REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DES SCIENCES

DEPARTEMENT DE MICROBIOLOGIE & BIOCHIMIE

N°:

DOMAINE : SCINCES DE LA NATURE ET DE LA VIE

FILIERE : SCIENCE BIOLOGIQUE

OPTION : BIOCHIMIE APPLIQUEE

Mémoire présenté pour l'obtention

Du diplôme de Master Académique

Par : Zidelkheir Nor Elhouda

Salamani Safia Marwa

Intitulé

Développement d'un modèle d'IA pour la prédiction de l'identité de métabolites inconnus à partir de l'analyse de données HPLC-MS

Soutenu devant le jury composé de :

Mr. Seifeddine Drif	Université Mohamed Boudiaf M'sila	Président
Dr. Abdenassar Harrar	Université Mohamed Boudiaf M'sila	Rapporteur
Dr. Mohammed Abdallah KHODJA	Université Mohamed Boudiaf M'sila	Examinateur

Année universitaire : 2024 /2025

« Le plaisir de la science est le plus noble des plaisirs, son désir dépasse toutes les convoitises, sa douceur surpasse toute autre douceur, et sa saveur transcende tous les goûts. Ses adeptes sont les véritables heureux, les authentiques réjouis ; ceux qui lui appartiennent sont comblés de délices. Car, dans la connaissance des choses, il est un plaisir qu'aucun autre ne peut égaler »

Imam al-Shatibi -*Al-Muwafaqat* (1/67)

Dédicace

Louange à Allah, le Tout-Puissant, qui m'a accordé la force, la patience et la persévérance nécessaires pour mener à bien ce travail. Sans Sa guidance et Sa bénédiction, rien n'aurait été possible.

À mon père, dont la vie ne m'a pas donné l'opportunité de croiser le chemin et que je n'ai pas eu la chance de connaître mais dont je suis intimement convaincue qu'il serait fier de chacune de mes réussites en ce monde, puisse Allah l'accueillir dans Sa miséricorde.

À ma mère, dont le cœur immense a été mon premier refuge et ma plus grande inspiration. Ton amour est un fleuve inépuisable qui nourrit mon âme et me pousse toujours plus loin. Chaque épreuve, chaque joie, tu l'as vécue à mes côtés, me tenant la main. Comme je te le dis souvent, tu es mon Alpha Ursae Minoris, cette étoile fixe et fidèle qui éclaire mes pas dans l'obscurité, l'astre sûr et constant qui, dans la voûte céleste de ma vie, m'a toujours indiqué la bonne direction.

À ma très chère grand-mère, gardienne de nos mémoires et conteuse de nos histoires, pour la douce mélodie de ta présence. Tes prières et ta tendresse sont des jardins secrets où je me ressource, des bénédictions qui m'accompagnent à chaque pas, Je vous remercie pour tout Ma**.**

À ma sœur Aya et ma précieuse cousine Amel, les Super Nanas, pour nos rires partagés éclatants comme des étoiles, et votre soutien indéfectible qui fut un phare éclairant les contours de chaque obstacle. Votre amour, votre présence constante et vos encouragements ont été des sources d'inspiration et de motivation, rendant ce chemin plus lumineux. Vous êtes véritablement mon Miracle et mes Espoirs les plus chers.

À mon cher oncle Saber et mes précieuses tantes, Naima, Ascia et Rafi, pour votre bienveillance infinie et vos encouragements constants, véritables piliers sur le chemin de mon parcours, Votre affection sincère a toujours été un havre de paix, une douce assurance qui m'a donné la force d'avancer. Par vos conseils avisés et votre présence chaleureuse, vous avez éclairé ma route. Je suis très fière d'appartenir à cette famille.

À ma précieuse Safia, Safpléia, dont le soutien indéfectible a été une source de force et de motivation inépuisable, un souffle vital dans les défis de cette thèse. Tu es pour moi l'équivalent terrestre des Pléiades, ce groupe d'étoiles sœurs qui brillent de mille feux dans la voûte céleste. Comme elles, ta présence est un éclat constant, un repère fiable et une source d'émerveillement. Guidée par la pureté, choisie par les étoiles, telle est l'essence que tu incarnes dans ma vie. Merci d'avoir été bien plus qu'une amie, d'avoir été ma sœur de cœur, un guide stellaire, un soutien inestimable qui a rendu ce voyage académique possible et plus doux.

À mes chers amis, Aya NIAZ, le TEAM et toute la promotion de biochimie, pour votre soutien indéfectible, votre écoute et les innombrables moments de joie qui ont illuminé ce parcours. Nos discussions passionnées sur l'avenir resteront gravées. Ce travail est aussi le fruit de votre amitié. Merci pour tout

Zidelkheir Norelhouda

Dédicace

Ma reconnaissance éternelle va à Allah, le Miséricordieux, qui a veillé sur moi à chaque étape, m'offrant la patience dans l'attente, la force dans l'épreuve, et la clarté dans l'obscurité. Sans Lui, rien n'aurait été possible. Car ma quête du savoir n'avait d'autre but que de plaire à Allah. Je Lui demande l'acceptation et la sincérité, et qu'il fasse de ce travail une œuvre dédiée à Lui seul, une lumière qui éclaire mes pas sur le chemin de la vie.

À mes parents, mes fondations et mon équilibre. ***À ma mère***, une femme de fer, mon modèle et mon inspiration. ***À mon père***, mon pilier et mon repère. Qu'Allah vous récompense de la meilleure des récompenses. Par Dieu, aucune parole ne saurait vous rendre justice. Même si tous les écrivains de ce monde unissaient leur talent, aucune plume ne pourrait décrire la grandeur de ce que vous êtes, ni la mesure de votre amour. C'est votre foi en moi qui a bâti la mienne.

À mes grands-parents, et tout particulièrement à celle dont les mains racontent plus que mille livres. C'est dans ses silences que j'ai appris la patience et dans ses yeux que j'ai trouvé la paix. Pour sa tendresse, ses encouragements et ses douces prières qui m'ont toujours accompagnée.

À celui qui a cru en moi ***Ali***, même les jours où je me tenais à peine debout. Tu n'as jamais attendu que je sois parfaite, seulement sincère. Merci d'avoir été l'épaule, la voix, et parfois même le souffle. Ma force et ma motivation au quotidien et pour la vie, Inch'Allah. Merci pour ton soutien, ton amour, ta patience, ton écoute dans les moments de doute et ta lumière dans les jours sombres. Ton regard fier me donne chaque jour la force d'aller plus loin.

À ceux qui ont tendu la main sans poser de questions, ceux dont la porte était ouverte bien avant que je frappe. Il y a des gestes qu'on n'oublie jamais ; les vôtres sont gravés. À ma tante ***Fatima*** et mon oncle ***Ahmed***, pour leur affection discrète, leurs mots bienveillants et leur soutien constant. Vous avez su, à votre manière, être une présence rassurante dans ma vie et cette aventure.

À mon frère Lakhedar et mes sœurs ***Jannah, Maryam, Dounia et Malak***, vous êtes mes racines solides, mes reflets, même avec nos défauts et nos petits désordres doux. On a grandi ensemble, parfois avec des hauts et des bas, souvent de manière chaotique, mais toujours unies. Vous êtes, et vous resterez, ma toute première équipe.

À Houda, mon étoile fixe dans ce ciel d'incertitudes. Houda est comme cette étoile massive au cœur de la Bubble Nebula, fille d'énergie et d'étoiles, lumineuse. Ma sœur de cœur, ma Kanka, dans une langue que seules nous comprenons. « Guided by purity, chosen by stars », c'est exactement toi. J'ai traversé des années inoubliables, pleines de rires, de défis et de lumière. Partenaire de nuits blanches et de moments partagés, on a porté ce travail comme on tient une promesse : ensemble, toujours. Et si tout était à refaire, je recommencerais avec toi, sans hésiter, sous le même ciel, à chercher les mêmes étoiles. Comme la Bubble Nebula, immense bulle de gaz illuminée par l'étoile puissante en son centre, tu es cette force invisible mais déterminante qui façonne tout autour d'elle. Ton énergie intense repousse les doutes et les obstacles, créant un espace où notre amitié peut grandir et s'épanouir, rappelant que c'est dans sa puissance que naissent les plus belles promesses et les plus solides liens, ceux qui font notre amitié.

À ma team de copines, Celles avec qui j'ai ri fort, parlé sans filtre, et partagé mille petits moments. Merci d'avoir été mon échappée dans le stress, et cette bulle de légèreté quand tout devenait trop sérieux. Avec vous, chaque pause avait un goût de fête.

Salamani Safia Marwa

Remerciement

Au seuil de ce travail, fruit de longs mois d'efforts et de persévérance, nos pensées et notre gratitude se tournent humblement vers Dieu le Tout-Puissant, qui nous a accordé la santé, la force et la volonté nécessaires pour mener cette aventure intellectuelle à son terme.

Ce mémoire, aboutissement de longs mois de recherche et de persévérance, a bénéficié du soutien inestimable de plusieurs contributeurs.

Nous exprimons notre profonde gratitude à notre directeur de mémoire, **Dr. Abdenassar Harrar**, dont la direction éclairée, la rigueur scientifique et la disponibilité constante ont constitué une boussole essentielle face aux complexités de cette recherche. Sa guidance, alliant expertise académique et soutien humain, fut déterminante.

Nos remerciements chaleureux s'adressent également aux membres du jury. Nous sommes particulièrement reconnaissantes envers le président, **Dr. Seifeddine Drif**, pour l'honneur de sa présidence et pour les échanges enrichissants qui ont souligné la dimension humaine de la science. Nous remercions également l'examinateur, **Dr. Mohammed Abdallah KHODJA**, dont l'évaluation attentive et les critiques constructives ont significativement contribué à parfaire ce travail.

Une mention spéciale est dédiée à **Mme Bouazziz**, "professeure de cœur", dont la bienveillance, le soutien moral et les encouragements constants ont été une source inestimable de réconfort et de motivation tout au long de notre parcours.

Une mention toute particulière est dédiée à la **Dr. Salamani Dalila**. Dans chaque voyage, une lumière veille, parfois discrète mais toujours essentielle, et pour nous, cette lumière a porté ton nom. Ton expertise et ton aide précieuse ont été un pilier fondamental, transformant les obstacles en défis surmontables et agissant comme une seconde boussole dans cette recherche. Nous t'en sommes profondément reconnaissantes.

Enfin, nous remercions l'ensemble du personnel administratif et enseignant du département pour leur contribution, directe ou indirecte, à la réussite de notre parcours universitaire.

Sommaire

Résumé i

Liste des abréviations ii

Liste des figures iii

Liste des tableaux iv

Introduction 1

Chapitre I. Chromatographie-MS et métabolites 2

I.1. Métabolites 2

I.2. Chromatographie 3

I.3. Spectrométrie de Masse 6

I.4. Couplage chromatographie-MS 13

Chapitre II. Concepts de l'intelligence artificielle 16

II.1. Introduction à l'intelligence artificielle 16

II.2. Principales approches de l'intelligence artificiels 17

II.3. Machine learning vs Deep learning 23

II.4. Synergie entre le Machine learning et le Deep learning 26

Chapitre III. Matériels et méthodes 29

III.1. Matériels 29

III.2. Méthodes 32

Chapitre IV. Résultats et discussion 38

IV.1. Résultats 38

IV.2. Discussion 51

IV.3. Limitations de l'étude 52

Conclusion 53

Perspective 53

Références bibliographiques 54

ملخص

تُعد عملية التعرف الدقيق على نواتج مطيافية الكتلة (MS) من التحديات الكبرى في علم الميتابولوميات. نقترح في هذه الدراسة نموذجا للتعلم العميق متعدّد المهام قادرًا على التنبؤ بأسماء المركّبات وصيغها الجزيئية في آن واحد اعتمادًا على أطياف MS. انطلاقًا من 122,512 طيفًا تمثل 18,332 كيانًا جزيئيًا فريدًا، تم استخلاص مجموعة بيانات مكوّنة من 15,930 طيفًا عالي الجودة (تعود إلى 255 مركبًا مميزًا) من مكتبة MassBank-NIST 2024.11، وقد جرى معالجتها مسبقًا باستخدام لغة Python 3.12.4 مع مكتبات *pandas* و*numpy* و*scikit-learn* من أجل تحويل البيانات وتطبيعها. تم تطوير شبكة عصبية التفافية أحادية البعد (D-CNN1) باستخدام *TensorFlow/Keras*، وتم تدريبها والتحقق من صحتها داخل بيئة Visual Studio 2022. حقق النموذج دقة بنسبة 93.31 % في التنبؤ بأسماء المركبات و94.76 % في التنبؤ بصيغها الجزيئية تُظهر هذه النتائج أن نموذج التعلم العميق من نوع CNN متعدد المهام يُسهم بفعالية في تعزيز تفسير الأطياف وتسريع عملية تحديد المركّبات في علم الميتابولوميات.

الكلمات المفتاحية: التعرف على المستقلبات، الذكاء الاصطناعي، كروماتوغرافيا السائلة عالية الأداء، التحليل الطيفي الكتلي، شبكة عصبية التفافية

Abstract

Accurate annotation of mass spectrometry (MS) spectra remains a major challenge in metabolomics. We propose a multitask deep learning framework that simultaneously predicts compound names and molecular formulas from MS spectra. From an initial 122,512 spectral entries representing 18,332 unique molecular entities, a curated dataset of 15,930 high-quality spectra (255 unique compounds) was extracted from the MassBank-NIST 2024.11 library and preprocessed using Python 3.12.4 with *pandas*, *numpy*, and *scikit-learn* for data transformation and normalization. A one-dimensional convolutional neural network (1D-CNN) was implemented in *TensorFlow/Keras*, trained and validated in Visual Studio 2022. The model achieved 93.31 % and 94.76 % accuracy for name and formula prediction, respectively. These results demonstrate the potential of deep multitask architectures to enhance spectral annotation and accelerate compound identification in metabolomics.

Keywords: Metabolite Identification, AI, HPLC, MS, CNN

Résumé

L’annotation précise des spectres de spectrométrie de masse (MS) demeure un défi majeur en métabolomique. Nous proposons un cadre d’apprentissage profond multitâche capable de prédire simultanément les noms des composés et leurs formules moléculaires à partir des spectres MS. À partir de 122 512 entrées spectrales représentant 18 332 entités moléculaires uniques, un jeu de données de 15 930 spectres de haute qualité (correspondant à 255 composés uniques) a été extrait de la bibliothèque MassBank-NIST 2024.11, puis prétraité à l’aide de Python 3.12.4 avec les bibliothèques pandas, *numpy* et *scikit-learn* pour la transformation et la normalisation des données. Un réseau de neurones convolutif unidimensionnel (1D-CNN) a été implémenté sous *TensorFlow/Keras*, entraîné et validé dans l’environnement Visual Studio 2022. Le modèle a atteint une précision de 93.31 % pour la prédiction des noms et de 94.76 % pour celle des formules. Ces résultats démontrent le potentiel des architectures profondes multitâches à améliorer l’annotation spectrale et à accélérer l’identification des composés en métabolomique.

