $\mathcal{U}$ the first VLM found. Then, we can safely exit the while loop since any other possible VLM sequence will be a subset of the one already found. Without this else branch, a VLM sequence that ends with the last peak presented by H would be missed by the algorithm.

virtualLockMassDetection$(\mathcal{S}, w)$;
Input: $\mathcal{S} = \{S1, S2, \ldots Sm\}$, a set of mass spectra.
Input: w, a window size parameter in relative units.
Output: The sequence of all isolated VLM points with respect to $(\mathcal{S}, w)$.
Data: H, a heap initialized with $H.init(\mathcal{S})$; thus containing the first peak of each spectra in $\mathcal{S}$.
Data: A, an active sequence initialized with $A.init(H, w)$; hence initially empty.
Data: $\mathcal{U}$, a sequence of m/z values, initially empty.
$found \leftarrow False$;
while $H.empyt() = False$ **do**
    if $A.isValid(H) = True$ **then** $found = True$;
    if $A.insert(H, \mathcal{S}) = False$ **then**
        if $found = True$ **then**
            $\mathcal{U}.append(A.get\mu_A())$;
            $found \leftarrow False$;
        end
        $A.advanceLowerBound()$;
    else
        if $H.empty() = True$ **then**
            while $A.empyt() = False$ **do**
                if $A.isValid(H) = True$ **then**
                    $\mathcal{U}.append(A.get\mu_A())$;
                    break;
                end
                $A.advanceLowerBound()$;
            end
        end
    end
end
return $deleteOverlaps(\mathcal{U}, w)$;

Algorithm 8.1The virtual Lock Mass Detection Algorithm

As explained in Supplementary Information, the running time of Algorithm 8.1 (i.e., the VLM detection algorithm) is in $O(n \log m)$ for a sequence of m spectra that contains a total of n peaks. This, however, is for a fixed value of window size w. Note that in order to obtain the most accurate correction (by interpolation) for the spectra in $\mathcal{S}$, we should use the largest number of isolated (i.e., non-overlapping) VLMs we can find. Consequently, the optimal value for w is the one for which Algorithm 8.1 will give the largest number of isolated VLMs. Moreover, note that if w is too small, very few VLMs will be detected as w will not be able to cover exactly one peak per spectra. If, on the other hand, w is too large, a large number of the VLMs found in the first phase of the algorithm will overlap and the remaining isolated VLMs will be rare. Consequently, because of this “unimodal" behavior, one can generally

find rapidly the best value for w. In our case, we never needed to try more than 20 different values.

8.4.2.6 An Algorithm for Virtual Lock Mass Correction

Given a set $\mathcal{S}$ of spectra and a widow size parameter w expressed in relative units, once the sequence $\mathcal{U}$ of all isolated VLM points w.r.t. $(\mathcal{S}, w)$ has been determined, the individual spectra in $\mathcal{S}$ can be corrected in a manner similar as it is usually done with traditional lock masses. Algorithm 8.2 performs the correction needed for each peak in a spectrum $S \in \mathcal{S}$.

virtualLockMassCorrection$(S, \mathcal{V}, w)$;
Input: $S = \langle (\mu_1, i_1), (\mu_2, i_2), \ldots, (\mu_m, i_m) \rangle$, a spectrum.
Input: $\mathcal{V} = \langle v_1, v_2, \ldots, v_n \rangle$, a sequence of m/z values (VLMs) sorted in increasing order.
Input: w, a window size parameter in relative units.
Output: A spectrum $S' = \langle (\mu'_1, i_1), (\mu'_2, i_2), \ldots, (\mu'_m, i_m) \rangle$ where each μ'_j is the corrected m/z value for the peak $(\mu_j, i_j) \in S$.
Data: $a = (a_1, a_2, \ldots, a_n)$, a vector if indexes (natural numbers).
//construction of a
$i \leftarrow 1$;
for $j = 1$ **to** m **do**
    if $\mu_j \in [v_i(1-w),\ v_i(1+w)]$ **then**
        $a_i \leftarrow j$; // peak (μ_j, i_j) is associated to VLM v_i
        $i \leftarrow i + 1$;
    end
end
//correct each μ_j such that $a_1 \leq j \leq a_n$
$j \leftarrow a_1$;
$i \leftarrow 1$;
while $i < n$ **do**
    //linear interpolation correction of $\mu_j$when $a_i \leq j \leq a_{i+1}$
    $slope \leftarrow \frac{v_{i+1} - v_i}{\mu_{a_{i+1}} - \mu_{a_i}}$;
    $b \leftarrow v_i - slope \times \mu_{a_i}$;
    while $j \geq a_i \wedge j \leq a_{i+1}$**do**
        //correction of μ_j
        $\mu'_j \leftarrow slope \times \mu_j + b$;
        $j \leftarrow j + 1$;
    end
    $i \leftarrow i + 1$;
end

Algorithm 8.2 Virtual Lock Mass Correction Algorithm

First, in the for loop, we identify each peak of S corresponding to a lock mass point $v_i \in \mathcal{V}$. Since $S \in \mathcal{S}$ and v_i is a VLM point w.r.t. $(\mathcal{S}, w)$, we are assured to find exactly one such peak $p_j \in S$ with an observed m/z value of μ_j such that μ_j lies in the interval $[(1-w)v_i, (1+w)v_i]$. For such μ_j, we assign the index j to a_i so that $a = (a, \ldots, a_n)$ is a vector of n indexes, each pointing to the peak in S associated to a VLM point. Note that for $\mu_j \in [(1-w)v_i, (1+w)v_i]$, its corrected m/z value must be equal to v_i. Instead of performing these corrections immediately in the for loop, we delay them to the linear interpolation step where all peaks having a m/z value μ_j such that $a_1 \leq j \leq a_n$ will be corrected.

Next, for each VLM v_i, we correct by linear interpolation all the m/z values μ_j such that $a_i \leq j \leq a_{i+1}$. To explain precisely this procedure, let $\mu'(\mu_j)$ denote the corrected value of μ_j. Linear interpolation consists at looking for a correction of the form

$$\mu'(\mu_j) = a\mu_j + b$$

where a is called the slope and b is the intercept. By imposing that $\mu'(\mu_j) = v_i$ for $j = a_i$ and $\mu'(\mu_j) = v_{i+1}$ for $j = a_{i+1}$, we find that

$$a = \frac{v_{i+1} - v_i}{\mu_{a_{i+1}} - \mu_{a_i}}$$

and $b = v_i = a\mu_{a_i}$. The nested while loops of the algorithm performs exactly these linear interpolation corrections for all μ_j such that $\alpha_i \leq j \leq \alpha_{i+1}$ for $i = 1$ to $n-1$.

Once all m/z values μ_j such that $a_1 \leq j \leq a_n$ have been corrected, the algorithm is done. Hence, we have decided not to correct any m/z value of S that is either smaller that $v_1(1-w)$ or larger than $v_n(1+w)$ because such a peak has only one adjacent VLM and, consequently, could only be corrected by extrapolation, which is much less reliable than interpolation. Therefore, we recommend removing all these peaks from S to perform statistical analyses or machine learning experiments. Finally, the intensities of the peaks remain unchanged. The running time complexity of this algorithm is $O(m)$ where m is the number of peaks in the spectrum S (see the full details in Supplementary Information).

8.4.2.7 From VLM correction to spectra alignment

After running the VLM detection and correction algorithms, all the peaks associated with VLM points will be perfectly aligned in the sense that each peak in different spectra associated to a VLM point v will have exactly the same m/z value v. However, all the other peaks corrected by Algorithm 8.2 will not be perfectly aligned in the sense that a molecule fragment responsible for a peak in different spectra will not yield exactly the same mass after correction. This is due to possibly many uncontrollable phenomena that vary each time a sample gets processed by a mass spectrometer, and by the fact that the correction of each peak was performed by an approximate numerical interpolation. However, if all the peaks have been corrected by Algorithm 8.2, we expect that the peaks corresponding to the same molecule fragment f across different spectra will have very similar masses and will all be localized within a very small window of m/z values. Moreover, we also expect that the m/z values of the peaks coming from another molecule fragment g having a different mass will not cross the m/z values coming from molecule fragment f.

More precisely, suppose that we have executed Algorithm 8.1 and Algorithm 8.2 with a window size parameter w (in relative units) on a set $\mathcal{S}$ of mass spectra. In addition, suppose that a molecule fragment f gives rise to a peak of m/z value μ_1 in spectrum S_1, and a peak of m/z value μ_2 in spectrum S_2, and so on for a sub-sequence of spectra in $\mathcal{S}$. Let $\mathcal{M}_f = \{\mu_1, \mu_2, \ldots\}$ be the set of these m/z values. Moreover, let μ_f be the average of the m/z values in $\mathcal{M}_f$. Then, we expect that there exists a window size θ in relative units, such that $0 < \theta < w$, and for which we have $\mu_i \in [\mu_f(1-\theta),\ \mu_f(1+\theta)]$ for all $\mu_i \in \mathcal{M}_f$. Moreover, if θ is sufficiently small, we expect that the sequence $\mathcal{M}_g$ referring to peaks produced by another molecule fragment g having a different mass will be such that each $\mu_j \in \mathcal{M}_g$ will not be located within $[\mu_f\ (1-\theta), \mu_f\ (1+\theta)]$.

Motivated by this hypothesis, let us introduce the following definitions. Given that Algorithm 8.1 and Algorithm 8.2 have been executed on a set $\mathcal{S}$ of mass spectra with window size parameter w in relative units, and given that we have another window size parameter $\theta \ll w$ in relative units, we say that a m/z value μ_f is an alignment point w.r.t. $(\mathcal{S}, \theta)$ if there exists a set $\mathcal{M}_f$ of peaks from $\mathcal{S}$ that satisfies the following properties.

1. Every peak in $\mathcal{M}_f$ comes from a different spectrum of $\mathcal{S}$.
2. The average of the m/z values of the peaks in $\mathcal{M}_f$ is equal to μ_f.