Mots-clés : Identification des Métabolites, IA, HPLC, MS, CNN

Liste des abréviations

AI : Artificial Intelligence

ANNs : Artificial Neural Networks (Réseaux de neurones artificiels)

CNN : Convolutional Neural Network (Réseau de neurones convolutifs)

DL : Deep Learning (Apprentissage profond)

DQN : Deep Q-Networks

GNN : Graph Neural Networks

GNPS : Global Natural Products Social Molecular Networking

GRU : Gated Recurrent Units

HPLC-MS : Chromatographie liquide haute performance couplée à la spectrométrie de masse

IA : Intelligence artificielle

LSTM : Long Short-Term Memory

METLIN : Base de données métabolomique

ML : Machine Learning (Apprentissage automatique)

MS : Spectrométrie de masse

NIST : National Institute of Standards and Technology

PLS-DA : Partial Least Squares Discriminant Analysis

QSAR : Quantitative Structure-Activity Relationship

RL : Reinforcement Learning (Apprentissage par renforcement)

RNA : Réseau de neurones artificiels

RNN : Recurrent Neural Network (Réseau de neurones récurrents)

SVM : Support Vector Machine

UHPLC : Ultra-High Performance Liquid Chromatography

UHPLC-MS : Ultra-High Performance Liquid Chromatography - Mass Spectrometry

XCMS : Plateforme pour l'analyse de données métabolomiques

Liste des figures

Figure I.1. Principe et Configuration d'un Système d'HPLC.....5

Figure I.2. Représentation schématique du principe de la technique de MS. 6

Figure I.3. Composants d'un spectromètre de masse 11

Figure I.4. Schéma de l'analyse en spectrométrie de masse en mode SRM. 12

Figure II.1. Les concepts de l'intelligence artificielle 17

Figure II.2. Taxonomies de l'approche en apprentissage automatique 19

Figure II.3. Neurone biologique et perceptron.....20

Figure II.4. Un modèle de perceptron multicouche 21

Figure II.5. Représentation schématique des réseaux de neurones convolutifs (CNNs). 22

Figure II.6. L'architecture du modèle Transformer 23

Figure II.7. Interconnexion entre le ML et le DL.....27

Figure IV.1. Accuracy pour la tâche de prédiction du nom au cours des époques. 46

Figure IV.2. Accuracy pour la tâche de prédiction de la formule chimique. 47

Figure IV.3. Évolution de la fonction de perte globale au cours des époques. 48

Figure IV.4. Évolution de la fonction de perte pour la prédiction du nom 49

Figure IV.5. Évolution de la fonction de perte pour la prédiction de la formule chimique 50

Liste des tableaux

Tableau I.1. Caractéristiques des différents analyseurs de masse. 11

Tableau II.1. Comparaison des architectures de ML et de DL) 24

Tableau II.2. Comparaison des techniques d'apprentissage de ML et de DL 26

Tableau III.1. Paramètres de classement descriptifs du dataset. 30

Tableau IV.1. Aperçu des données métabolomiques converties de MSP en CSV. 40

Tableau IV.2. Vue d'ensemble des données métabolomiques nettoyées. 41

Tableau IV.3. Aperçu des données métabolomiques transformées et encodées. 42

Introduction

Introduction

L'identification précise des métabolites inconnus demeure l'un des défis majeurs de la métabolomique, en particulier lors de l'analyse de données issues de la chromatographie liquide à haute performance couplée à la spectrométrie de masse (HPLC-MS) (Dührkop *et al.*, 2021). Cette technique, devenue incontournable pour l'analyse à haut débit des mélanges complexes, permet de générer de vastes ensembles de données contenant des milliers de signaux, dont une part significative correspond à des composés inconnus ou non référencés dans les bases de données existantes (Da Silva *et al.*, 2015). Malgré les avancées réalisées dans le développement de bibliothèques de spectres et d'outils d'annotation automatique, l'identification structurelle des métabolites repose encore largement sur la comparaison avec des standards de référence, une approche limitée par la couverture incomplète des bases de données et la disponibilité restreinte de standards commerciaux (Blaženović *et al.*, 2018).

La complexité des spectres MS/MS, la variabilité des conditions analytiques et la diversité structurelle des métabolites naturels compliquent davantage l'annotation fiable des signaux détectés (Nguyen *et al.*, 2019). Les méthodes conventionnelles, telles que la recherche par similarité de masse ou la prédiction de fragmentation, présentent des taux de faux positifs non négligeables et peinent à identifier les composés nouveaux ou faiblement représentés (Dührkop *et al.*, 2019). Par conséquent, une proportion importante de signaux détectés lors des analyses HPLC-MS demeure non identifiée, limitant la portée des études métabolomiques et la découverte de biomarqueurs ou de nouvelles entités chimiques (Da Silva *et al.*, 2015).

Face à ces limitations, l'intelligence artificielle (IA), et plus particulièrement l'apprentissage profond, s'impose comme une approche prometteuse pour l'interprétation et la prédiction de l'identité des métabolites à partir de données massives et hétérogènes (Ruttkies *et al.*, 2016). Les modèles d'IA, capables d'apprendre des relations complexes entre les caractéristiques spectrales et les structures moléculaires, ont récemment démontré leur efficacité pour l'annotation automatisée de métabolites inconnus, surpassant parfois les méthodes traditionnelles en termes de précision et de couverture (Dührkop *et al.*, 2021).

L'objectif principal de ce mémoire est de développer un modèle d'intelligence artificielle à sorties multiples (multi-output), capable de prédire simultanément le nom et la formule chimique d'un composé à partir de son spectre de masse. Ce modèle vise à offrir une précision élevée tout en réduisant la dépendance à l'expertise humaine pour l'interprétation des résultats analytiques.

Partie bibliographique
Chapitre I : Chromatographie-MS et métabolites

Chapitre I. Chromatographie-MS et métabolites

I.1. Métabolites

I.1.1. Définition

Les métabolites sont des molécules de faible poids moléculaire, généralement issues de la biosynthèse cellulaire, jouant des rôles essentiels dans les processus physiologiques. Ils peuvent agir comme intermédiaires ou produits finaux dans les voies métaboliques (Wellen et Thompson, 2012).

I.1.2. Classification

I.1.2.1. Métabolites primaires

Les métabolites primaires sont des composés essentiels au bon fonctionnement des cellules végétales. Ils participent directement à divers processus biochimiques et physiologiques, tels que la photosynthèse et la respiration, tout en fournissant l'énergie et les précurseurs nécessaires à la biosynthèse de nouvelles macromolécules indispensables au développement des plantes (Patel *et al.*, 2020).

I.1.2.2. Métabolites secondaires

Les métabolites secondaires sont des composés organiques qui ne participent pas directement à la croissance, au développement ou à la reproduction normale d'un organisme. Ils jouent souvent un rôle crucial dans les mécanismes de défense des plantes, les interactions avec d'autres organismes et les fonctions écologiques (Crozier *et al.*, 2006).

I.1.3. Fonctions et importances

Les métabolites produits par les plantes jouent un rôle fondamental dans les interactions écologiques. Ils peuvent influencer les relations avec d'autres organismes, tels que les herbivores et les pollinisateurs, et résultent souvent d'adaptations évolutives à des pressions environnementales spécifiques (Fernie et Pichersky, 2015).

Les métabolites remplissent de multiples fonctions, telles que la conversion de l'énergie, l'activité de signalisation ainsi que le rôle de cofacteurs. D'autres fonctions des métabolites ont également été observées, notamment leur influence épigénétique dans divers bioprocédés. Actuellement, les métabolites trouvent des applications dans le domaine clinique, où ils présentent un intérêt croissant pour les soins de santé. Ils peuvent, par exemple, être utilisés comme biomarqueurs pour la mesure du glucose dans le liquide cérébrospinal (LCS), en particulier dans les cas où l'on suspecte une infection, une inflammation ou un processus malin.

L'étude des métabolites permet l'intégration de différents types de données biologiques, enrichissant ainsi notre compréhension du métabolisme végétal. Cette intégration offre des perspectives sur les interactions entre les différents niveaux moléculaires et leur contribution au profil métabolique global de la plante (Wellen et Thompson, 2012) .

I.2. Chromatographie

I.2.1. Principe

La chromatographie est une technique analytique essentielle, fondée sur la séparation des composants d'un mélange en fonction de leurs différences d'interaction avec deux phases : une phase stationnaire et une phase mobile. Lorsqu'un mélange est appliqué sur la phase stationnaire, il est transporté par la phase mobile, qui peut être liquide ou gazeuse. Les différents constituants du mélange migrent à des vitesses distinctes en raison de leurs affinités respectives avec la phase stationnaire et la phase mobile, ce qui entraîne leur séparation progressive.

Le processus repose sur la partition différentielle des analytes entre ces deux phases : les molécules ayant une affinité plus forte pour la phase stationnaire restent plus longtemps dans le système et migrent donc plus lentement, tandis que celles qui préfèrent la phase mobile traversent plus rapidement et sont éluées en premier.

Les principaux facteurs influençant la séparation incluent la solubilité, l'adsorption, l'échange d'ions et l'affinité spécifique des molécules pour la phase stationnaire. La chromatographie peut être utilisée à des fins analytiques, pour identifier et quantifier les composants d'un mélange, ou à des fins préparatives, pour isoler et purifier des substances d'intérêt (Patel, 2018; Vij et Pathania, 2023).

I.2.2. Types

I.2.2.1. CCM

La chromatographie sur couche mince (CCM) est une méthode analytique employée pour séparer et identifier les constituants d'un mélange. Elle repose sur une phase stationnaire, généralement constituée d'une fine couche de matériau adsorbant déposée sur un support plat. L'échantillon à analyser est appliqué sur cette couche, puis une phase mobile, le plus souvent un solvant, migre à travers la phase stationnaire par capillarité. Ce déplacement permet aux différents composés du mélange de progresser à des vitesses différentes, entraînant ainsi leur séparation en fonction de leurs propriétés chimiques respectives. La CCM est largement utilisée pour l'analyse de diverses substances, tant dans les domaines environnemental et industriel que dans l'étude des matériaux végétaux et des extraits de plantes médicinales (Hameed *et al.*, 2023).

I.2.2.2. CG

La chromatographie en phase gazeuse (GC) est une technique analytique utilisée pour séparer et analyser des composés pouvant être vaporisés sans décomposition (Masucci et Caldwell, 2004). Le principe repose sur l'injection d'un échantillon dans un flux de gaz porteur (souvent de l'hélium ou de l'azote) qui transporte les molécules à travers une colonne contenant une phase stationnaire. Les composés se séparent en fonction de leur affinité avec la phase stationnaire : plus leur affinité est élevée, plus leur temps de rétention sera long. Le processus débute par la vaporisation de l'échantillon à haute température, suivi de son passage dans la colonne où il subit des interactions multiples entre la phase mobile et la phase stationnaire. La séparation est influencée par plusieurs facteurs tels que la nature de la phase stationnaire, la température du four, le débit du gaz porteur, la longueur et le diamètre de la colonne, ainsi que la sélection du détecteur, par exemple le détecteur à ionisation de flamme FID ou le détecteur de conductivité thermique TCD (Holley *et al.*, 1995).

I.2.2.3. LC

La chromatographie liquide est une technique de séparation où un liquide (phase mobile) transporte les composants d'un échantillon à travers un système contenant une phase stationnaire fixée dans une colonne, un capillaire ou sur une surface plane. Lors de l'introduction de l'échantillon, les molécules se répartissent entre les deux phases selon leurs affinités chimiques, et l'ajout continu de phase mobile provoque une série de transferts dynamiques entre les phases, permettant leur séparation. Cette séparation repose sur des interactions physico-chimiques telles que l'adsorption/partition, l'échange ionique, l'exclusion de taille ou une interaction d'affinité si la phase stationnaire est fonctionnalisée avec des groupes spécifiques (par exemple, des sites chiraux). L'élution peut être réalisée en mode isocratique (composition constante de la phase mobile) ou en gradient (composition modulée dans le temps). Polyvalente et adaptable, cette méthode est largement utilisée pour analyser des mélanges complexes, notamment dans des domaines comme la science du patrimoine, où elle permet d'étudier des matériaux fragiles sans altération majeure (Degano, 2019).

I.2.2.4. HPLC

La chromatographie liquide haute performance (HPLC) (Fig. I.1) est une technique analytique de séparation reposant sur l'injection d'un faible volume d'échantillon liquide dans une colonne remplie de particules de très petite taille (3 à 5 microns de diamètre), constituant la phase stationnaire. Sous l'effet d'une pression élevée exercée par une pompe, une phase mobile liquide traverse la colonne, entraînant les différents composants de l'échantillon. La séparation des

analytes repose sur des interactions chimiques et/ou physiques différentielles entre les molécules de l'échantillon et la matrice stationnaire de la colonne.

L'HPLC se distingue par sa capacité à analyser un large éventail de composés, y compris des substances non volatiles, ce qui en fait une méthode particulièrement adaptée à l'étude des macromolécules. Elle est couramment employée dans divers domaines scientifiques et industriels, notamment la pharmaceutique, la biotechnologie, les sciences environnementales et l'industrie agroalimentaire, aussi bien pour la séparation que pour la purification des analytes (Ali, 2022).

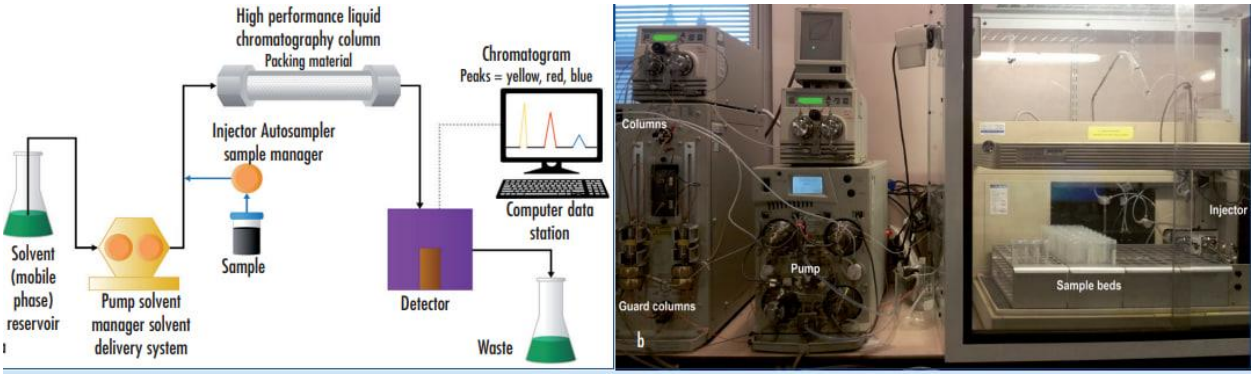

Figure I.1. Principe et Configuration d'un Système de Chromatographie Liquide Haute Performance (Blum, 2014).

I.2.2.5. SCF

La chromatographie en fluide supercritique (SFC) est une technique de séparation utilisant un fluide supercritique (phase mobile) comme le CO_2 sous pression et température élevées, combiné à une phase stationnaire fixée dans une colonne chromatographique. Lors de l'introduction de l'échantillon, les composants interagissent avec la phase stationnaire et le fluide supercritique, dont les propriétés hybrides (densité proche d'un liquide, viscosité semblable à un gaz) favorisent une séparation rapide et efficace. La séparation repose sur des interactions physico-chimiques telles que l'adsorption, la solubilité différentielle dans le fluide supercritique, ou des mécanismes d'affinité (ex. chiralité). L'élution peut être réalisée en mode isocratique (conditions constantes) ou en gradient (modulation de la pression, de la température ou de l'ajout de modificateurs comme le méthanol). Cette méthode combine les avantages de la chromatographie gazeuse (vitesse) et liquide (polyvalence), avec des applications en analyse de composés thermosensibles, séparations chirales ou purification de principes actifs, tout en étant éco-compatible grâce à l'utilisation réduite de solvants organiques (Taylor, 2009).

I.3. Spectrométrie de Masse

I.3.1. Principe

La spectrométrie de masse (Fig. I.2) (Fig. I.3) constitue une technique analytique de haute précision permettant de mesurer le rapport masse/charge (m/z) des ions. Le processus débute par l'ionisation des molécules présentes dans l'échantillon, laquelle peut être réalisée à l'aide de diverses méthodes, telles que l'ionisation par bombardement électronique, l'ionisation chimique ou l'ionisation laser. Une fois ionisées, les molécules sont converties en ions chargés, qui sont ensuite soumis à un processus d'accélération et de séparation en fonction de leur rapport m/z à travers l'application de champs électriques ou magnétiques.

Après cette séparation, les ions sont détectés, et les données recueillies sont analysées afin de fournir des informations détaillées sur la composition chimique et les propriétés de l'échantillon. La précision et la fiabilité des résultats en spectrométrie de masse dépendent en grande partie des étapes préalables de traitement de l'échantillon, telles que l'extraction, la purification et l'enrichissement des métabolites cibles, visant à éliminer les substances interférentes. Cette approche rigoureuse permet l'identification et la quantification de différents métabolites au sein d'échantillons biologiques complexes, positionnant la spectrométrie de masse comme un outil incontournable dans des domaines variés, tels que la métabolomique, le développement pharmaceutique et le diagnostic des maladies (Zhou, 2024).