3. Every peak in $\mathcal{M}_f$ has an m/z value in $[\mu_f(1-\theta),\ \mu_f(1+\theta)]$ and all other peaks of $\mathcal{S}$ have an m/z value outside this interval.
4. There does not exist another peak in $\mathcal{S}$ that we can add to $\mathcal{M}_f$ and still satisfy the above properties.

Whenever these criteria are satisfied, we say that $\mathcal{M}_f$ is the alignment set associated to alignment point μ_f. Given $\mathcal{S}$ and θ, an alignment point μ_f w.r.t. $(\mathcal{S},\theta)$ is said to overlap with another alignment point μg w.r.t. $(\mathcal{S},\theta)$ if and only if there exists a non-empty intersection between the m/z intervals $[\mu_f\ (1-\theta),\mu_f\ (1+\theta)]$ and $[\mu_g(1-\theta),\mu_g(1+\theta)]$.

Let $m \stackrel{\text{def}}{=} |\mathcal{S}|$. Note that there are only two differences between the definition of alignment point (and its associated alignment set) and the definition of VLM point (and its associated VLM set). The first difference is the fact that a VLM set must contain exactly m peaks, whereas an alignment set can contain any number of peaks between 1 to m (since the peaks in an alignment set may originate from a molecule fragment which is not present in all the samples for which we have a spectrum in $\mathcal{S}$). Hence, if we remove the constraint that each virtual lock mass must be formed of $|\mathcal{S}|$ peaks from the validation step, Algorithm 8.1 then finds all the maximum-length sub-sequence of peaks that satisfy the four criteria for a valid alignment set when it reaches the overlap deletion step. The second difference is that there is no intensity thresholds t_a and t_b applied to the peaks for alignment, as we wish to align every peak in the spectra if possible. Note that, generally, a lower intensity threshold is still applied to the peaks in order to remove peaks that are the result of background noise. Consequently, with that very minor change,

$$virtualLockMassDetection(\mathcal{S},\theta)$$

finds all isolated alignment points w.r.t. $(\mathcal{S},\theta)$ in $O(n\log m)$ time, where n is the total number of peaks in $\mathcal{S}$.

If the window size parameter θ is too large, then many alignment points will overlap and Algorithm 8.1 will return very few isolated alignment points. If θ is too small, then, in contrast with the VLM identification case, Algorithm 8.1 will return a very large number of isolated alignment points associated to aligned sets that contain only one point. Hence, in contrast

with the VLM identification case, the best parameter θ is not the one for which we obtain the largest number of alignment points.

What should then be the choice for θ? To answer this question, we consider each VLM point (and its associated sequence of peaks) found by Algorithm 8.1. If we leave out one VLM point v_i from the correction Algorithm 8.2 and use this algorithm to correct all the m/z values of the peaks associated to this VLM point, the maximum deviation from v_i among these m/z values will give us the smallest window size θ_i such that each m/z value will be located within $[v_i(1-\theta_i), v_i(1+\theta_i)]$. Essentially, this window size θ_i is the smallest one for which we can still recognize all the peaks associated to the same VLM v_i. It would then certainly be a very good choice for θ in that region of m/z values. We can then repeat this procedure for all isolated VLM points (except the VLMs with the smallest and largest m/z values) found by Algorithm 8.1 to obtain a sequence of θ_i values.

One interesting possibility for θ is the maximum among the θ_i values. However, this is clearly an overestimate of the maximum spreading of peaks associated to the same molecule fragment since all the VLMs will be used for the correction, including the one that was left out. Moreover, as we can see in Figure 8.2, we can recover a large fraction of the non-overlapping VLMs if we use a significantly smaller window size than the maxi θ_i. For that reason, we have decided to use, for the window size θ, the smallest value covering 95% of the non-overlapping VLMs, i.e., the 95th percentile. Alternatively, to attempt to maximize the accuracy of a learning algorithm, a percentile z can be selected by cross-validation along with the selection of the hyperparameters of the learning algorithm.

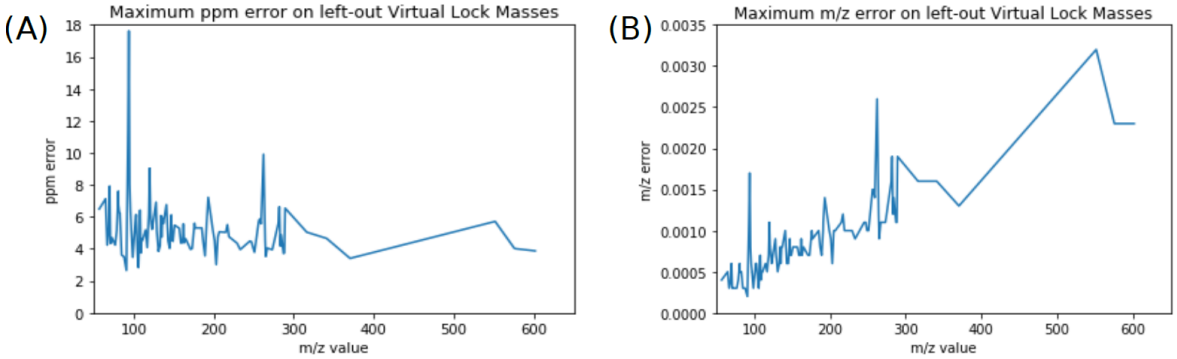


Figure 8.2 Error in ppm versus mass units. Subfigure (A) shows the error on left-out VLMs in ppms, while Subfigure (B) shows the error in Daltons. This data was acquired on the Days Dataset.

If we have r VLM points, each θ_i associated to the ith VLM point is found in $O(m)$ time for a sequence $\mathcal{S}$ of m spectra; thus implying a running time in $O(mr)$ to find every θ_i. Then, the 95th percentile is found by sorting the vector of θ_is in $O(r \log r)$ time. Assuming that we always have $\log(r) < m$, the total running time to find θ is in $O(mr)$, and hence in $O(n)$ when $\mathcal{S}$ contains a total of n peaks. Once the window size θ is found, we can then run Algorithm 8.1 just once on the full set $\mathcal{S}$ of spectra with that value of θ in $O(n \log m)$ time. Consequently, the total running time of the alignment algorithm, which includes the running time to find θ and to find all non overlapping alignment points w.r.t. $(\mathcal{S}, \theta)$, is in $O(n \log m)$.

Once we have the VLM points and the alignment points, these are used to provide a *representation* of the spectra which is well suited for running machine learning algorithms on them. Indeed, consider Figure 8.3. For any new spectrum S, the VLM points are first used correct the m/z value of each peak of S and, following that, the intensity of any corrected peak that fall into the window associated to an alignment point give a feature of S. Hence, the vector of these intensities provides a new representation of the spectrum S that we will use for the input into a classifier to predict the label of S.

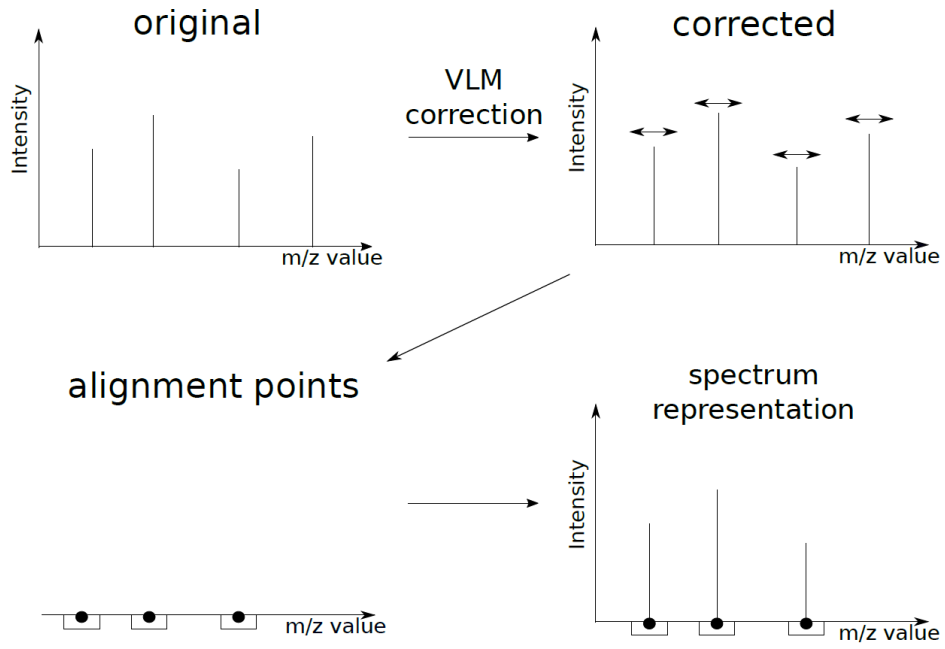


Figure 8.3 Workflow of the VLM and alignment algorithms. First, VLM points are detected in the original spectra in the dataset and VLM correction is applied. The alignment algorithm is then applied to the corrected spectra in order to obtain the alignment points.

Finally, it might be tempting to use a clustering approach to solve the problem of finding the isolated alignment points. However, we have to keep in mind that current trends lead to the processing of hundreds of spectra, each potentially containing thousands of peaks. The total number of peaks to be processed can thus reach a million peaks or more. In our case, the total running time of the full pipeline (finding all isolated VLMs, correcting all the mass spectra with the VLMs, and finding all the isolated alignment points) is in $O(n \log m)$. Hence, any algorithm running in $\Omega(n^2)$ time, will be completely surpassed by the proposed pipeline of algorithms. Currently, the running times of popular off-the-shelf clustering algorithms such as K-means and linkage-based clustering algorithms all require a running time in $\Omega(n^2)$. Moreover, all the clustering algorithms that we know have at least one parameter to tune, which often includes the number of desired clusters. Hence, with the current state of knowledge, a clustering-based algorithm is bound to be substantially less efficient than the proposed pipeline of algorithms.

An implementation of the algorithms for Python is available at https://github.com/francisbrochu/msvlm.

8.4.3 Dataset descriptions

Days Dataset Plasma from 20 healthy individuals was equally pooled together. The pooled plasma was aliquoted and kept at -20ºC. On two consecutive days, an acetonitrile crash was performed using 9 parts of acetonitrile (Fisher Optima) for 1 part of unfrozen plasma pool. The crash solution was centrifuged at 4000 rpm for 5 minutes. 2 µl of the solution was spotted on every well of a 96 wells Lazwell plate (Phytronix). The same experiment was repeated the next day. The dataset is then formed of the 96 aliquots unfrozen on the first day and the 96 aliquots of the second day, forming a 192 samples dataset.

Clomiphene-Acetaminophen Dataset One pill of acetaminophen (500 mg) was diluted in 50 ml of methanol and water (50:50). The solution was put in a sonicating bath for 20 minutes. The resulting solution was centrifuged and diluted 1:100 in water. For clomiphene, we used a solution of 100 µg/ml of clomiphene in methanol (Phytronix).

A pool of plasma was crashed as previously described. The solution was split in 3 parts. One received 10 µl of the acetaminophen solution, another 10 µl of the clomiphene solution and the last one stayed unmodified. Each type of sample was spotted 32 times.

Malaria Dataset *Plasmodium falciparum* parasites were put in culture in red blood cells and tightly synchronized. Culture was performed for 28-36 hours, until parasites are in the trophozoite stage and parasitaemia reached 5-10%. In the same conditions, red blood cells were kept uninfected. Cells were diluted to 2% hematocrit by adding the correct amount of pelleted cells to complete RPMI media. 200 µl of the cell suspensions was deposited in a 96 well plate in order to have 40 samples of infected cells and 40 samples of uninfected cells.

After 4 hours at 37ºC, the plate was spin at 800x g for 5 minutes. 180 µl of culture media was removed. Pellet was resuspended in the remaining 20 µl and 10 µl was transferred to a new 96 well plate. 100 µl of ice-cold methanol was quickly added to the plate and put on dry-ice to stop any metabolic reaction. The plate was vortexed 3 times, for 15 seconds each, over 15 minutes incubation on dry-ice. The plate then was placed for sonication in a water bath for 5 times 1 minute with 2 minutes breaks on dry-ice. Finally, the plate was centrifuged at 3200 rpm for 5 minutes at 4ºC. 30µL of the supernatant was transferred to another plate and kept at -80ºC until LDTD-MS analysis. For analysis, 2 µl of the metabolomic extract was spotted on a 96 well Lazwell plate and left at room temperature until dryness.

Cancer Dataset Plasma from patients diagnosed for breast cancer and from healthy patients were individually treated using the same acetonitrile crash protocol. A total of 96 samples from breast cancer patients were acquired. In addition, 96 plasma samples from healthy patients were also acquired in order to have control samples.

8.4.4 Data acquisition

All data were acquired on a Synapt G2-Si mass spectrometer. The instrument was operated in high resolution mode. Except if stated otherwise, data acquisition was performed in positive ionization. The acquisition method was MS^e with a scan time of 0.1 second. Calibration of the instrument was performed daily before data acquisition using a solution of sodium formate 0.5 mM. The instrument was operated with Mass Lynx software. The source

is a LDTD 960 ion source (Phytronix). The laser pattern used is the following: 2 seconds at 0%, ramp up to 65% in 6 seconds, hold at 65% for 2 seconds and back at 0% in 0.1 second.

8.4.5 Data conversion

Raw files produced by the mass spectrometer were converted to ion list using a continuous to centroid approach using the ProcessKernel software (Waters Corporation) using only the first function (low energy) present in the files. The resulting centroided peak list were used for data analysis.

For all experiments presented in this article, the ta threshold on intensity for virtual lock mass detection was set at 1000 counts. No t_b threshold was applied in the experiments since there was no saturation effect detected on the spectra in the datasets.

Use of human participants All participants provided written informed consent, and the study protocol was reviewed and approved by the Research ethics committee of the CHU de Québec-Université Laval Research Center. All experiments were performed in accordance with relevant guidelines and regulations.

8.5 Results

8.5.1 A consistent set of virtual lock masses can be detected in different batches

This experiment was conducted on the Days dataset (see Methods), which consist of 192 samples of pooled blood plasma. Half of the samples were acquired on a given day and the others were acquired the next day. Since the samples are of the same nature (as in, they are all of the same type of biofluid), we expect a high similarity apart from inter-batch variations. The goal of the experiment was to determine if a consistent set of virtual lock masses could be detected among similar datasets and within parts of the same dataset.

The VLM detection algorithm was independently applied to 1) every spectrum in the dataset, 2) only the spectra acquired on the first day, and 3) only the spectra acquired on the second day. The algorithm was applied with the same window size of 40 ppm in all cases. This

window size was determined by the procedure described in the Methods section, being the w that yielded the largest number of isolated VLMs on the entire dataset.

The detected VLMs were then compared in the following manner. We define that if we have two sets of spectra A and B, their detected VLMs will be $\mathcal{V}_A$ and $\mathcal{V}_B$. Each element of $\mathcal{V}_A$ is a VLM v_A that is composed of a single peak per spectrum for the spectra in A. If $B \subset A$, then a VLM $v_A \in \mathcal{V}_A$ and a VLM $v_B \in \mathcal{V}_B$ are homologous if the peaks forming $\mathcal{V}_B$ are a subset of the peaks forming $\mathcal{V}_A$. Additionally, we can define comparisons between the VLMs of subsets of A. If we have sets of spectra A, B and C, where $B \subset A$ and $C \subset A$, then we can define that VLMs v_B and v_C are homologous if v_B if homologous to v_A and v_C is homologous to v_A.

We compared the peak groups forming the VLMs in all spectra with the spectra acquired on the first day, and found that the 113 VLMs detected on all spectra have homologues in the set of 148 VLMs detected on the first day. Conversely, we observed that the 113 VLMs also have homologues within the set of 118 VLMs detected in the spectra acquired on the second day.

Hence, the algorithm finds common VLM points in all settings, corresponding to different days and multiple instrument recalibrations. This suggests that it correctly identifies landmark compounds that are present in a particular type of sample, which can be used as a common basis for correction. We therefore conclude that our detection algorithm behaves as expected.

8.5.2 Virtual lock mass correction improves machine learning analysis

Machine learning experiments were conducted on four binary classification tasks. The first two tasks consist of the detection of a single compound spiked in blood plasma samples from the Clomiphene-Acetaminophen dataset (see Methods). The third task is the detection of malaria infection in red blood cell culture samples from the Malaria dataset (see Methods). The fourth and final task consists of distinguishing plasma samples of patients with and without breast cancer in the Cancer dataset (see Methods).

8.5.3 Influence of the number of samples on virtual lock mass correction

An experiment was performed in order to evaluate the behavior of the VLM detection and correction algorithms on varying numbers of samples. In a first step, the VLM detection algorithm followed by the VLM correction algorithm was performed on the whole set of spectra. 25 spectra were randomly selected as a test set. These test spectra will be considered the “ground truth", i.e., the best correction that the algorithm can achieve for these 25 spectra.

The algorithm was subsequently applied to a part of the training set. This part was gradually increased from 10 to 160 spectra. At each point, the uncorrected test spectra were corrected and compared to the ground truth. The difference in m/z value between the homologous peaks is calculated in ppm. Then, the difference is squared and summed for all test spectra. Finally, this sum is divided by the number of peaks in the test spectra and the square root is taken. The difference in correction is thus expressed as the Root Mean Squared Error (RMSE) in ppm units for each peak. This experiment was repeated 50 times, with randomly re-partitioned test sets, in order to obtain statistically significant results.

Figure 8.4 shows the learning curves obtained on three different datasets. In each case, the trends is similar. When sub-sampling a low number of spectra as a training set for the VLM detection and correction algorithms, a higher number of lock masses is found. As the number of training spectra increases, the number of virtual lock masses found diminishes and starts to plateau near the number of lock masses found in the whole dataset. This is explained by the fact that when few spectra are in the training set, there is a higher number of candidates. As new spectra are added in the training set, there is a probability that one of the new spectra are missing at least one peak that was previously considered a virtual lock mass. These peaks could be missing because of strong noise, either on the m/z axis or in terms of intensity, rendering its intensity too small to be considered a VLM. A peak could also be missing simply because the compound or fragment generating that peak is not present in all samples.

The same trend is found in all three datasets for the Root Mean Squared Error (RMSE) in Figure 8.4-D. The error is initially high when few spectra are in the training set, but as more spectra are added in the training set it gradually decreases. In the case of the Days Dataset, the final average RMSE when using 160 spectra to train the algorithm is 0.56 ppms. For the

other two datasets (Clomiphene-Acetaminophen and Malaria), the final RMSEs are approximately 1.10 ppms. In each case, the RMSE drops under 2.0 ppms when using 100 spectra or more to train the correction algorithm. In conjunction with the results of inductive learning shown above, these results suggest that the VLM detection and correction algorithms can generalize the virtual lock masses and correction it learns to unseen spectra of the same nature, such as those of a new test set.