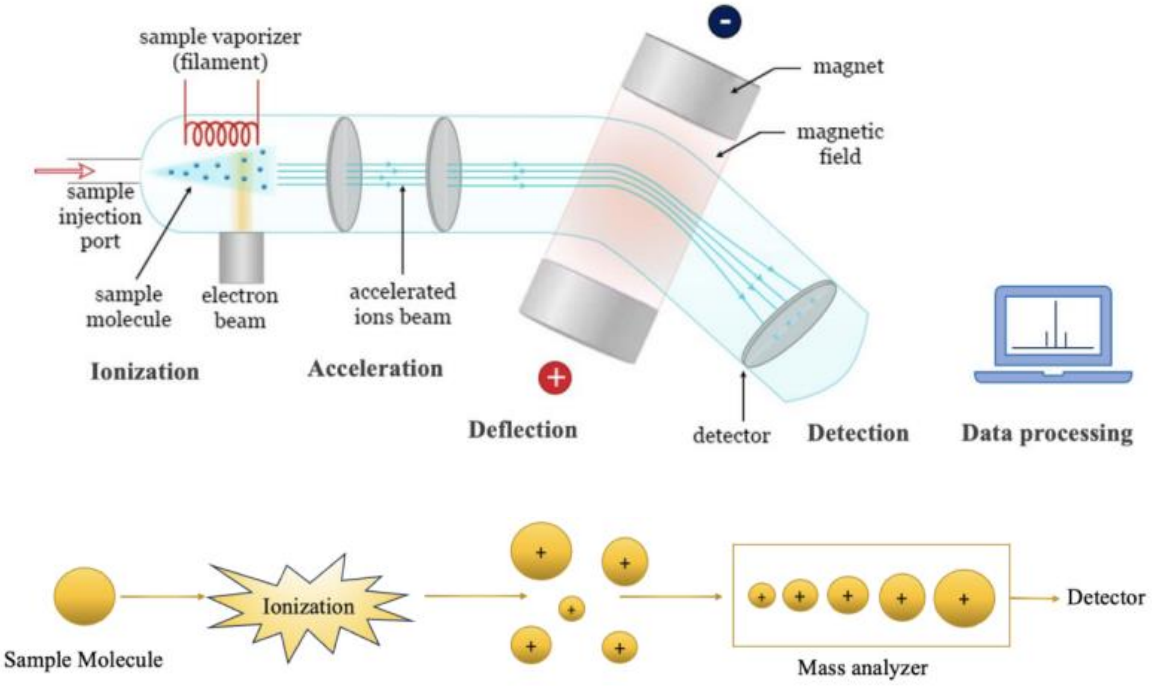

Figure I.2. Représentation schématique du principe de la technique de spectrométrie de masse et le processus d'ionisation (Kuril, 2024).

I.3.2. Types

La spectrométrie de masse (MS) se décline en plusieurs catégories selon les techniques d'ionisation employées et le type d'analyseur de masse utilisé.

I.3.2.1. Classification selon la méthode d'ionisation

- Impact électronique (EI)

L'ionisation par impact électronique (EI) est définie comme une méthode d'ionisation en spectrométrie de masse dans laquelle des électrons de haute énergie sont utilisés pour ioniser des molécules en phase gazeuse. Le processus implique l'interaction de ces électrons avec les molécules, conduisant à la formation d'ions moléculaires ainsi que de divers ions fragments en raison de l'énergie élevée transmise lors de la collision. Les ions résultants peuvent être détectés et analysés afin de fournir des informations sur la composition moléculaire de l'échantillon. Toutefois, la fragmentation inhérente à l'EI peut compliquer l'analyse, car elle peut entraîner la mauvaise identification des ions parents et affecter la quantification des analytes (Ikonomou et Rayne, 2002).

- Ionisation chimique (CI)

L'ionisation chimique (CI) est une technique d'ionisation douce utilisée en spectrométrie de masse, permettant de générer des pics d'ions moléculaires protonés facilement identifiables, ce qui est essentiel pour une analyse précise en métabolomique. Cette méthode limite la fragmentation des molécules analytes, favorisant ainsi la préservation des ions intacts (Capellades *et al.*, 2021).

La méthode CI complète le mode d'ionisation par impact électronique (EI), plus couramment utilisé, en offrant un moyen de générer des ions moléculaires tout en préservant l'intégrité de la structure moléculaire, souvent fragmentée lors d'une ionisation par impact électronique. Cette technique est particulièrement utile pour l'analyse de mélanges complexes et permet une meilleure compréhension des processus de spectrométrie de masse impliqués dans l'analyse chimique (Richer *et al.*, 2006).

- Electrospray (ESI)

L'ionisation par électrospray (ESI) est une technique largement utilisée en spectrométrie de masse, reposant sur la génération de gouttelettes chargées à partir d'un échantillon liquide. Dans le procédé ESI, une haute tension est appliquée à un liquide, produisant ainsi une fine brume de gouttelettes chargées dirigées vers le spectromètre de masse. Lors de leur trajet dans l'air, ces gouttelettes subissent un processus de désolvatation, au cours duquel le solvant s'évapore, ne laissant que des ions analytes chargés. Cette méthode est particulièrement efficace pour l'analyse

des biomolécules et autres composés de grande taille, car elle permet de produire des ions dans leur état natif sans entraîner de fragmentation significative. La technique est extrêmement polyvalente et peut être adaptée à diverses applications, notamment l'analyse de mélanges complexes et d'échantillons biologiques, ce qui en fait un outil essentiel dans des domaines tels que la protéomique et la métabolomique (Chipuk et Brodbelt, 2008).

- Désorption/Ionisation par matrice assistée par laser (MALDI)

La désorption/ionisation laser assistée par matrice (MALDI) est une technique puissante de spectrométrie de masse, principalement utilisée pour l'analyse de grandes biomolécules telles que les protéines, les peptides et les acides nucléiques. Le principe fondamental du MALDI repose sur l'utilisation d'une matrice, un composé chimique capable d'absorber l'énergie laser et de faciliter la désorption et l'ionisation des molécules d'analyte. Lorsqu'une impulsion laser est dirigée vers l'échantillon, la matrice absorbe l'énergie et se vaporise rapidement, entraînant avec elle les molécules d'analyte dans la phase gazeuse. Ce processus aboutit à la formation d'ions, qui peuvent ensuite être analysés par spectrométrie de masse. De manière générale, le MALDI est reconnu pour ses capacités d'analyse rapide et sa faculté à générer majoritairement des ions monovalents, ce qui le rend particulièrement adapté à l'analyse de mélanges complexes (Chang *et al.*, 2007).

- Ionisation à pression atmosphérique (APCI)

L'ionisation chimique à pression atmosphérique (APCI) est une technique d'ionisation en spectrométrie de masse qui fonctionne à pression atmosphérique, en utilisant une décharge corona pour ioniser les analytes en phase gazeuse. Le processus débute par l'application d'une haute tension sur une aiguille corona, générant une décharge corona qui produit un plasma énergétique. Ce plasma interagit avec le gaz environnant, ionisant ainsi les molécules d'analyte présentes dans la phase gazeuse. L'efficacité de l'ionisation est influencée par la morphologie de l'aiguille corona, en particulier par sa finesse, qui affecte la précision spatiale de la décharge. L'APCI se révèle particulièrement efficace pour l'analyse d'une large gamme de composés, qu'ils soient polaires ou apolaires, ce qui en fait un outil polyvalent pour les applications en spectrométrie de masse (Auvil et Bier, 2024).

I.3.2.2. Classification selon le type d'analyseur de masse

- Quadrupôle (Q)

La spectrométrie de masse à analyseur quadrupolaire (Q-MS) est une technique d'analyse en spectrométrie de masse qui repose sur la séparation des ions selon leur rapport masse/charge (m/z), à l'aide de quatre tiges métalliques parallèles soumises à un champ électrique oscillant. Le processus débute par l'introduction des ions générés en amont (par exemple via APCI ou ESI), qui

sont ensuite guidés à travers le quadrupôle. Ce dernier applique une combinaison de tensions continues et radiofréquences, permettant de filtrer sélectivement les ions en fonction de leur m/z. Seuls les ions stables pour un certain couple de tensions atteignent le détecteur. L'efficacité de séparation et de détection dépend fortement des paramètres du quadrupôle, tels que la fréquence de balayage, la résolution et la stabilité du champ. La spectrométrie quadrupolaire est particulièrement efficace pour les analyses quantitatives ciblées (comme le mode MRM), mais elle est aussi utilisée en screening non ciblé dans certaines configurations. Elle est largement utilisée pour sa robustesse, sa précision et sa compatibilité avec différents types de sources d'ionisation et d'analyseurs couplés (par exemple en tandem QqQ) (Alseekh *et al.*, 2021).

- Analyseur à temps de vol (TOF)

La spectrométrie de masse à analyseur à temps de vol (TOF-MS) repose sur la séparation des ions en fonction de leur rapport masse/charge (m/z) par mesure de leur temps de parcours dans un tube de vol sous vide. Après ionisation, les ions sont accélérés par un champ électrique pulsé vers un détecteur, et leur vitesse de déplacement est inversement proportionnelle à leur masse. Les ions légers atteignent le détecteur plus rapidement que les ions lourds, permettant ainsi leur différenciation temporelle. La précision de cette méthode repose sur la régularité de l'impulsion initiale et l'homogénéité du champ électrique. L'analyseur TOF offre une très large gamme de détection en m/z, une haute sensibilité, et une excellente précision en masse, bien que sa plage dynamique soit plus limitée que celle d'autres analyseurs. Il est particulièrement adapté aux analyses nécessitant une grande exactitude massique et une détection rapide, notamment dans les domaines de la protéomique, de la métabolomique exploratoire, et du criblage non ciblé en clinique (Strathmann et Hoofnagle, 2011).

- Piège à ions (Ion Trap, IT)

Les pièges à ions sont des dispositifs dans lesquels les ions sont confinés par des champs électromagnétiques pendant une durée prolongée dans un volume restreint, où la mesure de masse est également effectuée. Le mouvement des ions dans ce type de piège peut être décrit par l'équation de Mathieu, dont la solution donne accès à un diagramme de stabilité déterminé par deux paramètres, a et q, définissant les conditions nécessaires à leur confinement dans un champ quadrupolaire. Les ions peuvent être excités de manière résonante par un champ électrique faible oscillant à leur fréquence séculaire, ce qui permet leur fragmentation ou leur éjection. L'ajout d'hélium comme gaz tampon accroît l'efficacité de cette excitation résonante en favorisant la fragmentation des ions précurseurs. Tous les modes opératoires initialement développés pour les pièges tridimensionnels (3D) sont applicables aux pièges linéaires (2D), offrant ainsi une flexibilité accrue. Les pièges à ions modernes permettent d'atteindre des résolutions supérieures à

30 000 (FWHM) à condition que les vitesses de balayage soient suffisamment lentes. Grâce à leur polyvalence, ces dispositifs sont aujourd'hui largement utilisés pour analyser une grande variété de particules, allant des espèces atomiques aux protéines en passant par les nanoparticules et les molécules organiques (Nolting *et al.*, 2019),

- Orbitrap

L'analyseur Orbitrap est une technologie de spectrométrie de masse à haute résolution fondée sur la détection des oscillations d'ions piégés dans un champ électrostatique. Ce dispositif utilise une électrode centrale en forme de fuseau, autour de laquelle les ions injectés oscillent axialement sous l'effet d'un champ radial. Ces oscillations sont captées par des électrodes externes sous forme d'un courant d'image, puis transformées mathématiquement en spectre de masse à l'aide de la transformée de Fourier. L'Orbitrap est intégré dans des configurations hybrides telles que le quadrupôle-Orbitrap (Q-Orbitrap), qui combinent la capacité de sélection des ions du quadrupôle avec la haute résolution du détecteur Orbitrap. Ce système permet l'acquisition de spectres MS et MS/MS via une cellule de dissociation à haute énergie (HCD), assurant une fragmentation efficace des analytes. L'analyseur se distingue par sa grande précision massique (inférieure à 2 ppm), sa stabilité de calibration, et une résolution élevée, ce qui le rend adapté à l'analyse de composés complexes et faiblement concentrés. L'Orbitrap constitue ainsi un outil performant pour des applications avancées en spectrométrie de masse, notamment en protéomique, métabolomique et analyse structurale (Helfer *et al.*, 2015).

- Transformée de Fourier à cyclotron ionique (FT-ICR-MS)

La spectrométrie de masse par transformée de Fourier à cyclotron ionique (FT-ICR-MS) repose sur l'utilisation d'un champ magnétique intense associé à un champ électrique pour piéger les ions dans un dispositif appelé piège de Penning. Ce piège est constitué de quatre électrodes disposées dans un champ magnétique perpendiculaire au champ électrique. Les ions sont ainsi confinés et mis en rotation selon une fréquence dite cyclotron, directement proportionnelle à l'intensité du champ magnétique et inversement proportionnelle au rapport masse/charge (m/z). Ce mouvement génère un courant d'image capté par les plaques détectrices, produisant un signal appelé free induction decay (FID). Ce signal est ensuite converti en spectre de masse par transformée de Fourier (FT), fournissant ainsi une mesure très précise des ions présents. La FT-ICR-MS permet d'obtenir une résolution exceptionnelle, atteignant jusqu'à 10^6, ainsi qu'une précision massique dans la gamme des ppm à sub-ppm, ce qui autorise l'analyse de protéines intactes non digérées. Sa haute résolution se traduit également par une capacité de détection accrue, rendant possible l'observation de multiples signaux dans des échantillons complexes (Rajawat et Jhingan, 2019).

Tableau I.1. Caractéristiques des différents analyseurs de masse (Harwood et Claridge, 1997).

Méthode	Grandeur mesurée	Plage masse/charge (m/z)	Résolution à m/z = 1 000	Plage dynamique
Temps de vol (TOF)	Temps de vol	10^6	$10^3 – 10^4$	10^4
Résonance cyclotronique ionique (ICR)	Fréquence cyclotronique	10^5	10^6	10^4
Piège à ions	Fréquence	10^4	10^4	10^4
Filtre quadrupolaire	Filtre pour m/z	$10^3 – 10^4$	$10^3 – 10^4$	10^5

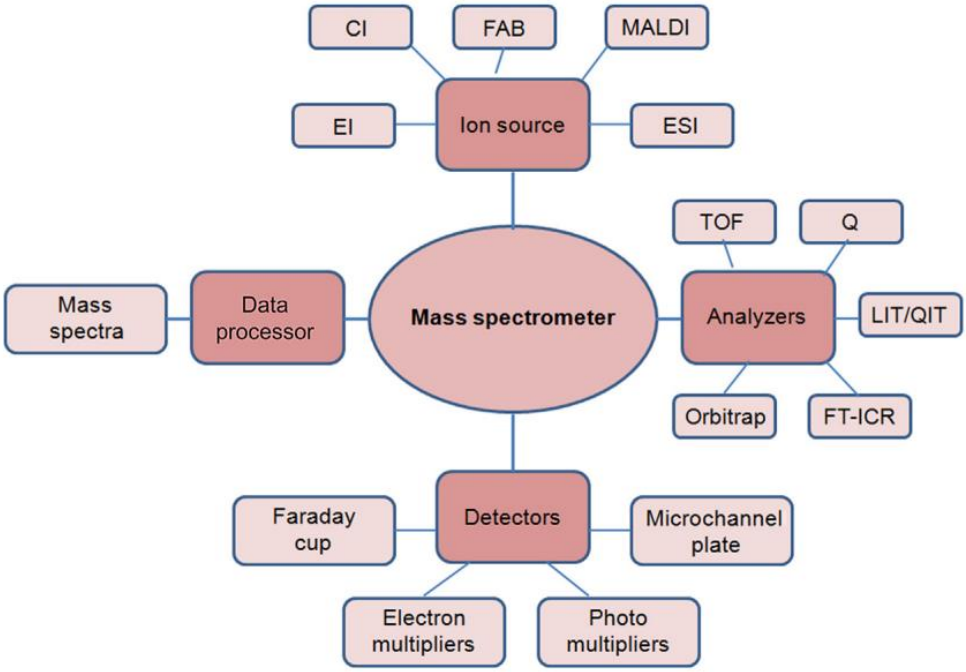

Figure I.3. Composants d'un spectromètre de masse (Rajawat et Jhingan, 2019).

I.3.2.3. Systèmes hybrides

Les systèmes hybrides en spectrométrie de masse (MS) ont considérablement élargi les capacités analytiques dans divers domaines, notamment en protéomique, en métabolomique et en pharmacologie. Parmi les instruments les plus représentatifs, on retrouve le Q-TOF, le Q-Orbitrap et le triple quadrupôle (QqQ), chacun combinant des technologies complémentaires pour optimiser la sensibilité, la sélectivité et la précision massique.

Le Q-TOF (Quadrupole-Time of Flight) est un système hybride qui associe un quadrupôle, servant à filtrer les ions selon leur rapport masse/charge, à un analyseur de type temps de vol (TOF). Cette combinaison permet une sélection ionique ciblée suivie d'une mesure précise du

temps de vol des ions filtrés, lequel est directement corrélé à leur masse. Cette configuration procure une excellente sensibilité et une haute précision de masse, ce qui en fait un outil particulièrement efficace pour l'analyse de mélanges complexes (Chernushevich *et al.*, 2001).

De son côté, le Q-Orbitrap associe également un quadrupôle à un autre analyseur de haute performance, l'Orbitrap. Le quadrupôle isole les ions précurseurs, tandis que l'analyseur Orbitrap réalise des mesures à très haute résolution et avec une précision massique généralement comprise entre 1 et 5 ppm. Ce type d'appareil est couramment utilisé dans les méthodes ciblées telles que le Parallel Reaction Monitoring (PRM), permettant la détection de tous les fragments issus d'un ion précurseur donné dans un spectre MS2 complet. Le Q-Orbitrap offre également la possibilité d'un multiplexage à l'échelle MS/MS, autorisant l'isolation simultanée de plusieurs précurseurs et la génération d'un spectre composite intégrant tous les fragments (Vidova et Spacil, 2017).