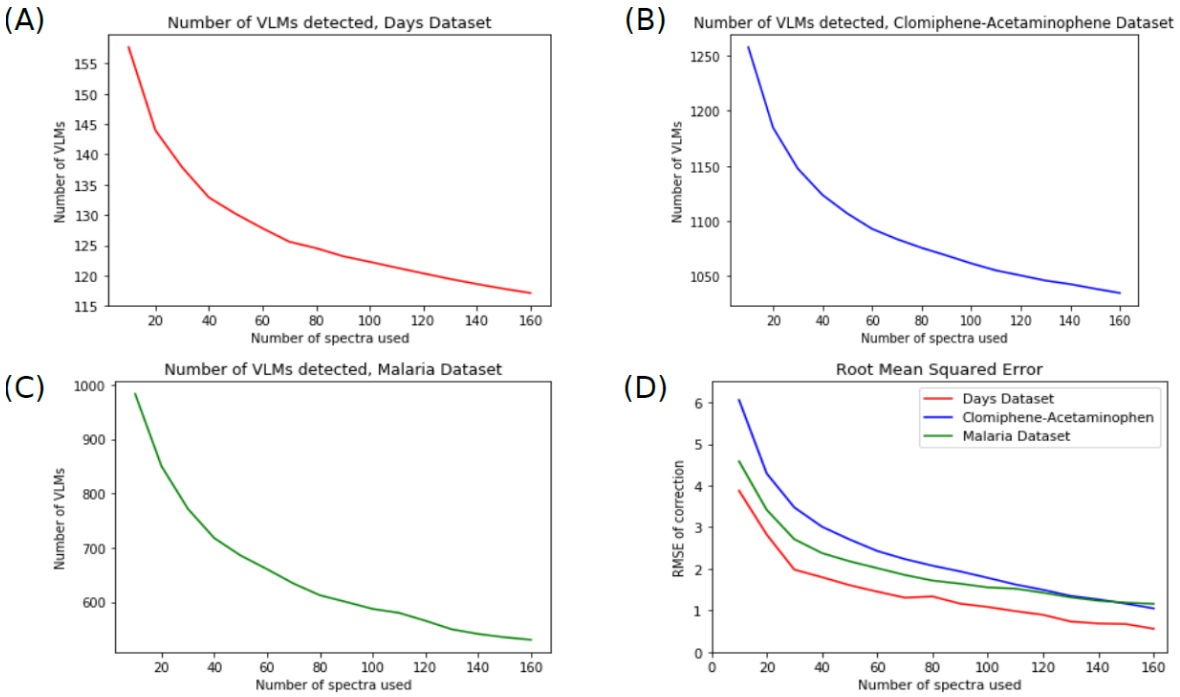


Figure 8.4 Learning Curves of Virtual Lock Mass Detection and Correction. Subfigures (A–C) show the learning curves for three different datasets ((A) Days, (B) Clomiphene-Acetaminophen and (C) Malaria). Subfigure (D) shows the Root Mean Square Error (RMSE) of VLM.

Figure 8.5 shows boxplots of the RMSE of the peaks at different points in the learning curve. Each subfigure shows the RMSE for peaks found in different mass ranges. In every range, the RMSE diminishes with the added spectra to the training set. Moreover, in addition to the average and median values decreasing, we can observe that the interquantile range also decreases. The outliers also tend to be of lesser values. Where multiple outliers have RMSE's greater than 15 when correcting based on 10 spectra, the highest values tend to be less than 10 to 12 (depending on the mass range) when correcting based on 150 spectra.

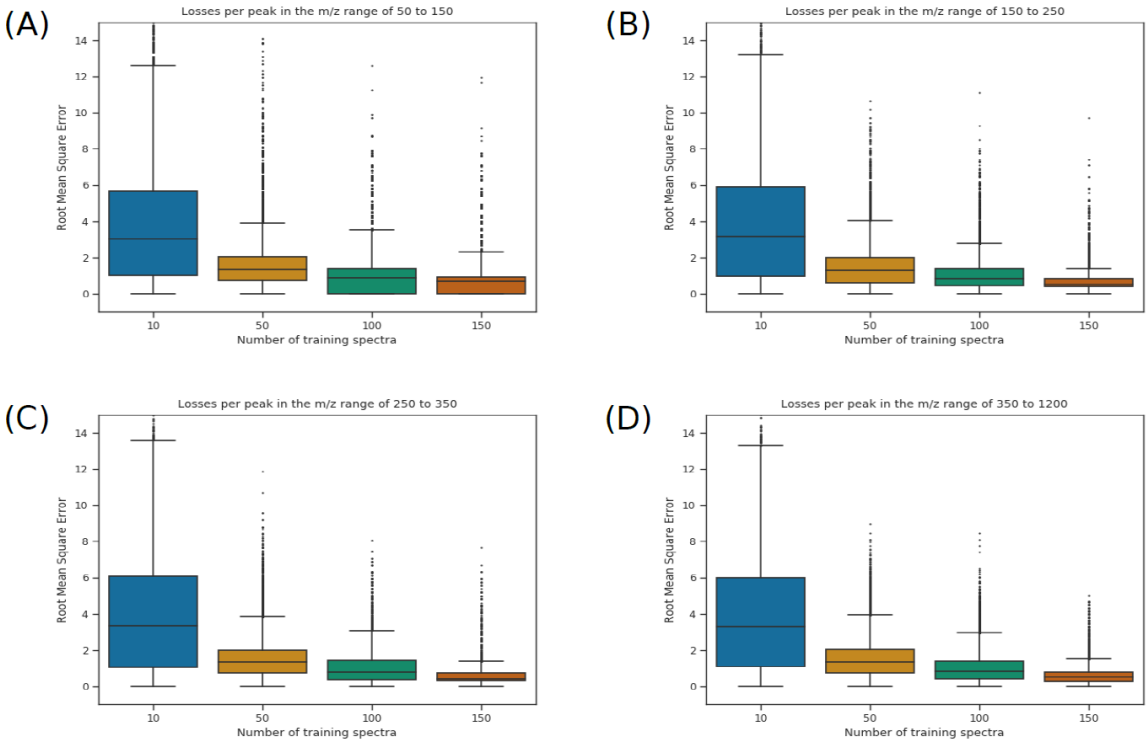


Figure 8.5 Loss per peak in different m/z ranges of the spectra. Each boxplot represents the RMSE of the peaks in a given region (50–150 in (A), 150–250 in (B), 250–350 in (C) and greater than 350 in (D)). Shown here are the results for the Days Dataset, in increasing order to training spectra, from 10 to 150. The outliers are shown as ticks over each box.

Multiple machine learning algorithms were applied to the spectra. The first algorithm used is the AdaBoost ensemble method [147]. This method learns a weighted majority vote on a set of simple pre-defined classifiers. A linear Support Vector Machine (SVM) was used with a L1-norm regularizer [148]. The latter is to ensure that the predictions are based on a small subset of the peaks. We also used decision tree and Set Covering Machine classifiers [149,150]. These algorithms have the advantage of producing interpretable classifiers that consist of a very small combination of simple rules on peak intensities. We used the Scikit-learn implementations for Python of the AdaBoost, CART, and the L1-regularized SVM algorithms [96]; whereas we used our own Python implementation of the SCM. This implementation is available at https://github.com/aldro61/pyscm.

8.5.4 Experimental protocol

For each experiment, the spectra were randomly partitioned into a training set and a test set. For the compound detection tasks (clomiphene and acetaminophen), the test set

consisted of 50 selected samples. For the cancer detection task, the same number of samples were included in the test set. Finally, for the malaria detection task, 100 samples were selected for the test set. The hyper-parameters of each learning algorithm were chosen by 5-fold cross-validation on the training set [60]. Each experiment was repeated 10 times independently on different partitions of the data.

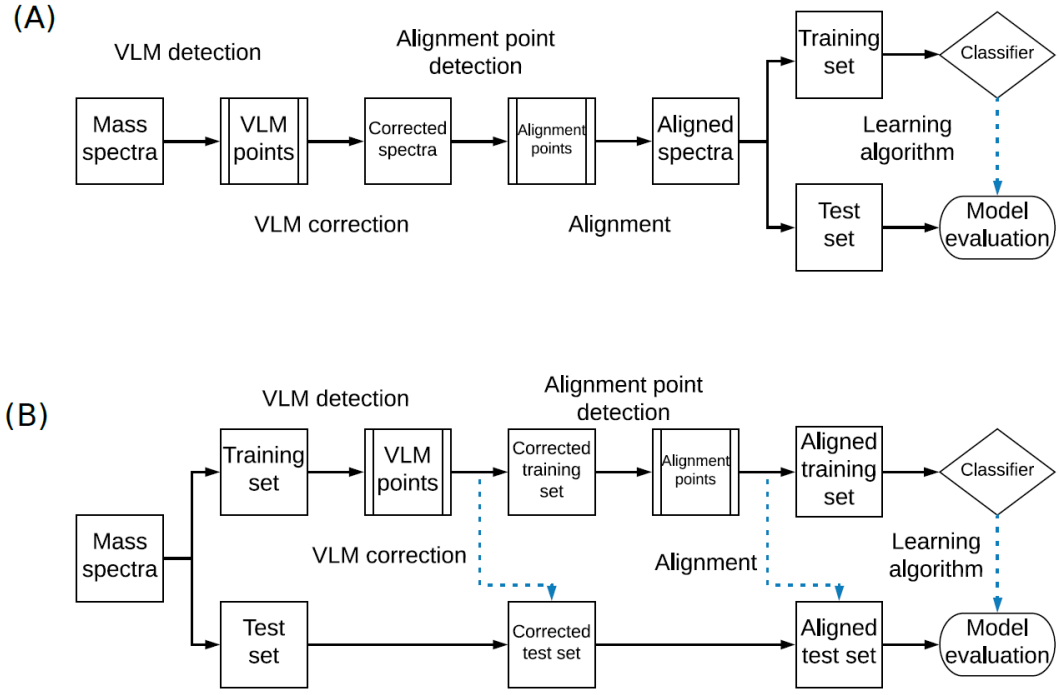


Figure 8.6 Transductive and inductive workflows. ***(A)*** *The transductive workflow, in which all spectra are corrected at once, prior to partitioning the data into a training and testing set.* ***(B)*** *The inductive workflow, where the data are first partitioned and only the spectra in the training set are used to learn a transformation that is applied to all spectra. The dotted blue arrows show where the algorithms were applied on unseen data, while the whole black arrows show the workflow of the training data. Thus, in the inductive workflow, the test set is formed of unseen data that is only used for the final evaluation of the model. In the transductive case, some information is taken from all samples, while only the learning part of the workflow separating a test set on which the algorithm does not learn.*

Two different experimental protocols were tested which are illustrated in Figure 8.6. First, the correction and alignment algorithms were applied in the transductive learning setting [151]. In this setting, the whole dataset is exposed to the pipeline of proposed algorithms (VLM detection + VLM correction + alignment point detection). The training and testing sets are then partitioned randomly. The second experimental protocol was conducted as the inductive learning setting, in which the pipeline of proposed algorithms were only applied to the training set. Hence the set of alignment points is found from the training set only. For the inductive learning protocol, the percentile parameter of the alignment algorithm is

considered an hyper-parameter and is thus cross-validated on the training set. For the transductive learning protocol, the percentile parameter is set at 95%. The features shown to the machine learning algorithms are the alignment points and their associated intensity values.