Enfin, le triple quadrupôle (QqQ) constitue un système linéaire intégrant trois quadrupôles successifs : Q1 et Q3 pour la sélection des ions, et q2 pour leur fragmentation. Utilisé principalement en mode Selected Reaction Monitoring (SRM) (Fig. I.4), ce dispositif permet une analyse hautement spécifique grâce à la détection ciblée de transitions spécifiques entre un ion précurseur et un ion fragment (Domon et Aebersold, 2006). Le QqQ se démarque par sa très grande sensibilité (jusqu'à 10 attomoles), sa compatibilité avec les méthodes chromatographiques rapides (telles que l'UHPLC), et sa capacité de multiplexage élevée, avec la possibilité d'analyser plus de 1000 transitions par cycle (Liebler et Zimmerman, 2013). Toutefois, sa nature ciblée impose une sélection préalable des transitions d'intérêt, ce qui limite l'exploration de nouveaux analytes après l'acquisition des données (Carr *et al.*, 2014).

Ainsi, chacun de ces instruments hybrides présente des avantages spécifiques selon les objectifs analytiques visés, qu'il s'agisse de précision massique, de sensibilité, de sélectivité ou de capacité de multiplexage.

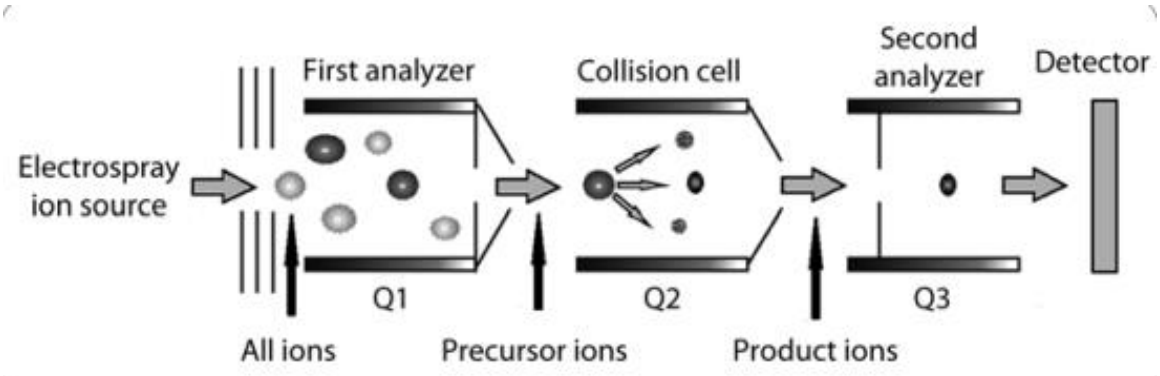

Figure I.4. Schéma de l'analyse en spectrométrie de masse en mode SRM (spectromètre de masse à triple quadrupôle, QqQ) (Faktor *et al.*, 2012).

I.4. Couplage chromatographie-MS

La plupart des utilisations de la spectrométrie de masse sont réalisées en combinaison avec une séparation chromatographique, principalement sous la forme des techniques GC/MS ou LC/MS. Ces associations ont été employées, par exemple, dans l'analyse organique en sciences de l'environnement ainsi que dans la caractérisation de composés biologiques, incluant la détermination de leur masse moléculaire (MM) et l'analyse de séquences de biopolymères (Burlingame *et al.*, 1996).

I.4.1. GC-MS

La chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS) repose sur le principe de la séparation des composés chimiques présents dans un mélange, suivie de leur identification par analyse de masse. Le processus commence par la vaporisation de l'échantillon, lequel est ensuite introduit dans une colonne chromatographique où les différents constituants sont séparés en fonction de leurs propriétés physiques et chimiques spécifiques. Après cette étape de séparation, les composés individuels sont ionisés dans le spectromètre de masse, permettant ainsi la détermination de leur rapport masse/charge. Cette combinaison de séparation et d'analyse de masse permet une identification précise de mélanges complexes, faisant de la GC-MS une technique analytique de premier plan. En somme, la GC-MS se définit comme une méthode analytique intégrant la chromatographie en phase gazeuse et la spectrométrie de masse, afin d'assurer une séparation et une identification de haute qualité des substances chimiques. Elle est largement utilisée dans divers domaines tels que la surveillance environnementale, l'analyse pharmaceutique ou encore la science médico-légale, en raison de son efficacité à analyser des mélanges complexes avec une grande précision et exactitude (Khalifea et Ali, 2025).

I.4.2. LC-MS

Le couplage chromatographie liquide – spectrométrie de masse (LC–MS) représente une technique analytique essentielle dans le domaine de la découverte de nouveaux médicaments. Elle repose sur l'association de deux systèmes complémentaires : la chromatographie liquide haute performance (HPLC), qui permet la séparation des analytes dans un échantillon complexe, et la spectrométrie de masse (MS), qui permet leur détection et identification avec une grande sensibilité. Le principe du LC–MS consiste à combiner ces deux technologies afin de bénéficier à la fois de la puissance de séparation de la chromatographie et de la capacité de détection précise de la spectrométrie de masse. Un système LC–MS typique se compose de quatre éléments principaux : un auto injecteur pour introduire les échantillons, un système HPLC chargé de séparer les composés d'intérêt, une source d'ionisation qui sert d'interface entre la HPLC et le spectromètre, et enfin le spectromètre de masse lui-même. Ces composants sont en général pilotés

via un même logiciel de contrôle. Durant l'analyse, les composés séparés par la HPLC sont acheminés vers la source d'ionisation, où ils sont convertis en ions, avant d'être détectés et analysés par le spectromètre de masse.

Un nouvel outil chromatographique, l'ultra-high performance liquid chromatography (UPLC), a récemment fait son apparition. L'UPLC utilise des colonnes garnies de particules de très petite taille ainsi que des pompes à haute pression, ce qui permet d'atteindre une résolution chromatographique nettement supérieure. Cette technologie présente un fort potentiel pour les applications d'identification des métabolites, et des références à des systèmes UPLC–MS sont désormais disponibles dans la littérature scientifique. À mesure que ces systèmes se démocratisent, ils deviendront des outils essentiels pour les chercheurs spécialisés dans l'identification des métabolites (Korfmacher, 2005).

Chapitre II : Concepts de l’intelligence artificielle

Chapitre II. Concepts de l'intelligence artificielle

II.1. Introduction à l'intelligence artificielle

II.1.1. Définition

Avec les avancées des technologies informatiques et l'émergence de nouveaux algorithmes intelligents, l'objectif de l'intelligence artificielle (IA) s'est rapproché de manière significative. L'IA désigne une forme d'intelligence simulée sur des machines programmables (Fig. II.1), visant à reproduire, dans une certaine mesure, les capacités cognitives du cerveau humain (Naeem *et al.*, 2020).

De manière plus générale, l'intelligence artificielle désigne la branche de l'informatique consacrée à la création de systèmes capables d'exécuter des tâches qui nécessitent habituellement l'intelligence humaine. Il s'agit d'un concept vaste qui englobe une grande diversité de sous-domaines et de techniques (Chartrand *et al.*, 2017).

Les systèmes d'intelligence artificielle (IA) peuvent être entraînés à accomplir diverses tâches, telles que la reconnaissance d'images et de la parole, la prise de décisions et la traduction linguistique (Litjens *et al.*, 2017). L'IA est devenue de plus en plus présente dans la société, avec des applications allant des soins de santé aux véhicules autonomes (Bohr et Memarzadeh, 2020). De plus, l'IA a permis de développer de nouveaux algorithmes et d'obtenir des perspectives innovantes qui ont contribué à l'amélioration de nombreux processus existants (Cordero *et al.*, 2020).

II.1.2. Historique

L'histoire de l'intelligence artificielle (IA) débute dans les années 1940 avec l'influence de la science-fiction, notamment la publication en 1942 de la nouvelle Runaround d'Isaac Asimov, qui introduit les Trois Lois de la Robotique (Haenlein et Kaplan, 2019). En parallèle, Alan Turing développe une machine de décryptage pendant la Seconde Guerre mondiale, avant de proposer en 1950 le célèbre test de Turing comme critère d'intelligence artificielle (TURING, 1950). Le terme « intelligence artificielle » est formalisé lors de la conférence de Dartmouth en 1956, marquant le début d'une phase d'optimisme et d'investissements (Haenlein et Kaplan, 2019).

La phase initiale de la recherche en IA s'est concentrée sur le développement d'algorithmes et de modèles capables d'imiter le raisonnement et la résolution de problèmes humains. Les chercheurs ont cherché à représenter les connaissances sous différentes formes, telles que la logique, les règles et le langage naturel, afin de créer des systèmes intelligents capables de réaliser des tâches nécessitant généralement une intelligence humaine.

Aujourd'hui, l'IA connaît un essor significatif, avec des applications touchant divers aspects de la vie quotidienne. Le domaine a évolué pour intégrer un large éventail de techniques et de théories visant à créer des artefacts intelligents capables de comprendre, percevoir et prendre des décisions, comblant ainsi le fossé entre les processus cognitifs humains et les capacités des machines (Zaraté, 2021).

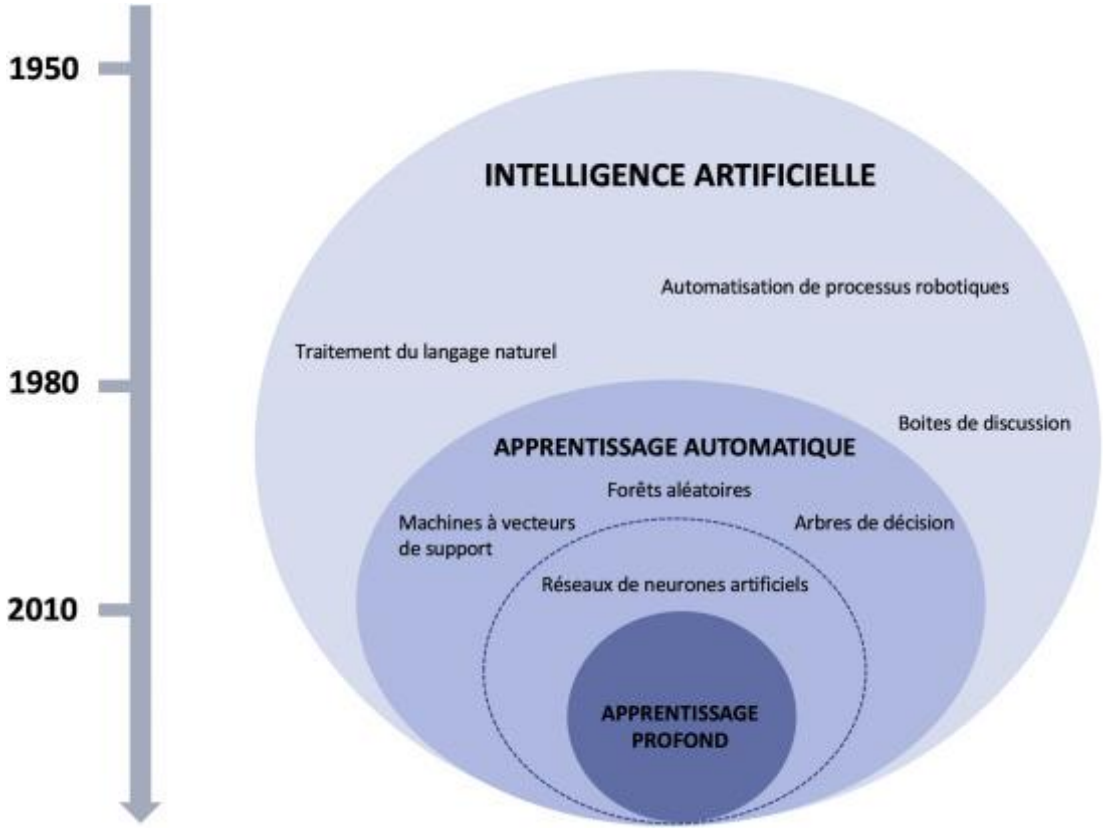

Figure II.1. Les concepts de l'intelligence artificielle (Bunod *et al.*, 2022).

II.2. Principales approches de l'intelligence artificiels

II.2.1. Apprentissage automatique (Machine learning)

L'apprentissage automatique (ML) (Fig. II.2) est un sous-domaine de l'intelligence artificielle (IA) qui met l'accent sur l'aspect apprentissage de l'intelligence. Il consiste à développer des programmes informatiques capables d'apprendre et de s'améliorer à partir de l'expérience, sans nécessiter une programmation explicite. Cette approche contraste avec les programmes informatiques traditionnels, qui requièrent des instructions spécifiques détaillant chaque étape que le programme doit suivre (Ahuja, 2019).

II.2.1.1. Apprentissage Supervisé

L'apprentissage supervisé est considéré comme une approche orientée par les tâches, où l'on dispose d'entrées et de sorties souhaitées. Un modèle mathématique est conçu pour établir une

correspondance entre les entrées et les sorties attendues. Les données non vues doivent être assignées à une classe de la manière la plus précise possible sur la base de ce modèle. L'objectif principal est de minimiser le risque ou l'erreur (Prakash *et al.*, 2018). Les algorithmes d'apprentissage supervisé sont capables d'effectuer à la fois des tâches de régression (prédiction d'une variable continue) et de classification (prédiction d'une variable discrète) (Kotsiantis, 2007).

II.2.1.2. Apprentissage Non Supervisé

L'apprentissage non supervisé étudie la manière dont les systèmes peuvent apprendre à représenter certains motifs d'entrée de manière à refléter la structure statistique de l'ensemble des données observées.

Les seules informations disponibles pour les méthodes d'apprentissage non supervisé sont les motifs d'entrée observés, souvent supposés être des échantillons indépendants d'une distribution de probabilité sous-jacente inconnue ainsi que certaines connaissances explicites ou implicites a priori sur ce qui est jugé important (Reddy *et al.*, 2018).

II.2.1.3. Méthodes Avancées

Il existe de nombreuses autres méthodes d'apprentissage automatique, plus complexes et performantes, qui sont particulièrement adaptées à la conception de petites molécules

II.2.1.3.1. Apprentissage Semi-Supervisé

L'apprentissage semi-supervisé (SSL) est une technique d'apprentissage automatique (ML) qui se situe à mi-chemin entre l'apprentissage supervisé et l'apprentissage non supervisé, c'est-à-dire que le jeu de données est partiellement étiqueté. L'apprentissage semi-supervisé permet d'exploiter des données partiellement annotées en apprenant des relations entre caractéristiques et étiquettes tout en utilisant les informations des données non annotées. Cette approche améliore la précision des modèles tout en réduisant l'effort d'annotation manuelle. Il est largement utilisé en conception de petites molécules, notamment pour la prédiction d'activité chimique, l'analyse métabolique, et la prédiction des cibles médicamenteuses (Bahi et Batouche, 2018).

II.2.1.4. Apprentissage Par Renforcement

Il s'agit d'une forme d'apprentissage située entre l'apprentissage supervisé et l'apprentissage non supervisé. Elle est utilisée pour les mêmes applications que l'apprentissage supervisé. De grandes quantités de données non étiquetées et de petites quantités de données étiquetées sont couramment exploitées (Aggarwal *et al.*, 2022). Un problème de RL est résolu via un processus d'apprentissage par essai-erreur, qui émule le comportement d'apprentissage humain. Un agent de

RL interagit avec un, dans le but de maximiser la récompense cumulative issue de ses actions (Datta *et al.*, 2021).

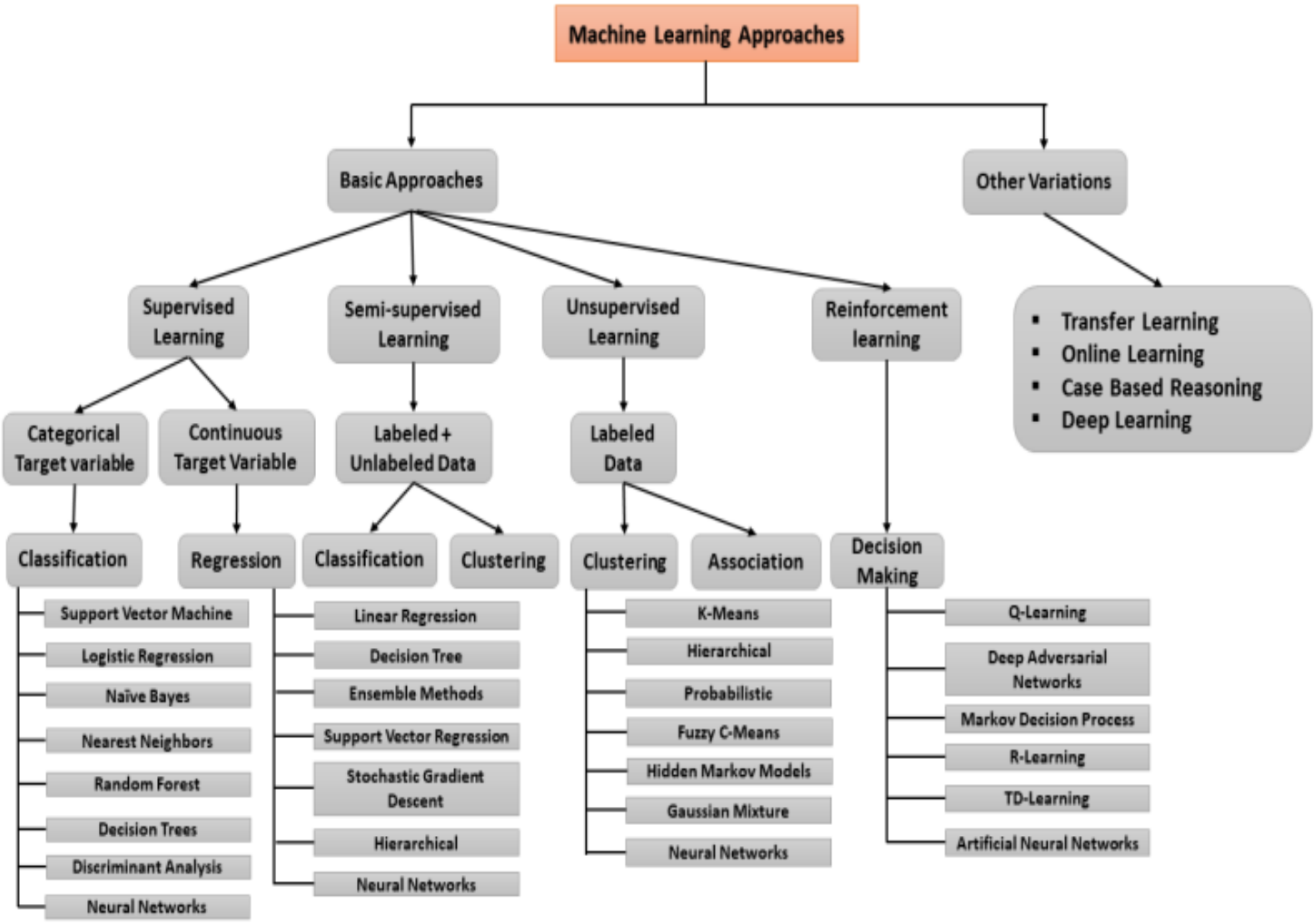

Figure II.2. Taxonomies de l'approche en apprentissage automatique (Rahman *et al.*, 2024).

II.2.1.5. Réseaux de neurones artificiels

Une classe d'algorithmes d'apprentissage automatique, appelée réseaux de neurones artificiels (Fig. II.3) (ANNs), a gagné en popularité en raison des récentes avancées en puissance de calcul, notamment sous forme d'unités de traitement graphique (GPU). Cette méthode est extrêmement polyvalente et capable à la fois de classification et de régression. Elle peut utiliser des cadres d'apprentissage supervisé, non supervisé et par renforcement, et donne les meilleurs résultats sur de grands ensembles de données (Lindley *et al.*, 2024).

Il convient toutefois de préciser que la notion des réseaux neuronaux superficiels (à faible nombre de couches cachées) relève historiquement du champ général du machine learning, tandis que le deep learning désigne spécifiquement leur extension vers des architectures comportant plusieurs couches de traitement empilées (Schmidhuber, 2015).

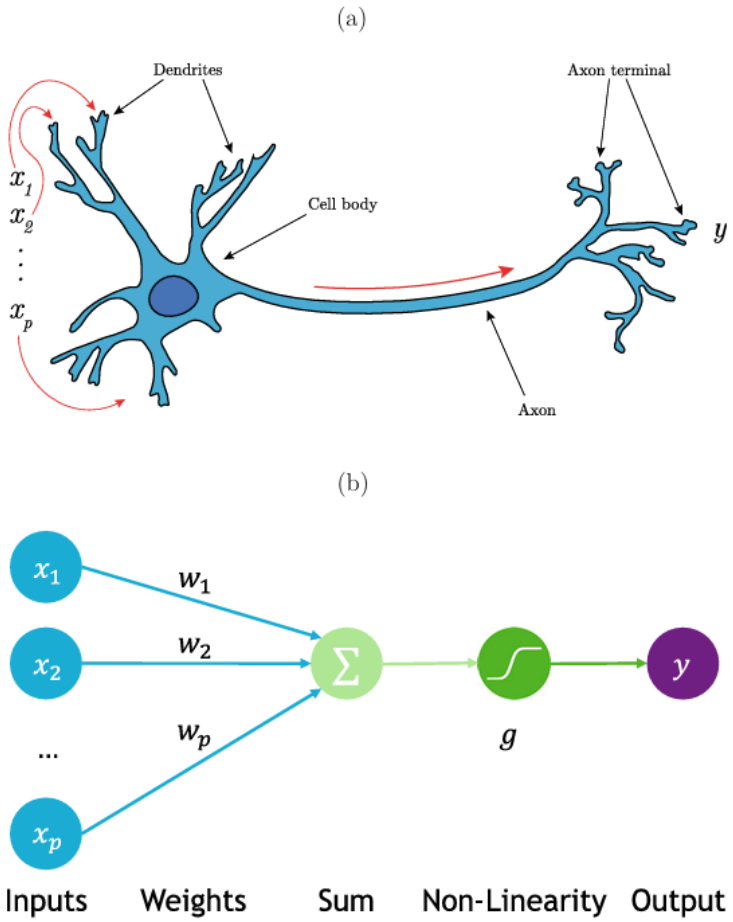

Figure II.3. Neurone biologique et perceptron (Colliot, 2023).

II.2.2. Apprentissage en profondeur (Deep learning)

Parmi les techniques relevant du domaine de l'apprentissage automatique (ML), l'apprentissage profond (DL) (Fig. II.4) s'est imposé comme l'une des plus prometteuses. En effet, le DL est une technique appartenant au ML, qui lui-même fait partie de la famille plus large de l'intelligence artificielle (Pesapane *et al.*, 2018).

L'apprentissage profond, également appelé apprentissage par réseaux neuronaux profonds, est un domaine de recherche récent et prometteur qui produit des résultats remarquables et connaît une croissance rapide en utilisant des réseaux neuronaux comportant de nombreuses couches généralement plus de 20 pour apprendre automatiquement des caractéristiques à partir des données (Erickson *et al.*, 2017).

« Profond » est un terme technique qui fait référence au nombre de couches dans un réseau de neurones artificiels (RNA). Il existe trois types de couches : la couche d'entrée (qui reçoit les données d'entrée), la couche de sortie (qui produit le résultat du traitement des données) et la couche cachée (qui extrait les motifs au sein des données).

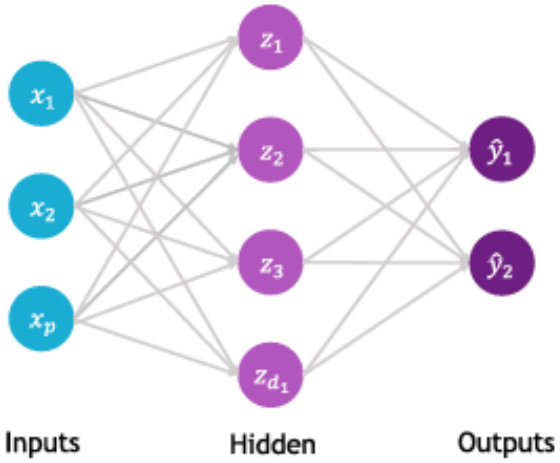

Figure II.4. Un modèle de perceptron multicouche (Colliot, 2023).

Un RNA profond se distingue du RNA superficiel (à une seule couche cachée) par le fait qu'il possède un grand nombre de couches cachées, ce qui lui permet d'accomplir des tâches plus complexes (Pesapane *et al.*, 2018).

II.2.2.1. Architectures fondamentales des réseaux neuronaux profonds

II.2.2.1.1. Réseaux de neurones convolutifs (CNN) et traitement de l'image

Un réseau de neurones convolutifs (Convolutional Neural Network) (Fig. II.5) est un type de modèle conçu pour le traitement de données présentant une topologie en grille, telles que les images. Les CNN disposent d'une grande capacité d'apprentissage, ajustable en modifiant leur profondeur et leur largeur. Ils reposent sur des hypothèses fortes concernant la nature des images, notamment la stationnarité des statistiques et la localité des dépendances entre pixels (Krizhevsky *et al.*, 2017).

Les CNN s'inspirent du fonctionnement biologique, en imitant le processus de perception visuelle humaine. Ils sont composés de plusieurs éléments fondamentaux, tels que les couches de convolution, les fonctions d'activation et les couches de regroupement (pooling), qui collaborent pour traiter efficacement les données d'entrée et en apprendre les représentations pertinentes (Li *et al.*, 2022).

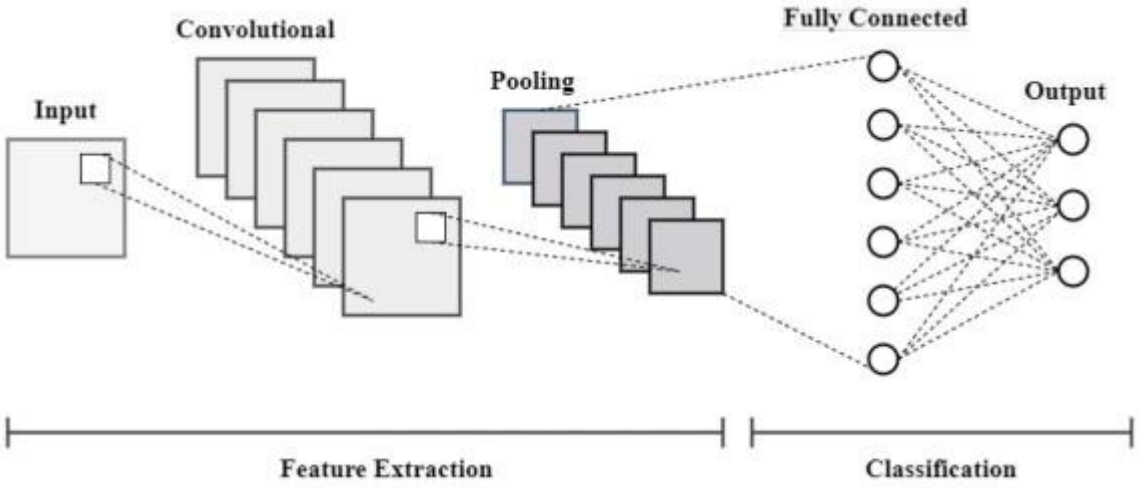

Figure II.5. Représentation schématique des réseaux de neurones convolutifs (CNNs) (Alharbi *et al.*, 2024).

II.2.2.1.2. Réseaux récurrents (RNN) et modélisation séquentielle

Les réseaux de neurones récurrents (Recurrent Neural Networks, RNNs) constituent une catégorie de modèles d'apprentissage profond spécifiquement conçus pour le traitement de données séquentielles. Les RNNs se distinguent par leur capacité à mémoriser les entrées précédentes grâce à l'utilisation d'un état interne (ou mémoire), leur permettant de traiter efficacement des séquences d'informations. Cette particularité les rend particulièrement adaptés à des applications telles que le traitement automatique du langage naturel, la reconnaissance vocale ou encore la prévision de séries temporelles, où le contexte et l'ordre des données jouent un rôle essentiel (Mienye *et al.*, 2024).

II.2.2.1.3. Transformeurs : révolution dans le traitement du langage naturel

Les transformeurs (Transformers) (Fig. II.6) sont une architecture de réseau neuronal profond reposant entièrement sur un mécanisme d'attention, et plus spécifiquement d'auto-attention (self-attention), permettant de capturer les relations contextuelles et les dépendances globales entre les éléments d'une séquence, sans recourir à des structures récurrentes ni convolutives. Cette approche innovante autorise un traitement parallèle des données, ce qui entraîne des temps d'apprentissage réduits et une amélioration des performances sur des tâches telles que la traduction automatique (Vaswani *et al.*, 2017).

Grâce à leur capacité à apprendre à partir de segments entiers d'une séquence, les transformeurs se montrent particulièrement efficaces pour gérer les dépendances à long terme, et ont démontré un potentiel remarquable dans divers domaines tels que le traitement automatique du langage naturel, la vision par ordinateur et d'autres encore (Islam *et al.*, 2024).

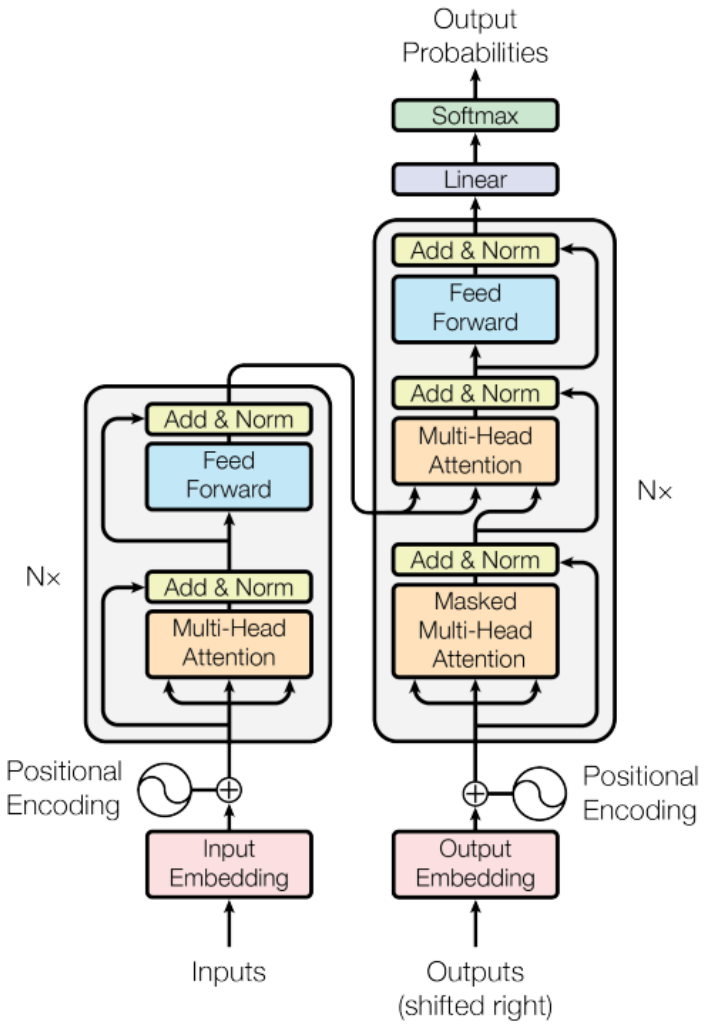

Figure II.6. L'architecture du modèle Transformer (Vaswani *et al.*, 2017).

II.3. Machine learning vs Deep learning

Il existe une synergie entre les deux approches pour atteindre les objectifs fixés : l'apprentissage automatique (ML) et l'apprentissage profond (DL) (Tab. II.1) (Tab. II.2). Bien que dépendant de l'extraction de caractéristiques et d'un format de données structuré pour une performance optimale, les architectures de DL s'adaptent particulièrement bien aux données non structurées (images et textes), pour lesquelles aucune extraction manuelle de caractéristiques n'est nécessaire. Les deux approches ont évolué afin de répondre à des défis spécifiques. Par ailleurs, tandis que les modèles de base du ML sont faciles à interpréter et souvent très efficaces, le DL

permet d'atteindre un niveau de performance supérieur grâce à l'utilisation de réseaux de neurones profonds dans ses formes les plus avancées (Howard *et al.*, 2017).

En comparant les deux modèles, on distingue des différences en termes de charge computationnelle, de capacité à s'étendre à d'autres problèmes, ainsi que d'interprétabilité des modèles. À cela s'ajoutent les domaines d'application les plus appropriés pour chaque type. La connaissance de ces distinctions est précieuse car elle permet de déterminer dans quel contexte utiliser le ML ou le DL.

En tant que sous-domaines de l'intelligence artificielle (IA), l'apprentissage automatique **(**ML**)** et l'apprentissage profond (DL) se concentrent sur la création de modèles capables de reconnaître des motifs dans les données et de prendre des décisions ou de formuler des prédictions sans programmation explicite. Selon la technique et l'application utilisées, les architectures d'apprentissage automatique peuvent varier considérablement. Dans le cadre de l'apprentissage profond, les structures de données complexes sont modélisées à l'aide de réseaux de neurones multi-couches. Composées de couches interconnectées de nœuds (ou neurones), ces architectures s'inspirent du fonctionnement du cerveau humain (Choudhary et Choudhary, 2024).