For each task, we compared the performance of classifiers according to their preprocessing. We thus compared (a) simply binning the spectra, (b) using the VLM detection and correction algorithms and then binning the mass spectra and (c) using the VLM detection and correction algorithms before using the alignment algorithm. Binning is a commonly used technique in mass spectrometry analysis consisting in grouping peaks and intensities found in a larger bin on the m/z axis into a single point or peak [44].

8.5.5 Results for transductive learning

Table 8.1 shows the results of the machine learning experiments in the transductive setting for different tasks. Let us first consider the case of the clomiphene detection task. In all conditions, we observe excellent results, with accuracies over 90% in almost every case. However, we know that the solution to this problem is the appearance of a single additional molecule and its fragments in the spectra, since a solution of water and clomiphene is added in the plasma samples. Thus, it is expected that a single peak (feature) should be sufficient to classify the spectra. Considering this information, we see that a single peak is used for classification only when applying the VLM correction and alignment algorithms when using the Decision Tree and SCM. We also see a decrease in the number of features used for the AdaBoost classifier when using the VLM correction and alignment algorithms. In the case of the L1-regularized SVM, the sparsest solution (with an average of 2.6 features used) was obtained when the VLM correction algorithm was applied in addition to binning.

Consider now the results for acetaminophen detection. In this case, an acetaminophen pill was added to the blood plasma samples. Thus, it is expected that multiple molecules and their fragments appear in the spectra in this case, at extremely high concentration not normally found in physiological blood plasma. It is then not surprising that most algorithms can identify acetaminophen with the use of a singe feature (peak). Note that in the case of the L1-regularized Support Vector Machine, the best results, both in terms of accuracy and sparsity, are obtained when the VLM correction and alignment algorithms were used.

Table 8.1 Machine learning results in the transductive setting.The percentage in each column is the average accuracy of classifiers on 10 repeats of the experiment. The number shown in parentheses is the average number of features used by the classifiers. The algorithms tested were AdaBoost, the Decision Tree algorithm, the Set Covering Machine (SCM) and a L1-norm Support Vector Machine (L1-SVM).

Condition	AdaBoost	Decision Tree	SCM	L1-SVM
Clomiphene Detection				
Binning only	98.0% (4.7)	98.6% (1.8)	95.2% (1.1)	89.6% (52.0)
VLM + Binning	98.2% (4.9)	97.0% (2.3)	97.0% (1.2)	**93.6%** (**2.6**)
VLM + Alignment	**98.8%** (**2.3**)	**99.4%** (**1.0**)	**99.4%** (**1.0**)	92.8% (138.6)
Acetaminophen Detection				
Binning only	99.2% (1.0)	99.2% (**1.0**)	99.2% (1.2)	97.6% (97.5)
VLM + Binning	99.2% (**1.0**)	99.2% (**1.0**)	**99.4%** (**1.0**)	99.0% (121.0)
VLM + Alignment	**99.8%** (**1.0**)	**100.0%** (**1.0**)	**99.4%** (**1.0**)	**99.6%** (**63.4**)
Malaria Detection				
Binning only	92.4% (51.8)	82.5% (**4.3**)	84.6% (2.2)	92.6% (150.1)
VLM + Binning	93.3% (**39.7**)	**88.7%** (4.6)	**89.4%** (**2.0**)	**95.4%** (133.2)
VLM + Alignment	**93.8%** (65.3)	86.1% (4.8)	85.4% (2.3)	95.2% (**131.4**)
Cancer Detection				
Binning Only	**70.4%** (69.2)	**63.8%** (6.4)	55.6% (**1.9**)	56.8% (**113.6**)
VLM + Binning	70.2% (43.9)	61.6% (4.8)	53.6% (2.2)	69.4% (138.6)
VLM + Alignment	67.4% (**30.0**)	62.6% (**2.3**)	**59.6%** (2.2)	**74.6%** (135.2)

The next two tasks represent more realistic problems with unknown solutions. Let us then consider the malaria detection task. For each algorithm, applying the VLM correction algorithm yields an increase in prediction accuracy. For the AdaBoost classifier, we observe an increase of about 1% and the best sparsity in the case of the VLM correction applied before binning, with a slight increase in accuracy with the alignment algorithm. The Decision Tree classifier increases its accuracy by approximately 5% with the VLM correction algorithm, both with alignment and with binning. We see a similar increase in accuracy for the Set Covering Machine in the case of VLM correction with binning. Finally, the L1-regularized SVM obtains a 3% increase in accuracy with the VLM correction algorithm applied, and a better sparsity.

Finally, let us consider the results for the cancer detection task. This classification problem is much harder, with few machine learning algorithms having a prediction accuracy over 70%. Still, both the AdaBoost and Decision Tree classifiers have similar results in all cases, with slight losses in accuracy but improved sparsity with the proposed algorithms applied. The Set Covering Machine sees its accuracy increased by 4% with both correction and alignment algorithms applied and with comparable sparsity. However, in the case of the L1-

regularized SVM, the classifier accuracy increases of almost 20% with the proposed algorithms compared to binning only.

8.5.6 Results for inductive learning

In Table 8.2, we compare the effect of using the proposed algorithms in the transductive setting versus the inductive setting. For the compound detection tasks, there is very little difference between the two approaches for both clomiphene detection and acetaminophen detection. The inductive setting yields slightly sparser classifiers, but the results are very similar. For the malaria detection task, the difference in sparsity is not significant for the Decision Tree and Set Covering Machine algorithms. The AdaBoost classifier is sparser for the inductive setting, while the L1 SVM has a significant advantage in the transductive setting. The results are also very similar in terms of accuracy for both settings, with very slightly better accuracies in the transductive setting. Finally, the transductive setting appears to be the best setting for cancer detection. The AdaBoost classifier is sparser in this case, with a slight decrease in accuracy. The Decision Tree and Set Covering Machine have better accuracies in the transductive setting, though the SCM is sparser in the inductive setting. The L1-regularized SVM is, on the other hand, much more accurate and slightly sparser in the transductive setting, with an increase in accuracy of about 6%.

Table 8.2 Comparison of transductive and inductive learning of the VLM and Alignment algorithms. The algorithms tested were AdaBoost, the Decision Tree algorithm, the Set Covering Machine (SCM) and a L1-norm Support Vector Machine (L1-SVM).

Condition	AdaBoost	Decision Tree	SCM	L1-SVM
Clomiphene Detection				
Transductive	98.8% (2.3)	99.4% (1.0)	**99.4%** (**1.0**)	92.8% (138.6)
Inductive	**99.4%** (**1.0**)	**99.4%** (**1.0**)	**96.4%** (**1.0**)	**93.4%** (**90.0**)
Acetaminophen Detection				
Transductive	99.8% (**1.0**)	**100.0%** (**1.0**)	99.4% (**1.0**)	**99.6%** (63.4)
Inductive	**100.0%** (**1.0**)	99.2% (**1.0**)	**99.6%** (**1.0**)	98.6% (**30.0**)
Malaria Detection				
Transductive	**93.8%** (65.3)	86.1% (4.8)	**85.4%** (2.3)	**95.2%** (**131.4**)
Inductive	92.9% (**54.3**)	**87.8%** (**4.7**)	84.2% (**2.2**)	95.1% (151.0)
Cancer Detection				
Transductive	67.4% (**30.0**)	**62.6%** (**2.3**)	**59.6%** (2.2)	**74.6%** (**135.2**)
Inductive	**69.2%** (63.9)	61.2% (6.7)	57.4% (**1.6**)	68.2% (145.4)

Finally, and perhaps not surprisingly, we can see (for AdaBoost and L1-SVM) that cancer and malaria detection need far more features then clomiphene and acetaminophen detection.

8.5.7 Stability of virtual lock masses in datasets

This experiment was conducted in order to verify that virtual lock masses detected on a given dataset will be found in unseen spectra of the same type. The algorithm for VLM detection was also cross-validated on the *Days* Dataset and the *ClomipheneAcetaminophen* Dataset. Each dataset was randomly partitioned into k folds. The VLM detection algorithm was applied to the first $k-1$ folds, the training folds. The detected VLMs on the training folds are then used for VLM correction of the spectra in last remaining fold, the testing fold. When the correction is applied, we note if every VLM is found in the spectra of the testing fold. The algorithm is scored according to the ratio of detected VLMs on the training folds that are also found in the testing fold. This process is repeated k times so that each fold serves as a test fold once. Multiples values of k were used in the experiment, such that $k \in \{3, 5, 8, 10, 15, 20\}$.

In each case, we found that every VLM point detected on the training set was detected on the testing set. This thus results in a ratio of VLMs found in the testing set over the VLMs detected on the training set of 100% in all cases. This provides empirical evidence of the stability of VLM points across different sets of spectra.