Tableau II.1. Comparaison des architectures d'apprentissage automatique (ML) et d'apprentissage profond (DL) (Choudhary et Choudhary, 2024).

Aspect	**Architectures d'apprentissage automatique (ML)**	**Architectures d'apprentissage profond (DL)**
Structure générale	Structures plates ou peu profondes (ex. : arbres de décision, modèles linéaires) avec un nombre limité d'unités interconnectées.	Comprend plusieurs couches, souvent hiérarchiques (ex. : CNN, RNN), avec de nombreux neurones par couche.
Flux de données	Flux généralement direct ou séquentiel, avec un passage unique des entrées aux sorties (ex. : régression linéaire).	Flux de données en couches, souvent avec rétropropagation à travers plusieurs couches cachées.

Types de composants	Nœuds représentant des décisions (ex. : arbres de décision) ou des hyperplans (ex. : SVM).	Inclut des neurones, convolutions, fonctions de regroupement et d'activation (ex. : ReLU, SoftMax).
Configuration des couches	Modèles tels que les arbres de décision ou la régression linéaire ont souvent une seule couche.	Architectures multicouches ; les CNN utilisent des couches de convolution ; les RNN utilisent des boucles et connexions de rétroaction.
Connexions	Modèles indépendants sans récurrence (ex. : les SVM ne font pas référence aux états précédents).	CNN : connexions locales (filtres) ; RNN : boucles récurrentes (pour les données séquentielles).
Processus d'apprentissage	Entraînement avec des algorithmes simples tels que la descente de gradient ou des mises à jour basées sur des règles.	Utilise la rétropropagation et la descente de gradient à travers plusieurs couches pour la mise à jour des poids.
Exemples d'architectures	Régression linéaire, arbres de décision, k-NN, forêts aléatoires.	CNN pour la vision, RNN pour les données séquentielles, GAN pour la génération de contenu.

Tableau II.2. Comparaison des techniques d'apprentissage automatique et d'apprentissage profond (Choudhary et Choudhary, 2024).

Catégorie de technique	Techniques ML	Techniques DL
Modèles linéaires	Régression linéaire, régression logistique	N/A
Modèles basés sur les arbres	Arbres de décision, forêts aléatoires, gradient boosting	N/A (utilisés dans les architectures ensemblistes)
Apprentissage basé sur les instances	K plus proches voisins (k-NN)	N/A
Modèles probabilistes	Naïve Bayes	N/A
Machines à vecteurs de support	SVM	N/A
Techniques de regroupement	K-means, regroupement hiérarchique	N/A
Modèles d'extraction de caractéristiques	ACP, analyse factorielle	Autoencodeurs, autoencodeurs variationnels (VAE)
Réseaux de neurones	N/A	Perceptron multicouche (MLP), CNN, RNN
Architectures avancées	N/A	LSTM, GAN, Transformateurs, mécanismes d'attention
Apprentissage par renforcement	Q-Learning, SARSA	DQN, DDPG

II.4. Synergie entre le Machine learning et le Deep learning

Le deep learning (DL) et le machine learning (ML) sont deux approches qui coexistent et interagissent dans l'ensemble des domaines de l'intelligence artificielle (Fig. II.7).

La synergie entre le ML, repose sur des algorithmes tels que les arbres de décision et les machines à vecteurs de support pour le traitement de données structurées, et le DL, qui utilise des réseaux de neurones pour traiter des données non structurées telles que les images et le langage naturel. La combinaison de ces paradigmes au sein de modèles hybrides ML-DL a permis

d'améliorer la précision des prédictions, la scalabilité et l'automatisation dans divers domaines tels que la santé, la finance, le traitement automatique du langage naturel et la robotique (Choudhary et Choudhary, 2024).

Les technologies de diagnostic et de prédiction (branche bleue) s'appuient sur l'apprentissage automatique (machine learning, ML) comme fondement de la plupart des tâches en intelligence artificielle, telles que l'analyse prédictive, la classification supervisée et non supervisée, la prise de décision et le regroupement. En outre, l'apprentissage profond (deep learning, DL) vient renforcer les capacités offertes par le ML en intégrant des réseaux neuronaux avancés tels que les CNN (Convolutional Neural Networks) et les RNN (Recurrent Neural Networks), permettant d'améliorer automatiquement les caractéristiques des données à partir d'un volume massif d'informations (Woschank *et al.*, 2020).

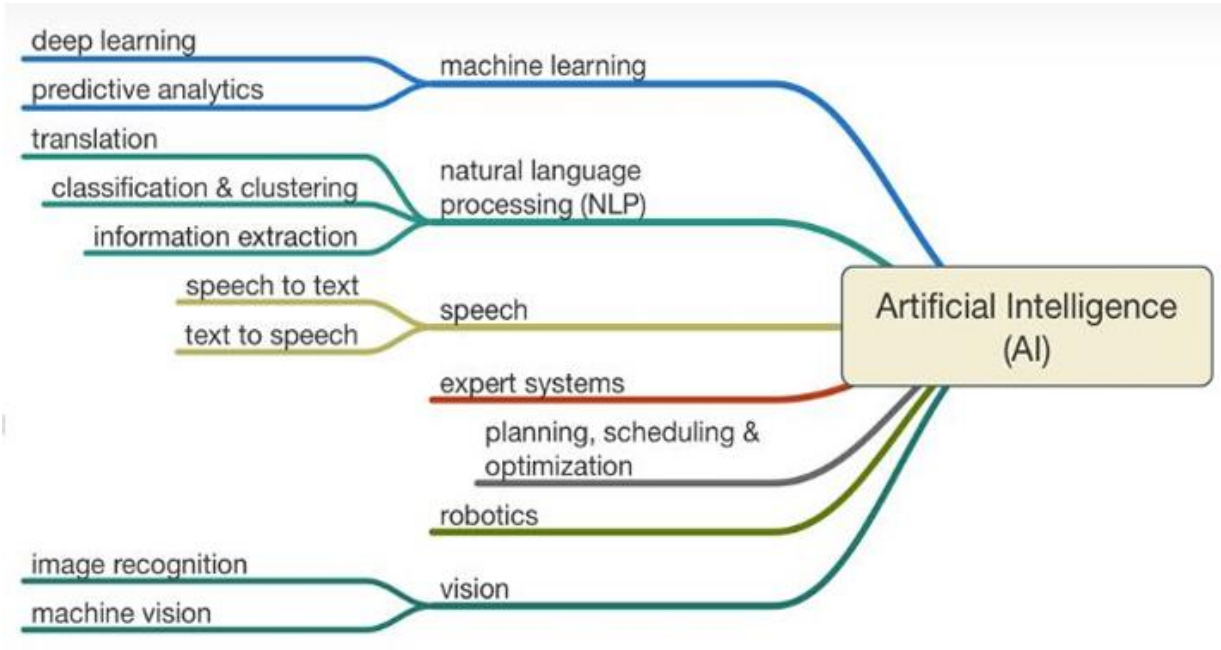

Figure II.7. Interconnexion entre l'apprentissage automatique et l'apprentissage profond dans les domaines de l'intelligence artificielle (Badmus *et al.*, 2024).

La combinaison de l'apprentissage automatique (ML) avec l'apprentissage profond (DL) présente certains avantages, mais comporte également plusieurs inconvénients qu'il convient de surmonter autant que possible. Un problème majeur réside dans la complexité computationnelle. Les modèles de grande envergure, tels que les réseaux de neurones convolutifs (CNN) et les transformeurs, exigent une puissance de calcul considérable, ce qui accroît l'utilisation des ressources ainsi que les coûts organisationnels. Toutefois, lorsqu'ils sont intégrés aux algorithmes de ML, le temps nécessaire à l'entraînement ainsi que la consommation énergétique tendent à

augmenter, ce qui confère une importance cruciale à l'efficacité des optimiseurs. Par ailleurs, les exigences en matière de données sont généralement élevées, comme cela a été observé dans diverses études de contextualisation. Les modèles DL sont généralement plus performants sur de grands ensembles de données, tandis que les modèles ML sont mieux adaptés aux petits jeux de données. Néanmoins, la combinaison des deux types de modèles se heurte à la nécessité de disposer de jeux de données volumineux pour que chaque composant fonctionne efficacement (Bishop, 2006).

Partie pratique

Chapitre III : Matériels et méthodes

Chapitre III. Matériels et méthodes

III.1. Matériels

III.1.1. Data set

MassBank est une bibliothèque spectrale de masse en libre accès dédiée à l'identification de petites molécules chimiques d'intérêt en métabolomique, expozomique et sciences de l'environnement. La grande majorité des données actuellement disponibles dans MassBank sont issues de la spectrométrie de masse à haute résolution, bien que tous types de données spectrales soient acceptés. Un large éventail d'options de recherche permet d'explorer la base de données de manière flexible et approfondie.

La bibliothèque MassBank repose sur des enregistrements au format texte, incluant des métadonnées descriptives et les informations spectrales correspondantes selon le format standardisé MassBank. L'ensemble des données est archivé sur GitHub et Zenodo, tandis que le code source est également disponible sur GitHub. La bibliothèque peut être téléchargée dans différents formats, notamment en fichiers texte, bases de données (fichiers SQL) ou fichiers MSP.

Le projet MassBank est maintenu et développé par le consortium MassBank, avec le soutien de l'association NORMAN, du FNR (Fonds National de la Recherche) et de l'initiative NFDI4Chem. Les principaux contributeurs au développement et à la maintenance de MassBank et des outils associés sont les équipes de Steffen Neumann (Leibniz Institute of Plant Biochemistry, IPB, Halle/Saale, Allemagne), Michael Stravs (Swiss Federal Institute of Aquatic Science and Technology, EAWAG, Dübendorf, Suisse), Emma Schymanski (Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, LCSB, Belvaux, Luxembourg) et Tobias Schulze (Helmholtz Centre for Environmental Research, UFZ, Leipzig, Allemagne).

Le jeu de données principal utilisé pour le développement du modèle prédictif a été extrait de MassBank_NIST_version_2024.11.msp, une base de données en libre accès de haute qualité spécialisée dans la spectrométrie de masse. Il comprend 122 512 entrées, chacune correspondant à un spectre MS unique ou répliqué, annoté avec des métadonnées structurales et instrumentales détaillées, acquises dans des conditions expérimentales standardisées. En regroupant les entrées redondantes selon leur identifiant InChIKey (International Chemical Identifier Key), utilisé comme empreinte structurale unique, le jeu de données révèle un total de 18 332 entités moléculaires distinctes, chacune correspondant à une structure chimique unique.

Chaque enregistrement comporte 20 attributs, couvrant à la fois des descripteurs chimiques et des paramètres spectrométriques (Tab. III.1). Ces attributs incluent les noms des composés, les

formules moléculaires, les masses exactes monoisotopiques, les poids moléculaires, ainsi que des listes de pics haute résolution. Le jeu de données renseigne également des variables expérimentales telles que la masse sur charge (m/z) de l'ion précurseur, le type d'ion précurseur, le mode d'ionisation, l'énergie de collision et les caractéristiques de l'instrument utilisé.

Dans l'ensemble, ce jeu de données constitue une base solide, chimiquement diversifiée et structurée de manière cohérente, pour le développement de modèles d'intelligence artificielle appliqués à l'identification des métabolites.

Tableau III.1. Paramètres de classement descriptifs du dataset.

N°	Paramètre	Description
1	Name	Nom du métabolite ou du composé. Présent dans la quasi-totalité des entrées.
2	Synon	Synonymes du composé. Présents dans environ 50 % des entrées.
3	DB#	Identifiants internes à la base MassBank.
4	InChIKey	Identifiant chimique unique (clé InChI) utilisé pour distinguer les structures moléculaires. Présent dans >95 % des entrées.
5	InChI	Chaîne InChI (International Chemical Identifier), représentant la structure exacte. Presque complet.
6	SMILES	Représentation linéaire de la structure chimique (Simplified Molecular Input Line Entry System). Très largement présente.
7	Precursor_type	Type d'ion précurseur (ex. : [M+H]+, [M−H]−), utile pour l'interprétation du spectre MS/MS.
8	Spectrum_type	Type de spectre (ex. : MS2, MSn). Disponible pour presque toutes les entrées.
9	PrecursorMZ	Rapport masse/charge (m/z) de l'ion précurseur. Présent dans la majorité des cas.
10	Instrument_type	Catégorie de l'instrument (ex. : QTOF, Ion Trap, Orbitrap).
11	Instrument	Dénomination spécifique de l'instrument utilisé pour l'acquisition. Donnée bien remplie dans la base.
12	Ion_mode	Mode d'ionisation utilisé (positif ou négatif). Quasi systématiquement renseigné.
13	Collision_energy	Énergie de collision appliquée lors de la fragmentation (en eV ou arbitraire).
14	Formula	Formule chimique brute du composé (ex. : C6H12O6).
15	MW	Poids moléculaire (masse moyenne) calculé à partir de la formule chimique.
16	ExactMass	Masse exacte calculée, prenant en compte les masses isotopiques réelles.
17	Comments	Remarques techniques ou conditions spécifiques de l'expérience.
18	Splash	Identifiant universel de spectre (SPLASH), permettant la recherche de spectres similaires.
19	Num Peaks	Nombre de pics dans le spectre MS/MS associé.
20	Peaks	Liste des pics spectrométriques : paires m/z – intensité normalisée. Format standardisé dans toutes les entrées.

III.1.2. Machine

Tous les travaux de prétraitement des données spectrales, de développement des modèles et d'entraînement des réseaux de neurones ont été réalisés sur un Apple MacBook Air M1 (2020). Cette machine est équipée d'une puce Apple Silicon M1, intégrant un CPU à 8 cœurs (4 cœurs haute performance et 4 cœurs à haute efficacité énergétique), ainsi que 8 Go de mémoire unifiée, accompagnés d'un SSD NVMe pour un accès rapide aux données. Le système d'exploitation utilisé était macOS Sequoia, et l'ensemble des scripts a été exécuté dans un environnement Python, à l'aide de Visual Studio Code comme interface de développement. Cette configuration s'est avérée suffisante pour des expérimentations en apprentissage multitâche visant la prédiction conjointe des noms moléculaires et des formules chimiques à partir de spectres de masse, démontrant ainsi sa robustesse et son efficacité dans les flux de travail bioinformatiques et chimiométriques.

III.1.3. Logiciels et bibliothèques

Le développement du modèle d'intelligence artificielle et l'analyse des données spectrales ont été réalisés dans un environnement de programmation moderne, stable et adapté aux besoins des applications en bioinformatique, chimio-informatique et science des données. Les outils principaux utilisés sont les suivants :

- Visual Studio 2022 (version 17.10) : utilisé comme environnement de développement intégré (IDE) pour l'écriture, l'exécution et le débogage du code Python. Il intègre des fonctionnalités avancées telles que l'auto-complétion, la gestion des environnements virtuels et le support de Git, facilitant un développement structuré et reproductible.
- Python (version 3.12.4) : langage principal utilisé pour le traitement des données et la conception du modèle d'intelligence artificielle. Sa syntaxe claire et son vaste écosystème scientifique assurent une compatibilité optimale avec les bibliothèques de calcul, de modélisation et de visualisation.

Par ailleurs, plusieurs bibliothèques spécialisées ont été mobilisées pour le traitement, l'analyse et la modélisation des données, notamment :

- *Pandas* : Bibliothèque centrale pour la manipulation des données tabulaires (import/export CSV, filtrage, traitement des colonnes).
- *Numpy* : Fournit des structures de données numériques performantes (vecteurs, matrices) et des opérations mathématiques rapides, essentielles pour le traitement des spectres binned.
- *Pickle* : Sert à sérialiser/désérialiser les objets Python comme les modèles entraînés, les encodeurs de classes et les normalisateurs, afin de les réutiliser pour l'inférence.

- *Scikit-learn* (sklearn) : Bibliothèque d'apprentissage automatique fournissant des outils pour le prétraitement des données :
  - LabelEncoder : encode les noms et formules moléculaires en classes numériques.
  - StandardScaler : normalise les vecteurs spectraux (z-score) avant l'entraînement du modèle.
  - Train_test_split : divise les données en ensembles d'entraînement et de test avec stratification.
  - Shuffle : mélange les données pour éviter les biais liés à l'ordre initial.
- *Tensorflow* et *tensorflow.keras* : Plateforme principale pour le développement, la compilation et l'entraînement du modèle de deep learning :
  - Layers : construction des couches du réseau (Conv1D, ReLU, Dense, Dropout, etc.).
  - Models : assemblage du modèle CNN avec double sortie.
  - Callbacks : inclut les mécanismes de sauvegarde du meilleur modèle (ModelCheckpoint), réduction du taux d'apprentissage adaptative (ReduceLROnPlateau), et arrêt précoce (EarlyStopping).

III.2. Méthodes

III.2.1. Préparation des données

Dans l'analyse des métabolites, les données de spectrométrie de masse (MS) sont souvent stockées au format MSP (Mass Spectral Profile), qui contient des informations détaillées sur chaque molécule, y compris son nom, des informations sur le précurseur, et les rapports masse/charge (m/z) avec les intensités correspondantes. Pour faciliter l'analyse des données, en particulier lors de l'utilisation de modèles d'apprentissage automatique, il est essentiel de convertir ces données en un format structuré et facilement accessible tel que le CSV.

La conversion du format MSP en CSV est cruciale pour plusieurs raisons. Les fichiers CSV sont largement pris en charge par divers outils et logiciels d'analyse de données, rendant les données plus accessibles pour les chercheurs et les analystes. Le format CSV fournit une manière structurée de représenter les données, facilitant ainsi leur manipulation et analyse. De plus, les modèles d'apprentissage automatique et les analyses statistiques nécessitent souvent des données d'entrée sous forme tabulaire, ce que les fichiers CSV fournissent facilement. Enfin, les fichiers CSV peuvent gérer efficacement de grands ensembles de données et peuvent être traités par de nombreux langages de programmation et outils.

Nous avons utilisé un script qui s'appuie sur les bibliothèques *csv* et *tqdm* pour convertir efficacement les données MSP en format CSV, fournissant ainsi une représentation structurée et

accessible des données de spectrométrie de masse. La bibliothèque *csv* gère la lecture et l'écriture des fichiers CSV, tandis que la bibliothèque *tqdm* offre une visualisation de la progression. Le script garantit que les données sont aplaties en un seul vecteur de caractéristiques pour chaque molécule, facilitant ainsi une analyse plus approfondie et les applications d'apprentissage automatique.

III.2.2. Nettoyage des données

Après la conversion du jeu de données au format .csv, une étape de nettoyage a été entreprise afin d'assurer la qualité, la cohérence et la pertinence des données utilisées pour l'entraînement du modèle. Cette étape visait à filtrer les entrées redondantes, incomplètes ou sous-représentées.

Le fichier .csv contenant l'ensemble des spectres a d'abord été chargé à l'aide de la bibliothèque *Pandas*. Seules les colonnes pertinentes pour l'analyse ont été conservées, à savoir : le nom de la molécule (Name), la masse moléculaire (MW), la masse exacte (ExactMass), la liste des pics spectrométriques (Peaks), le nombre de pics (Num Peaks), ainsi que la formule brute (Formula).

Afin d'assurer un équilibre des classes lors de l'entraînement supervisé, seules les molécules apparaissant au moins 50 fois dans l'ensemble de données ont été retenues. Cette opération permet de minimiser le biais induit par des classes rares ou marginales.

Enfin, une vérification de cohérence a été réalisée sur l'étendue du nombre de pics enregistrés par molécule, permettant de détecter d'éventuelles anomalies. L'ensemble des données filtrées a ensuite été sauvegardé dans un nouveau fichier .csv prêt pour l'étape de vectorisation.

III.2.3. Transformation des données

Afin de rendre les classes catégorielles exploitables par les modèles d'apprentissage automatique, un encodage numérique a été appliqué aux colonnes *Name* (nom de la molécule) et *Formula* (formule moléculaire brute).

Les noms des molécules ont été encodés en étiquettes entières à l'aide de la classe *LabelEncoder* de la bibliothèque *scikit-learn*. Chaque molécule distincte s'est vue attribuer un identifiant unique sous forme d'entier stocké dans la colonne y_name. Le nombre total de classes moléculaires encodées correspond ainsi au nombre de composés uniques retenus après filtrage.

De manière analogue, les formules moléculaires ont également été encodées à l'aide d'un second encodeur *LabelEncoder,* produisant la variable y_formula représentant les formules sous forme d'entiers.

Pour garantir la reproductibilité et permettre une inférence ultérieure cohérente, les deux objets encodeurs ont été sauvegardés au format .pkl (*Pickle*) :

- label_encoder_name_name_formula_pred.pkl : encodeur des noms de molécules.
- label_encoder_formula_name_formula_pred.pkl : encodeur des formules moléculaires.

Ce prétraitement assure une compatibilité totale avec les réseaux de neurones en transformant les catégories textuelles en classes numériques exploitables pour l'entraînement supervisé.

Pour exploiter les informations contenues dans les spectres de masse, une série d'opérations de parsing et de vectorisation a été appliquée à la colonne Peaks, qui encode les couples m/z (rapport masse/charge) et intensité sous forme textuelle.

III.2.4. Mélange et séparation des données

Le mélange (*shuffling*) et la division (*splitting*) du jeu de données en ensembles d'entraînement et de test constituent une étape essentielle pour la construction de modèles d'apprentissage automatique fiables.

Pourquoi le mélange est-il important ?

Il permet d'éviter les biais liés à l'ordre des données. En effet, les données peuvent être initialement triées selon des attributs comme le nom de la molécule, la masse moléculaire ou d'autres caractéristiques. Sans mélange préalable, le modèle risque d'être entraîné sur une seule classe ou un groupe homogène, puis testé sur un ensemble totalement différent, ce qui compromettrait l'équité de l'évaluation et empêcherait toute généralisation. Le mélange assure une répartition aléatoire des classes (molécules) entre les ensembles d'entraînement et de test.

Pourquoi la division en ensembles d'entraînement/test est-elle indispensable ?

- Évaluation de la généralisation : le modèle apprend des structures à partir de l'ensemble d'entraînement. L'ensemble de test joue le rôle de données inédites, permettant d'évaluer la capacité du modèle à généraliser au-delà des exemples appris.
- Prévention du surapprentissage (*overfitting*) : l'évaluation sur les données d'entraînement peut conduire à une précision artificiellement élevée, le modèle risquant de mémoriser les données au lieu d'en apprendre les schémas. Un ensemble de test indépendant permet de détecter cette dérive.
- Reproductibilité expérimentale : diviser et conserver les mêmes ensembles d'entraînement et de test permet de reproduire les expériences, de faciliter le débogage et de comparer les performances entre différentes approches de manière rigoureuse.

III.2.5. Normalisation des données

La normalisation des données constitue une étape cruciale dans la préparation des entrées pour les modèles d'apprentissage automatique, en particulier ceux reposant sur des distances (k-Nearest Neighbors) ou des réseaux de neurones. Sans cette opération, les caractéristiques ayant des valeurs numériques élevées peuvent dominer celles de plus petite échelle, faussant ainsi les calculs de distance ou les gradients pendant l'optimisation.

Dans ce travail, les vecteurs de caractéristiques spectraux ont été standardisés selon une distribution centrée réduite (moyenne = 0, écart-type = 1) à l'aide de la méthode *StandardScaler*. L'ajustement (*fit*) a été réalisé uniquement sur l'ensemble d'apprentissage, puis appliqué à l'ensemble de test pour éviter toute fuite d'information.

Enfin, l'objet *scaler* a été sauvegardé sous format binaire (scaler_name_formula_pred.pkl) afin de garantir une normalisation identique lors des phases ultérieures d'inférence.

III.2.6. Choix du modèle

- Choix d'une architecture CNN 1D (Convolutional Neural Network One-Dimensional) idéale pour le traitement de signaux unidimensionnels tels que les données spectrales.
- Implémentation de blocs résiduels composés de deux couches *Conv1D* avec connexion de saut (skip connection), permettant un entraînement efficace des réseaux profonds en évitant le problème de gradients disparaissants.
- Conception d'un modèle à sorties multitâches avec deux branches : Une pour prédire le nom du composé, une autre pour prédire la formule chimique. Cela permet un transfert inductif en exploitant des caractéristiques partagées pour améliorer l'apprentissage.
- Utilisation de couches *Conv1D* pour détecter des motifs locaux dans des bins m/z adjacents : 32 filtres apprennent des cartes de caractéristiques, une taille de noyau de 3 capture des motifs sur trois bins consécutifs.
- Application de la normalisation par lot (*BatchNormalization*) après chaque convolution afin de stabiliser les activations et améliorer la convergence lors de l'entraînement.
- Utilisation de la fonction d'activation *ReLU* pour introduire de la non-linéarité et favoriser la parcimonie des activations.
- Ajout de blocs résiduels permettant au réseau d'apprendre des mappings identitaires, facilitant ainsi la propagation des gradients dans les couches profondes.
- Incorporation de couches *MaxPooling1D* pour sous-échantillonner les cartes de caractéristiques, réduisant la charge computationnelle et favorisant l'abstraction des représentations.

- Utilisation de la fonction *softmax* en sortie, adaptée à la classification multi-classes mutuellement exclusive, produisant une distribution de probabilités pour chaque classe.

III.2.7. Compilation et entraînement du modèle

Le modèle est construit à l'aide de la fonction *build_1dcnn* en spécifiant la longueur d'entrée, le nombre de classes pour le nom et pour la formule chimique.

La compilation du modèle utilise l'optimiseur *Adam* (Adaptive Moment Estimation), choisi pour sa capacité à adapter le taux d'apprentissage pour chaque paramètre, avec un taux initial fixé à 1e-3, considéré comme un bon point de départ.

La fonction de perte utilisée est la *sparse_categorical_crossentropy*, qui calcule la log-vraisemblance négative de la classe correcte, adaptée à la classification multi-classes avec des labels entiers. La métrique d'évaluation est l'accuracy pour les deux sorties.

Trois callbacks sont définis pour l'entraînement :

- ModelCheckpoint : sauvegarde uniquement le meilleur modèle basé sur la précision de validation pour la prédiction du nom.
- ReduceLROnPlateau : réduit le taux d'apprentissage lorsque la précision de validation stagne, permettant au modèle d'échapper aux minima locaux.
- EarlyStopping : stoppe prématurément l'entraînement si aucune amélioration n'est observée pendant un certain nombre d'époques (patience de 10), afin d'éviter le surapprentissage, tout en restaurant les poids du meilleur modèle.

L'entraînement se fait sur 500 époques, avec une taille de batch de 32 et une validation sur 10 % des données d'entraînement. Les callbacks sont activés pour optimiser le processus.

III.2.8. Évaluation du modèle

Le modèle est évalué sur l'ensemble de test, retournant la perte totale, les pertes spécifiques à chaque tâche (nom et formule), ainsi que les précisions correspondantes. Les précisions sur la prédiction du nom et de la formule sont affichées avec 4 décimales.

III.2.9. Enregistrement du modèle et des résultats

Dans cette étape, le modèle convolutif entraîné est sauvegardé sous le format HDF5 (.h5) à l'aide de la méthode *save()* de l'objet *tf.keras.Model*, ce qui permet de préserver à la fois l'architecture du réseau, les poids appris et l'état de l'optimiseur. Ensuite, l'historique de l'entraînement (pertes et précisions par époque), stocké dans l'attribut *history.history*, est enregistré dans un fichier binaire au format .pkl à l'aide du module Python *pickle*. Par ailleurs,

l'ensemble de test, incluant les caractéristiques d'entrée (sous deux formats : brut et redimensionné pour *Conv1D*) ainsi que les étiquettes de sortie (nom et formule), est sauvegardé sous deux formats : un fichier .pkl pour une réutilisation ultérieure par programme, et un fichier .csv pour une lecture humaine et une analyse descriptive. Ces opérations de sauvegarde utilisent les fonctions *to_pickle()* et *to_csv()* de la bibliothèque *pandas*.

Chapitre IV : Résultats et discussion

Chapitre IV. Résultats et discussion

IV.1. Résultats

IV.1.1. Préparation des données

À l'issue de la phase de conversion et de structuration des données issues du fichier MSP (MassBank, version NIST 2024.11), un fichier CSV exploitable a été généré, intégrant 122 512 entrées spectrales correspondant à différents profils MS. Chaque enregistrement regroupe jusqu'à 20 paramètres physico-chimiques et instrumentaux. Bien que certaines molécules soient représentées par plusieurs spectres (réplicats expérimentaux), une analyse basée sur les identifiants InChIKey révèle que ces spectres concernent 18 332 entités moléculaires distinctes (Tab. VI.1).

La table ainsi obtenue constitue un jeu de données structuré, compatible avec les outils de data science et les algorithmes d'apprentissage automatique. Les colonnes incluent notamment : le nom du composé, sa formule chimique, sa masse exacte, son spectre MS2 sous forme de couples (m/z, intensité), ainsi que des paramètres expérimentaux comme l'énergie de collision, l'instrument utilisé ou le mode d'ionisation.

Cette étape de préparation a permis d'optimiser l'intégration du jeu de données dans le pipeline de modélisation, en assurant la normalisation des entrées, la gestion des valeurs manquantes et l'uniformisation des champs structurés nécessaires à l'apprentissage supervisé.

IV.1.2. Nettoyage des données

La phase de nettoyage a permis de consolider un sous-ensemble de données de haute qualité, prêt pour l'entraînement du modèle d'apprentissage automatique. À partir des 122 512 entrées initiales, un filtrage a été appliqué pour conserver uniquement les enregistrements présentant une complétude adéquate des champs critiques : nom du composé (Name), formule moléculaire (Formula), masse moléculaire (MW), masse exacte (ExactMass), nombre de pics (Num Peaks) et liste des pics (Peaks).

L'analyse a révélé que plusieurs molécules étaient représentées à de multiples reprises, reflétant différentes conditions expérimentales. Afin de garantir une représentativité suffisante des classes dans le cadre de l'apprentissage supervisé, seules les molécules apparaissant au moins 50 fois dans le jeu de données ont été retenues. Après application de ce critère, 255 molécules distinctes ont été sélectionnées. Ces composés sont associés à un total de 15 930 entrées spectrales, qui constituent le jeu final utilisé pour l'entraînement du modèle.

Un contrôle de cohérence a également été mené sur le champ Num Peaks, permettant d'exclure les spectres comportant un nombre anormalement faible ou excessif de signaux, susceptibles d'introduire du bruit dans le modèle.

Au terme de ce processus, un nouveau fichier CSV (Tab. VI.2) contenant les données filtrées a été généré. Ce fichier constitue l'entrée du pipeline de vectorisation et servira à la génération des matrices d'apprentissage pour la prédiction du nom et de la formule chimique des métabolites à partir de leurs spectres MS.

IV.1.3. Transformation des données

Les colonnes Name et Formula ont été encodées en classes numériques à l'aide de la méthode *LabelEncoder* de *scikit*-learn. L'encodage du nom des molécules et des formules chimiques a permis d'identifier 255 classes distinctes. Les encodeurs ont été sauvegardés aux formats Pkl. Parallèlement, les spectres de la colonne Peaks, initialement sous forme textuelle, ont été transformés en vecteurs numériques normalisés, constituant la matrice d'entrée X utilisée pour l'apprentissage supervisé (Tab. VI.3).

Tableau IV.1. Vue d'ensemble des données métabolomiques préparées après la conversion des fichiers MSP en format CSV.

Name	Pyrophen	Decarestrictine F	Roquefortine A
Synon	N-[(1S)-1-(4-methoxy-6-oxopyran-2-yl)-2-phenylethyl]acetamide	(1S,3R,8Z,10R)-3-methyl-4,11-dioxabicyclo[8.1.0]undec-8-ene-5,7-dione	[(6aR,9S,10R,10aR)-7,9-dimethyl-6,6a,8,9,10,10a-hexahydro-4H-indolo[4,3-fg]quinoline-10-yl] acetate
DB#	MSBNK-AAFC-AC000854	MSBNK-AAFC-AC000767	MSBNK-AAFC-AC000552
InChIKey	VFMQMACUYWGDOJ-AWEZNQCLSA-N	MXRJZFNJVFPSQN-NQRYBKARSA-N	GJSSYQDXZLZOLR-QMHBMSAFSA-N
InChI	InChI=1S/C16H17NO4/c1-11(18)17-14(8-12-6-4-3-5-7-12)15-9-13(20-2)10-16(19)21-15/h3-7,9-10,14H,8H2,1-2H3,(H,17,18)/t14-/m0/s1	InChI=1S/C10H12O4/c1-6-4-9-8(14-9)3-2-7(11)5-10(12)13-6/h2-3,6,8-9H,4-5H2,1H3/b3-2-/t6-,8-,9+/m1/s1	InChI=1S/C18H22N2O2/c1-10-9-20(3)15-7-12-8-19-14-6-4-5-13(16(12)14)17(15)18(10)22-11(2)21/h4-6,8,10,15,17-19H,7,9H2,1-3H3/t10-,15+,17+,18+/m0/s1
SMILES	CC(=O)N[C@@H](CC1=CC=CC=C1)C2=CC(=CC(=O)O2)OC	C[C@@H]1C[C@H]2[C@H](O2)/C=C\C(=O)CC(=O)O1	C[C@H]1CN([C@@H]2CC3=CNC4=CC=CC(=C34)[C@H]2[C@@H]1OC(=O)C)C
Precursor_type	[M+H]+	[M+H]+	[M+H]+
Spectrum_type	MS2	MS2	MS2
PrecursorMZ	288.1225	197.0803	299.1749
Instrument_type	LC-ESI-ITFT	LC-ESI-ITFT	LC-ESI-ITFT
Instrument	Q-Exactive Orbitrap Thermo Scientific	Q-Exactive Orbitrap Thermo Scientific	Q-Exactive Orbitrap Thermo Scientific
Ion_mode	POSITIVE	POSITIVE	POSITIVE
Collision_energy	30(NCE)	10(NCE)	30(NCE)
Formula	$C_{16}H_{17}NO_4$	$C_{10}H_{12}O_4$	$C_{18}H_{22}N_2O_2$
MW	287	196	298
ExactMass	287.11575	196.07355	298.16813
Comments	Parent=288.1225	Parent=197.0803	Parent=299.1749
Splash	splash10-004j-1940000000-b1bd14eb30f6afd2739e	splash10-0a4i-0900000000-6881f47b296a28ed74f6	splash10-000b-0090000000-61fb8ffc68987b8be791
Num Peaks	8	4	6
Peaks	[91.0542, 245.0, 125.0233, 999.0, 154.0499, 80.0, 155.0577, 355.0, 185.0961, 349.0, 200.107, 45.0, 229.0859, 142.0, 246.1125, 734.0]	[85.0284, 54.0, 137.0597, 54.0, 155.0703, 999.0, 197.0808, 323.0]	[144.0808, 37.0, 168.0808, 32.0, 196.1121, 82.0, 208.1121, 54.0, 239.1543, 877.0, 299.1754, 999.0]

Tableau IV.2. Vue d'ensemble des données métabolomiques après le processus de nettoyage.