8.6 Discussion

The algorithms proposed in this article aim to render mass spectra more comparable for large datasets acquired in single or multiple batches. The VLM detection algorithm is stable and detects virtual lock masses reliably in datasets. It also detects peaks that are present in mass spectra of the same type but that are not part of the training set. In addition, applying the proposed pipeline of algorithms (VLM detection + VLM correction + alignment point detection) on sets of mass spectra before statistical and machine learning analyses generally yields classifiers with increased accuracy and sometimes with increased sparsity, leading to interpretable models that could serve for biomarker discovery. The proposed pipeline of algorithms has a very low running time complexity of $O(n \log m)$ for a collection of m spectra containing a total of n peaks which, as argued, cannot be surpassed by algorithms based on clustering (with the current state of knowledge). Open-source implementations of the algorithms in Python are also made publicly available.

However, the algorithms, as presented, have a number of drawbacks. Since the virtual lock masses are assigned the average m/z value of the peaks associated to it, the correction algorithm does not correct the peaks to the exact m/z value of the ion. The alignment algorithm has also the same property. However, the virtual lock mass approach is compatible with any external lock masses added to the spectra. Thus, by applying both methods, any shift away from the exact (and known) m/z value of an external lock mass can be corrected. Some situations are also unsuitable for the proposed algorithms. In order for the VLM detection algorithm to function properly and detect virtual lock masses, the mass spectra forming the dataset must be of the same "nature" so that the algorithm can detect a sufficient number of peaks that are common to all spectra. Additionally, the correction algorithm works best in a situation where there are more peaks than spectra. In the cases where each spectrum contains very few peaks, there is a much lower probability that that algorithm can find peaks present in all spectra of the set.

Future works The algorithms, as presented here, can only be applied to mass spectra represented by a list of peaks of the form (μ, ι) where μ is the m/z value of the peak and ι its intensity. Hence, the algorithms are currently not applicable with mass spectra having additional dimensions for the peaks, such as ion mobility. It is also not applicable to mass spectra paired with chromatography. It is thus relevant to investigate if the proposed approach, based on virtual lock masses, can be extended to incorporate these extra dimensions.

8.7 Acknowledgements

We thank Waters Corporation for their support and expertise that helped to the design of the proposed algorithms and for the support using their instruments. We also thank Phytronix Technologies Inc. for their support with their instruments. Computations were made on the supercomputer Colosse from Université Laval, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). Computations were also made on the supercomputer Graham from Waterloo University, managed by Compute Canada. The authors thank the participants for their generosity and providing samples. Plasma samples and the biobanking infrastructure were supported by

grants from the Canadian Breast Cancer Research Alliance and the Fondation du cancer du sein du Québec and the Banque de tissus et données of the Réseau de recherche sur le cancer of the Fond de recherche du Québec – Santé (FRQS), associated with the Canadian Tumor Repository Network (CTRNet). DR lab is funded by a CIHR grant MOP130359. DR is a Fonds de la Recherche du Québec-Santé Junior 2 fellow.

8.8 Author contributions statement

F.B., P-L.P., A.D., M.M. and F.L. conceived the algorithm. F.B., M.M. and F.L. conceived the experiments. F.B., P-L.P., A.D. and M.M. programmed the algorithm. F.B. conducted the experiments and analyzed the results. D.G., D.R., F.D. and C.D. validated the results and reviewed the manuscript. J.C. participated to the initial design and reviewed the manuscript. F.B., P-L.P., A.D., M.M. and F.L. wrote and reviewed the manuscript.

8.9 Additional information

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing Interests Statement

The authors declare no competing interests.

Annexe B- Matériel supplémentaire accompagnant le chapitre 5

Tableau 9.1 Moyenne des scores de la fonction d'efficacité du récepteur (AUC score) pour les classifications sur 30 séparations de Monte-Carlo pour quatre algorithmes. La valeur entre parenthèses correspond à l'écart-type

	DT		RF		SCM		SVM	
EXPÉRIENCE	Entrain.	Test	Entrain.	Test	Entrain.	Test	Entrain.	Test
2017-02-10	0,90 (0,05)	0,77 (0,05)	1,00 (0,00)	0,75 (0,06)	0,95 (0,01)	0,76 (0,07)	1,00 (0,00)	0,77 (0,06)
2017-03-02	0,81 (0,05)	0,74 (0,06)	1,00 (0,00)	0,74 (0,06)	0,92 (0,01)	0,72 (0,06)	1,00 (0,00)	0,76 (0,05)
2017-08-02	0,92 (0,04)	0,77 (0,04)	1,00 (0,00)	0,79 (0,06)	0,88 (0,02)	0,73 (0,05)	1,00 (0,00)	0,76 (0,06)
2019-02-14	0,71 (0,17)	0,52 (0,04)	0,99 (0,03)	0,50 (0,03)	0,56 (0,03)	0,50 (0,02)	0,99 (0,02)	0,54 (0,04)

Tableau 9.2 Liste des ions sélectionnés par au moins deux modèles et trois algorithmes différents pour l'expérience 2017-03-02

Section du diagramme de Venn	m/z et mode d'ionisation
Intersection RF, DT et SVM-L1	178,054 (+)
	174,023 (+)
	172,106 (-)
Intersection RF, DT et SCM	160,172 (+)
	178,089 (+)
	86,099 (+)
	150,097 (+)
	122,101 (+)
	313,174 (-)
Intersection R.F., SCM et SVM-L1	514,340 (-)

Annexe C – Supporting information for Predicting Ion Mobility Collision Cross Sections Using a Deep Neural Network: DeepCCS

10.1 Method supplementary information

The five-fold cross-validation was performed using only the training set, trying a total of 340 different combinations for a total of 1700 different models trained. The complete list of hyper-parameters that were tested and their range are available in **Table 10.1**. The hyperparameter combination giving the best cross-validation score was kept. Finally, the last maximum pooling layer stride parameter was increased to 2 in order to reduce the total number of learnable parameters in the network. During training, the Adam optimizer was used with a learning rate of $1e^{-4}$ for 150 epochs and a batch size of 2. Further details about the neural network structure can be found in **Table 10.2** and on the github repositories available at github.com/plpla/DeepCCS and github.com/plpla/DeepCCS_paper.

10.2 Outliers description

Five outliers are visible in Figure 5 of the original manuscript (Figure 6.5 of this thesis). Here, we describe each outlier with either a confirmed or highly probable explanation.

A. Name: Methyl behenate
   Class: Lipid
   Ion: M-H
   Reference CCS: 156.1 $Å^2$
   DeepCCS: 192.7 $Å^2$
   Explanation: A repeated measurement of methyl behenate using the standardized CCS measurement protocol described by Stow et al. (*Analytical Chemistry* 89(17), 9048-9055, 2017) gives a CCS value of 186.2 $Å^2$ (**Figure 10.1-A**). This confirms the reference value error.

B. Name: 1,2-Diacyl-sn-glycero 3-phosphocholine
   Class: Lipid
   Ion: M+H
   Reference CCS: 189.6 $Å^2$
   DeepCCS: 251.0 $Å^2$

Explanation: A similar lipid (PC 34:2), the double bound unspecified form of 1,2-Diacyl-sn-glycero 3-phosphocholine, was measured with a CCS value of 279.5 $Å^2$ (**Figure 10.1-B**), a value closer to the predicted value. A low abundance ion signal appears at lower CCS and it is suspected that a similar signal artifact is the source of the reference value error.

C. Name: D-Maltose
Class: Sugar

Ion: M-H

Reference CCS: 205.9 $Å^2$

DeepCCS: 169.1 $Å^2$

Explanation: A repeated measurement of D-maltose gives a CCS value of 168.8 $Å^2$ (**Figure 10.1-C**). This updated measurement is very close to the DeepCCS prediction and confirms the reference value error. Carbohydrates are prone to aggregation, and these multimers readily dissociate during transfer from the IM to the MS stage, resulting in multiple ion signals appearing at higher CCS values (c.f., **Figure 10.1-C**). It is suspected that the large CCS reported for the reference value is in fact a multimer signal.

D. Name: Sophorose
Class: Sugar

Ion: M-H

Reference CCS: 187.1 $Å^2$

DeepCCS: 168.7 $Å^2$

Explanation: As was observed with D-maltose (above) it is believed that the large discrepancy between the reference and predicted CCS values for sophorose is due to an aggregate that dissociated into the monomer mass prior to the MS measurement, although no measurements are available to confirm this hypothesis.

E. Name: L-Threonine
Class: Amino acid

Ion: M-H

Reference CCS: 141.4 $Å^2$

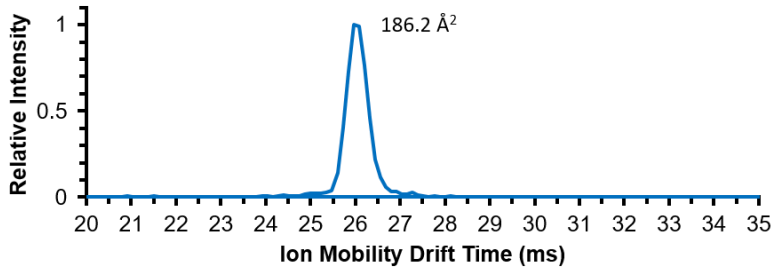
DeepCCS: 125.1 $Å^2$

Explanation: The reference CCS value is higher than other amino acids reported in the reference set. A repeated IM measurement of L-threonine gives a CCS value of 127.6 $Å^2$ (**Figure 10.1-D**), which is much closer to the predicted value. This confirms the reference error. L-threonine exhibits several high CCS artifacts in the IM

spectrum (c.f., **Figure 10.1-D**), and these are likely the source of the erroneously high reference CCS.

(A) Methyl Behenate ($C_{23}H_{46}O_2$)

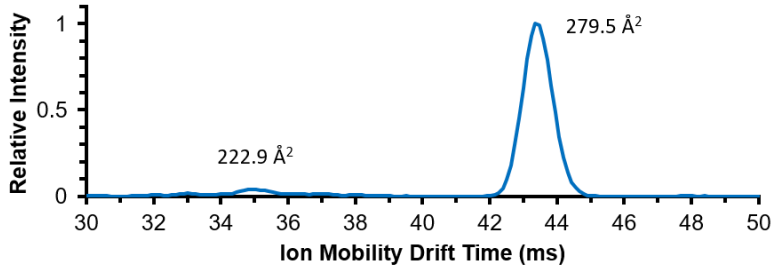
$[M-H]^-$ = 353.3419 Da (measured = 353.3412; 2.1 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	185.9
2	186.3
3	186.3
4	186.4
5	186.2
6	186.1
7	185.9
Average=	**186.2**
Std. Deviation=	**0.2**

(B) Phosphatidylcholine 34:2 ($C_{42}H_{80}NO_8P$)

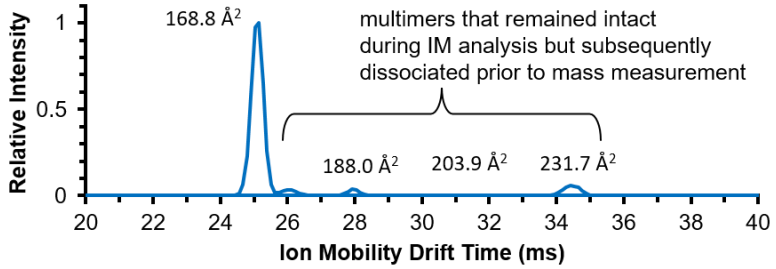
$[M+H]^+$ = 758.5699 Da (measured = 758.5665; 4.5 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	278.8
2	280.6
3	279.4
4	279.0
5	279.9
6	279.2
7	279.4
Average=	**279.5**
Std. Deviation=	**0.6**

(C) D-Maltose ($C_{12}H_{22}O_{11}$)

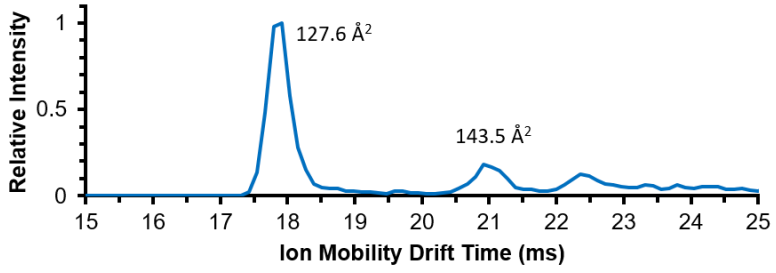
$[M-H]^-$ = 341.1084 Da (measured = 341.1088; 1.2 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	168.6
2	168.9
3	169.1
4	169.0
5	168.9
6	168.7
7	168.6
Average=	**168.8**
Std. Deviation=	**0.2**

(D) L-Threonine ($C_4H_9NO_3$)

$[M+H]^+$ = 120.0661 Da (measured = 120.0657; 3.0 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	127.5
2	127.6
3	127.8
4	127.7
5	127.6
6	127.5
7	127.5
Average=	**127.6**
Std. Deviation=	**0.1**

Figure 10.1 Repeat measurements of the IM spectra and CCS values for (A) methyl behenate, (B) PC 34:2, (C) D-maltose, and (D) L-threonine.

Table 10.1 Possible CNN hyper-parameters values during the random-search cross-validation

Parameter	Values
Max. number of epochs	50, 100 or 150
Batch size	2, 5, 10, 15 or 20
Dropout rate	0.0 to 0.5
Number of convolution layers	Between 1 and 10
Convolution layer width	64, 128, 256 or 384
Convolution filter size	3, 4 or 5
Maximum pooling filter size	2, 3 or 4
Number of dense layers	Between 1 and 10
Dense layers width	64, 128, 256 or 384
Add normalization layer	True or False

Table 10.2 CNN multi-output model performances on the HMDB molecular properties prediction problem

Property	R^2	Median relative error (%)
Polar surface area	0.9998	0.1670
logS	0.9860	-0.6311
Refractivity	0.9999	0.0995
Polarizability	0.9997	0.2005
logP (ALOGPS)	0.9891	0.6419
logP (Chemaxon)	0.9966	0.1574

Table 10.3 DeepCCS neural network structure

Layer (type)	Output Shape	Param #	Connected to
smile (InputLayer)	(None, 250, 37)	0	
conv1d_1 (Conv1D)	(None, 247, 64)	9536	smile[0][0]
conv1d_2 (Conv1D)	(None, 244, 64)	16448	conv1d_1[0][0]
max_pooling1d_1 (MaxPooling1D)	(None, 243, 64)	0	conv1d_2[0][0]
conv1d_3 (Conv1D)	(None, 240, 64)	16448	max_pooling1d_1[0][0]
max_pooling1d_2 (MaxPooling1D)	(None, 239, 64)	0	conv1d_3[0][0]
conv1d_4 (Conv1D)	(None, 236, 64)	16448	max_pooling1d_2[0][0]
max_pooling1d_3 (MaxPooling1D)	(None, 235, 64)	0	conv1d_4[0][0]
conv1d_5 (Conv1D)	(None, 232, 64)	16448	max_pooling1d_3[0][0]
max_pooling1d_4 (MaxPooling1D)	(None, 231, 64)	0	conv1d_5[0][0]
conv1d_6 (Conv1D)	(None, 228, 64)	16448	max_pooling1d_4[0][0]
max_pooling1d_5 (MaxPooling1D)	(None, 227, 64)	0	conv1d_6[0][0]
conv1d_7 (Conv1D)	(None, 224, 64)	16448	max_pooling1d_5[0][0]
max_pooling1d_6 (MaxPooling1D)	(None, 112, 64)	0	conv1d_7[0][0]
flatten_1 (Flatten)	(None, 7168)	0	max_pooling1d_6[0][0]
adduct (InputLayer)	(None, 4)	0	
concatenate_1 (Concatenate)	(None, 7172)	0	flatten_1[0][0], adduct[0][0]
dense_1 (Dense)	(None, 384)	2754432	concatenate_1[0][0]
dense_2 (Dense)	(None, 384)	147840	dense_1[0][0]
dense_3 (Dense)	(None, 1)	385	dense_2[0][0]

Table 10.4 CNN structure for HMDB chemical properties prediction

Layer	(type)	Output Shape	Param #	Connect to
smiles	(InputLayer)	(None, 250, 58)	0	
conv1d_1	(Conv1D)	(None, 247, 64)	14912	smiles[0][0]
conv1d_2	(Conv1D)	(None, 244, 64)	16448	conv1d_1[0][0]
max_pooling1d_1	(MaxPooling1D)	(None,243, 64)	0	conv1d_2[0][0]
conv1d_3	(Conv1D)	(None, 240, 64)	16448	max_pooling1d_1[0][0]
max_pooling1d_2	(MaxPooling1D)	(None, 239, 64)	0	conv1d_3[0][0]
conv1d_4	(Conv1D)	(None, 236, 64)	16448	max_pooling1d_2[0][0]
max_pooling1d_3	(MaxPooling1D)	(None, 235 , 64)	0	conv1d_4[0][0]
conv1d_5	(Conv1D)	(None, 232, 64)	16448	max_pooling1d_3[0][0]
max_pooling1d_4	(MaxPooling1D)	(None, 231 , 64)	0	conv1d_5[0][0]
conv1d_6	(Conv1D)	(None, 228, 64)	16448	max_pooling1d_4[0][0]
max_pooling1d_5	(MaxPooling1D)	(None,227, 64)	0	conv1d_6[0][0]
conv1d_7	(Conv1D)	(None, 223, 64)	16448	max_pooling1d_5[0][0]
max_pooling1d_6	(MaxPooling1D)	(None, 112, 64)	0	conv1d_7[0][0]
flatten_1	(Flatten)	(None, 7168)	0	max_pooling1d_6[0][0]
dense_1	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_3	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_5	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_7	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_9	(Dense)	(None, 384)	2752896	flatten_1[0][0]

dense_11	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_2	(Dense)	(None, 384)	147840	dense_1[0][0]
dense_4	(Dense)	(None, 384)	147840	dense_3[0][0]
dense_6	(Dense)	(None, 384)	147840	dense_5[0][0]
dense_8	(Dense)	(None, 384)	147840	dense_7[0][0]
dense_10	(Dense)	(None, 384)	147840	dense_9[0][0]
dense_12	(Dense)	(None, 384)	147840	dense_11[0][0]
polar_surface_area	(Dense)	(None, 1)	385	dense_2[0][0]
logs	(Dense)	(None, 1)	385	dense_4[0][0]
refractivity	(Dense)	(None, 1)	385	dense_6[0][0]
polarizability	(Dense)	(None, 1)	385	dense_8[0][0]
logp_alogps	(Dense)	(None, 1)	385	dense_10[0][0]
logp_chemaxon	(Dense)	(None, 1)	385	dense_12[0][0]

Table 10.5 Effect of repetitive SMILES-ion combination on the single split experiment

Dataset	Single split		Single split no repetitions	
	R^2	Median relative error (%)	R^2	Median relative error (%)
Global	**0.976 (0.001)**	**2.67 (0.18)**	**0.976 (0.001)**	**2.67 (0.18)**
MetCCS Agilent pos.	0.960 (0.005)	2.02 (0.24)	0.957 (0.006)	2.50 (0.12)
MetCCS Agilent neg.	0.969 (0.005)	3.11 (0.49)	0.978 (0.005)	2.87 (0.53)
Astarita pos.	0.901 (0.013)	4.86 (0.30)	0.901 (0.011)	5.05 (0.33)
Astarita neg.	0.955 (0.006)	3.13 (0.48)	0.961 (0.005)	3.11 (0.44)
Baker	0.954 (0.006)	2.43 (0.11)	0.948 (0.008)	2.64 (0.11)
McLean	0.995 (0.001)	1.49 (0.14)	0.992 (0.001)	1.71 (0.20)
CBM 2018	0.930 (0.010)	2.26 (0.28)	0.930 (0.010)	2.26 (0.28)
Repetitives	-	-	0.960 (0.005)	2.69 (0.34)

Table 10.6 ClassyFire classification at the class level of the datasets used to train and test DeepCCS

Class	**Number**	**Class**	**Number**
Flavonoids	17	Tetracyclines	1
Ergoline and derivatives	1	Organic phosphonic acids and derivatives	1
Pyrimidine nucleotides	18	Nucleoside and nucleotide analogues	1
('Unknown	2	Harmala alkaloids	1
6,7-benzomorphans	1	Benzothiadiazoles	1
Strychnos alkaloids	1	Keto acids and derivatives	7
Fatty Acyls	92	Diarylheptanoids	1
Glycerophospholipids	38	Pyridine nucleotides	2
Isoflavonoids	6	Glycerolipids	3
Organofluorides	3	Organic phosphoric acids and derivatives	5
Diazanaphthalenes	2	Piperidines	7
Pyridines and derivatives	22	Diazines	12
Imidazothiazoles	1	Tetrahydroisoquinolines	1
Triazines	3	Carboxylic acids and derivatives	211
Thioethers	1	Lactones	3

Organic sulfuric acids and derivatives	2	Benzothiopyrans	5
Pyrrolopyrazines	1	Azoles	14
Anthracenes	3	Tetrapyrroles and derivatives	5
Benzofurans	1	Organic sulfonic acids and derivatives	1
Organooxygen compounds	135	Purine nucleosides	14
Piperazinoazepines	2	Carboximidic acids and derivatives	2
Indanes	1	Cycloheptathiophenes	1
5'-deoxyribonucleosides	4	Pyrimidine nucleosides	9
Dibenzocycloheptenes	5	Oxazinanes	1
Ribonucleoside 3'-phosphates	1	Benzene and substituted derivatives	134
Phenols	21	Imidazopyrimidines	23
Indenes and isoindenes	1	Cinnamaldehydes	1
Biotin and derivatives	1	Lactams	1
Benzodioxoles	6	Phenanthrenes and derivatives	5
Tetralins	1	Dithiolanes	1
('Unclassified	54	Organic carbonic acids and derivatives	2
Cinnamic acids and derivatives	6	Peptidomimetics	7
Yohimbine alkaloids	2	Stilbenes	2
Benzazepines	8	Benzoxazoles	1
Purine nucleotides	30	Morphinans	12
Benzoxazepines	2	Indoles and derivatives	36
Polypeptides	7	Naphthalenes	8
Organonitrogen compounds	29	Coumarins and derivatives	3
Pyrroles	3	Thienodiazepines	1
Phenol ethers	10	Quinolines and derivatives	6
Organic oxoanionic compounds	2	Benzothiazepines	3
Isoquinolines and derivatives	1	Benzimidazoles	2
Benzocycloheptapyridines	2	Dihydrofurans	1
Phenylpropanoic acids	5	Flavin nucleotides	2

Imidazole ribonucleosides and ribonucleotides	1	Amaryllidaceae alkaloids	1
Organic dithiophosphoric acids and derivatives	1	Benzodiazepines	32
Phthalide isoquinolines	1	Non-metal oxoanionic compounds	1
Pyridopyrimidines	2	Cinchona alkaloids	2
Macrolactams	1	Steroids and steroid derivatives	54
Benzothiepins	1	Hydroxy acids and derivatives	7
Diazinanes	10	(5'->5'-dinucleotides	6
Linear 1,3-diarylpropanoids	3	Tropane alkaloids	1
Benzoxepines	2	Pteridines and derivatives	10
Sphingolipids	5	Benzopyrans	2
Prenol lipids	22	Benzothiazines	16

Table 10.7 ClassyFire classification at the subclass level of the datasets used to train and test DeepCCS

Subclass	**Number**	**Subclass**	**Number**
Pterins and derivatives	8	Fatty aldehydes	1
Benzylethers	2	1-hydroxy-2-unsubstituted benzenoids	2
Chalcones and dihydrochalcones	1	Pheniramines	1
Unknown	132	Retinoids	3
Delta valerolactones	1	Pyridinecarboxylic acids and derivatives	7
Short-chain keto acids and derivatives	3	Quinoline carboxylic acids	2
Dibenzoxazepines	2	Phosphosphingolipids	1
Anilides	9	Phenylpiperidines	3
Cyclopyrrolones	1	Piperidinecarboxylic acids and derivatives	2
Gamma butyrolactones	2	Fatty acid esters	9
Phenylbutylamines	4	Fentanyls	1
Quinone and hydroquinone lipids	6	Steroid esters	1
Purine nucleotide sugars	4	Purine ribonucleotides	18
Diphenylmethanes	21	Carbazoles	3

Pyridine carboxaldehydes	3	Imidazoles	8
Hydroxyindoles	2	Carbonyl compounds	19
Pregnane steroids	4	Beta hydroxy acids and derivatives	4
Sulfated steroids	2	Indolyl carboxylic acids and derivatives	5
Monoterpenoids	4	Aminoquinolines and derivatives	2
Organic pyrophosphates	2	N-phenylureas	2
Benzodifurans	1	Linear diarylheptanoids	1
Fatty acids and conjugates	53	Unclassified	54
Dicarboxylic acids and derivatives	3	Gamma-keto acids and derivatives	2
Terpene glycosides	2	Pyrazoles	3
Galanthamine-type amaryllidaceae alkaloids	1	Aniline and substituted anilines	2
Amino acids, peptides and analogues	195	1-benzopyrans	2
Hydroxypyridines	1	Beta lactams	1
Androstane steroids	6	Carboxylic acid derivatives	5
O-methylated flavonoids	5	Styrenes	1
Diterpenoids	2	Hydroxysteroids	8
Purines and purine derivatives	23	Phenylpropanes	3
Benzenediols	9	Fatty acyl thioesters	16
Triradylcglycerols	1	Vitamin D and derivatives	1
Purine deoxyribonucleotides	6	Benzylisoquinolines	1
Pyrimidine nucleotide sugars	5	Benzenesulfonamides	17
Indolines	1	Cholestane steroids	3
Trifluoromethylbenzenes	2	1,3, 5-triazines	2
Bipyridines and oligopyridines	1	O-methylated isoflavonoids	3
Benzoic acids and derivatives	31	Flavones	5
Purine 2-deoxyribonucleosides	3	Alloxazines and isoalloxazines	1
Medium-chain keto acids and derivatives	2	Quinolones and derivatives	1
Pyridoxamines	2	Aryl thioethers	1
Phosphate esters	5	Pyrimidine ribonucleotides	6

Halobenzenes	10	Aminophenyl ethers	1
1-ribosyl-imidazolecarboxamides	1	5-deoxy-5-thionucleosides	3
Benzyl alcohols	2	Methoxybenzenes	4
Isoprenoid phosphates	2	1-benzothiopyrans	5
Substituted pyrroles	3	Sulfanilides	1
Benzylpiperidines	1	Pyridoindoles	1
Indolecarboxylic acids and derivatives	1	Dibenzothiazepines	1
Indoloquinolines	1	Flavonoid glycosides	3
Fatty acyl glycosides	1	Phenoxy compounds	2
Piperazines	10	Methoxyphenols	8
Bile acids, alcohols and derivatives	23	Alpha-halocarboxylic acids and derivatives	1
Indoles	6	Pyrimidine 2-deoxyribonucleosides	3
Cinnamic acids	2	Glycerophosphoglycerols	2
Hydroxycinnamic acids and derivatives	4	Tryptamines and derivatives	15
Amines	21	Glycerophosphocholines	19
Hydroxycoumarins	3	Pyrrolidinylpyridines	1
Carboximidic acids	2	Dibenzodiazepines	2
Porphyrins	3	Quaternary ammonium salts	6
Biflavonoids and polyflavonoids	1	Lysergic acids and derivatives	1
Anthraquinones	1	Steroid lactones	1
Xylenes	3	Lineolic acids and derivatives	6
Anisoles	5	Butyrophenones	2
1, 4-benzodiazepines	27	Thiazoles	3
Eicosanoids	6	Cyclohexylphenols	1
Tricarboxylic acids and derivatives	6	Dithiophosphate O-esters	1
Flavans	3	Pyrimidines and pyrimidine derivatives	12
Ureas	2	Short-chain hydroxy acids and derivatives	1
Furanones	1	Phenylacetamides	1

Phenylpyruvic acid derivatives	1	Alpha hydroxy acids and derivatives	2
Glycerophosphoethanolamines	13	Steroidal glycosides	1
Triterpenoids	1	Tetracarboxylic acids and derivatives	1
Carbohydrates and carbohydrate conjugates	107	Naphthoylindoles	1
Tyrosols and derivatives	2	Lipoamides	1
Glycerophosphates	2	Non-metal pyrophosphates	1
Benzenesulfonic acids and derivatives	1	Nicotinamide nucleotides	1
Cyclic pyrimidine nucleotides	1	Benzoylindoles	1
Ceramides	2	Bilirubins	2
Sesquiterpenoids	2	Dibenzothiepins	1
Naphthoquinones	2	Pyrimidine deoxyribonucleotides	6
Nicotinic acid nucleotides	1	Benzodiazines	2
Biphenyls and derivatives	2	Dibenzoxepines	2
Glycerophosphoserines	2	2,6-dimethyl-3-benzazocines	1
Benzylamines	1	Cyclic purine nucleotides	2
Rotenoids	1	Diradylglycerols	1
Hydropyridines	4	Pyridoxines	2
Alcohols and polyols	8	Estrane steroids	4
Guanidines	2	Isoflav-2-enes	2
Dibenzazepines	7	Organosulfonic acids and derivatives	1
Phenothiazines	14	Morpholines	1
Glycosylglycerols	1	Phenethylamines	5
Hybrid peptides	7	Aminotriazines	1
Organic phosphonic acids	1	Arylsulfates	2
Glycosphingolipids	2		