Name	Pyrophen	Decarestrictine F	Roquefortine A
Formula	$C_{16}H_{17}NO_4$	$C_{10}H_{12}O_4$	$C_{18}H_{22}N_2O_2$
MW	287	196	298
ExactMass	287.11575	196.07355	298.16813
Num Peaks	8	4	6
Peaks	[91.0542, 245.0, 125.0233, 999.0, 154.0499, 80.0, 155.0577, 355.0, 185.0961, 349.0, 200.107, 45.0, 229.0859, 142.0, 246.1125, 734.0]	[85.0284, 54.0, 137.0597, 54.0, 155.0703, 999.0, 197.0808, 323.0]	[144.0808, 37.0, 168.0808, 32.0, 196.1121, 82.0, 208.1121, 54.0, 239.1543, 877.0, 299.1754, 999.0]

Tableau IV.3. Vue d'ensemble des données métabolomiques après transformation et encodage.

Name	Pyrophen	Decarestrictine F	Roquefortine A
Formula	$C_{16}H_{17}NO_4$	$C_{10}H_{12}O_4$	$C_{18}H_{22}N_2O_2$
MW	287	196	298
ExactMass	287.11575	196.07355	298.16813
Num Peaks	8	4	6
Peaks	[91.0542, 245.0, 125.0233, 999.0, 154.0499, 80.0, 155.0577, 355.0, 185.0961, 349.0, 200.107, 45.0, 229.0859, 142.0, 246.1125, 734.0]	[85.0284, 54.0, 137.0597, 54.0, 155.0703, 999.0, 197.0808, 323.0]	[144.0808, 37.0, 168.0808, 32.0, 196.1121, 82.0, 208.1121, 54.0, 239.1543, 877.0, 299.1754, 999.0]
EncodedName	19645	11637	19834
EncodedFormula	2515	92	3259
ProcessedPeaksNumerical	91.0542,245.0,125.0233,999.0,154.0499,80.0,155.0577,355.0,185.0961,349.0,200.107,45.0,229.0859,142.0,246.1125,734.0	85.0284,54.0,137.0597,54.0,155.0703,999.0,197.0808,323.0	144.0808,37.0,168.0808,32.0,196.1121,82.0,208.1121,54.0,239.1543,877.0,299.1754,999.0

IV.1.4. Mélange et séparation des données

Après encodage et vectorisation, le jeu de données final, composé de 15 930 entrées, a été mélangé aléatoirement afin d'assurer une répartition homogène des classes. Ce mélange a permis d'éviter tout biais de distribution initial, notamment ceux liés à l'ordre des molécules ou de leurs caractéristiques spectrales.

Une division standard a ensuite été appliquée avec 80 % des données (12 744 échantillons) affectés à l'ensemble d'entraînement et 20 % (3 186 échantillons) à l'ensemble de test. Cette séparation a été réalisée de manière stratifiée sur la variable cible *y_name*, garantissant une représentation proportionnelle des 255 classes moléculaires dans les deux ensembles. Ce protocole assure une évaluation rigoureuse de la performance du modèle et limite les risques de surapprentissage.

IV.1.5. Normalisation des données

Les vecteurs spectraux ont été normalisés par standardisation (moyenne = 0, écart-type = 1) à l'aide de *StandardScaler*. Le scaler a été ajusté sur l'ensemble d'apprentissage (12 744 entrées) puis appliqué à l'ensemble de test (3 186 entrées), évitant toute fuite d'information. L'objet a été sauvegardé (scaler_name_formula_pred.pkl) pour une normalisation identique en phase d'inférence.

IV.1.6. Choix du modèle

Un modèle CNN 1D multitâche a été construit pour exploiter la structure locale des spectres de masse et prédire simultanément le nom et la formule des composés. L'architecture comprend :

- Une couche d'entrée adaptée aux signaux 1D vectorisés (binned spectra).
- Des couches Conv1D avec 32 filtres et noyaux de taille 3, capturant des motifs locaux dans les bins m/z.
- Deux blocs résiduels avec normalisation (*BatchNormalization*) et activation *ReLU*, facilitant l'apprentissage profond.
- Des couches *MaxPooling1D* pour réduire la dimensionnalité et extraire des caractéristiques invariantes.
- Un dense layer avec dropout (0.3) pour limiter le surapprentissage.
- Deux couches de sortie *softmax* : l'une pour prédire le nom (255 classes), l'autre pour la formule brute, permettant un apprentissage multitâche avec transfert inductif entre les deux cibles.

L'entrée a été reshaped en (échantillons, longueur, 1) pour être compatible avec *Conv1D*. Ce modèle combine profondeur, robustesse et efficacité pour la classification spectrale supervisée.

IV.1.7. Compilation et entrainement du modèle

Le modèle CNN a été compilé avec l'optimiseur Adam (learning rate = 1e-3) et la fonction de perte sparse_categorical_crossentropy pour les deux sorties. L'entraînement a été effectué sur 500 époques avec un batch size de 32, en utilisant 10 % des données d'apprentissage pour la validation.

Grâce aux callbacks activés : Le meilleur modèle (selon la précision de validation sur le nom) a été sauvegardé. Le taux d'apprentissage a été réduit automatiquement lors des plateaux de performance. L'entraînement s'est arrêté prématurément après stagnation, avec restauration des meilleurs poids. Le processus d'apprentissage s'est montré stable, avec une convergence progressive des métriques de validation, évitant à la fois le surapprentissage et l'oubli catastrophique.

IV.1.8. Evaluation du modèle

Le modèle a été évalué sur l'ensemble de test, produisant une précision de 93.31 % (Fig. VI.1) pour la prédiction du nom et de 94.76 % (Fig. IV.2) pour la formule chimique. Les pertes spécifiques à chaque tâche, ainsi que la perte globale, ont également été calculées. Ces performances reflètent la capacité du modèle à généraliser correctement sur des données inédites tout en conservant une bonne séparation des tâches.

L'analyse des courbes d'apprentissage confirme cette observation. Les courbes de perte globale (Fig. IV.3) montrent une diminution rapide de la perte pour les ensembles d'entraînement et de validation au cours des premières époques, suivie d'une stabilisation autour de la 15^{e} époque. Cette convergence indique une phase d'apprentissage efficace et stable. La légère différence entre la perte d'entraînement et celle de validation suggère un faible écart de généralisation, sans signe apparent de surapprentissage.

De plus, les courbes de perte spécifiques aux deux sorties du modèle illustrent un comportement cohérent :

Perte pour la prédiction du nom (Fig. IV.4) : Une dynamique similaire est observée. La perte de validation se stabilise autour de 0,6, et bien qu'un léger écart subsiste par rapport à la courbe d'entraînement, l'évolution est régulière et sans oscillations majeures, ce qui témoigne d'un apprentissage stable.

Perte pour la prédiction de la formule chimique (Fig. IV.5) : La perte de validation se stabilise autour de 0,65 après 15 époques, tandis que la perte d'entraînement continue de diminuer légèrement, ce qui traduit une bonne capacité de généralisation. Aucune divergence marquée n'est observée entre les deux courbes, ce qui est rassurant.

Perte totale du modèle (Fig. IV.3) : La convergence conjointe des deux tâches est bien capturée par la perte globale, qui diminue de manière soutenue au cours des premières itérations et se stabilise sans sursaut significatif, signe d'une optimisation multi-tâches réussie.

Ainsi, les courbes confirment que le modèle est robuste, qu'il parvient à équilibrer l'apprentissage des deux tâches et qu'il est bien adapté à la complexité des représentations biochimiques traitées dans ce travail.

IV.1.9. Enregistrement du modèle et des résultats

Le modèle convolutif final a été sauvegardé au format HDF5 (.h5) via *model.save()*, incluant l'architecture, les poids et l'état de l'optimiseur. L'historique d'apprentissage (pertes et précisions par époque) a été enregistré au format binaire pkl à l'aide du module pickle.

L'ensemble de test a été exporté sous deux formats pkl pour une réutilisation programmatique (données brutes, données redimensionnées, étiquettes), csv : pour une lecture humaine et une analyse descriptive. Les fonctions *to_pickle()* et *to_csv()* de pandas ont été utilisées pour ces opérations.

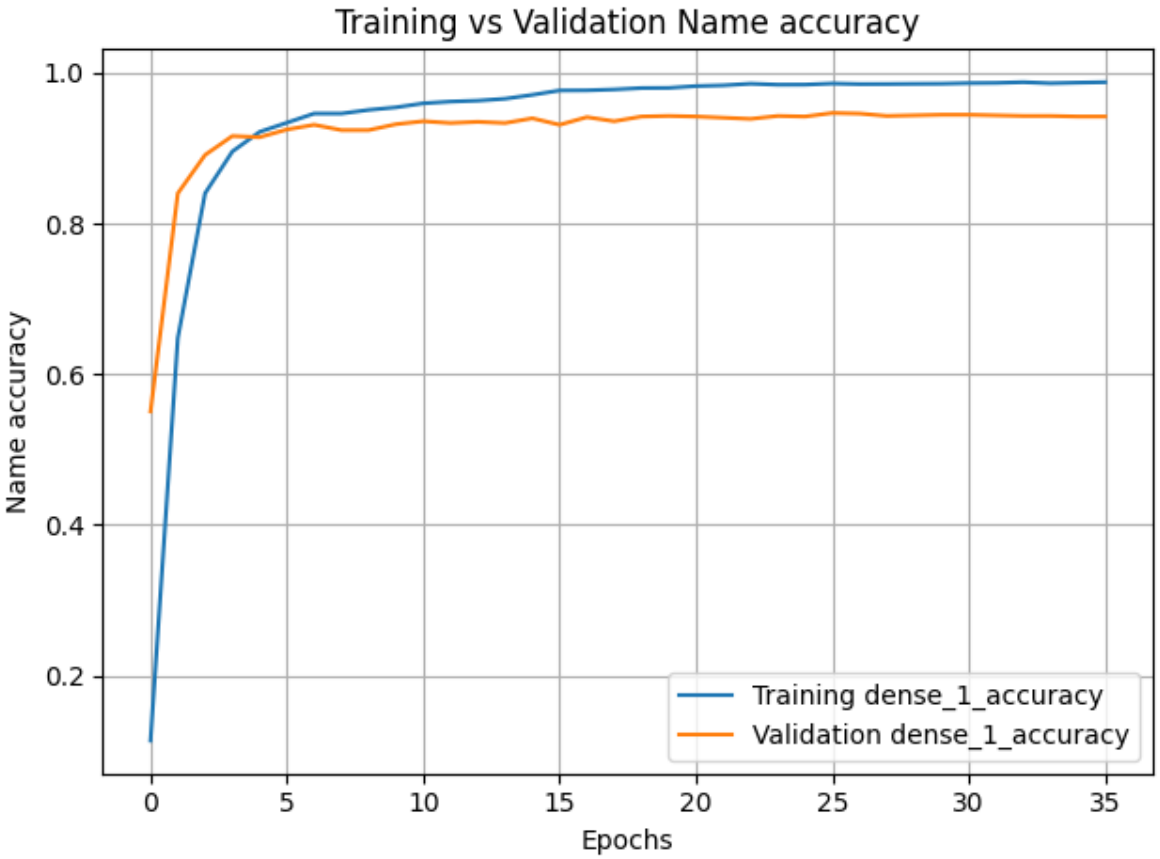

Figure IV.1. Précision (Accuracy) en entraînement et en validation pour la tâche de prédiction du nom au cours des époques.

Cette figure illustre l'évolution de la précision en entraînement et en validation pour la sortie de prédiction du nom dans un modèle neuronal à sorties multiples. Les courbes montrent une progression rapide de la précision au début de l'entraînement, suivie d'une phase de stabilisation, indiquant une convergence efficace. La précision en entraînement atteint environ 99 %, tandis que la précision en validation se stabilise autour de 94 %, suggérant un surapprentissage minime. L'écart modéré entre les deux courbes reflète une bonne capacité de généralisation du modèle sur des données non vues.

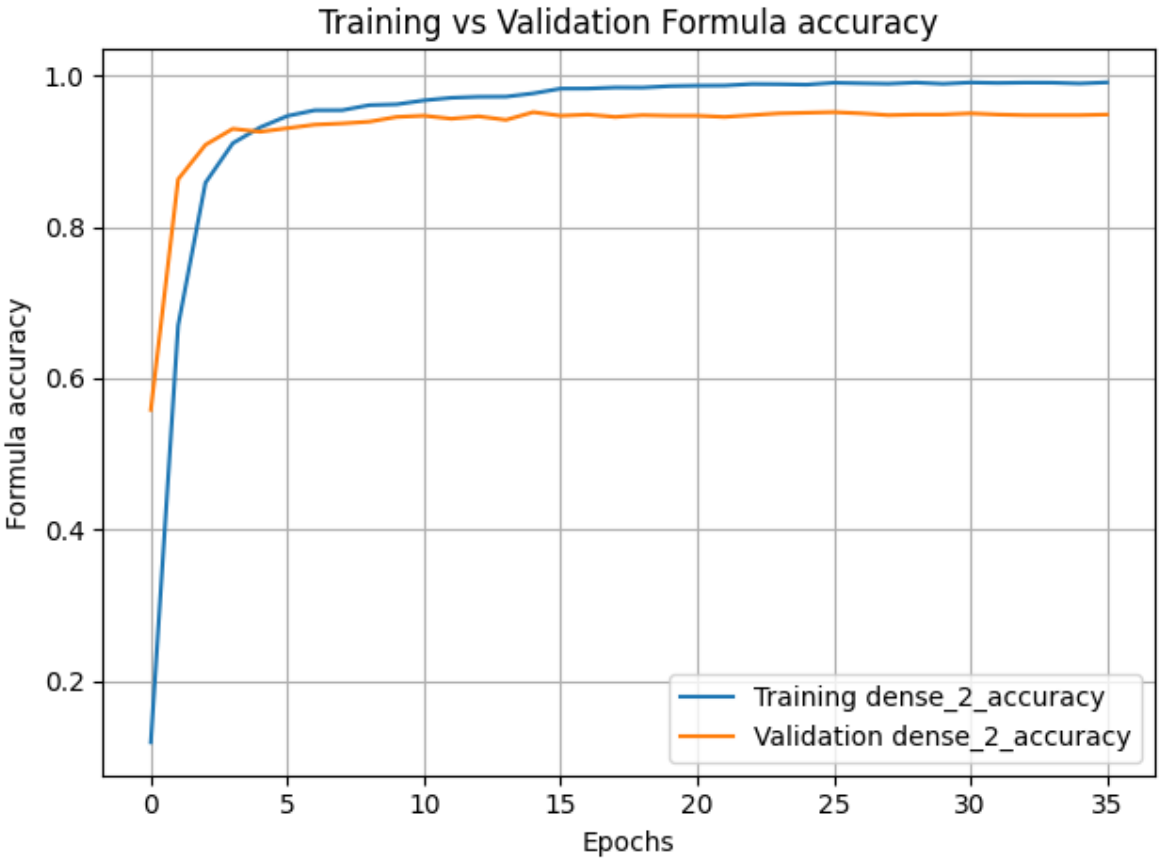

Figure IV.2. Précision (Accuracy) en entraînement et en validation pour la tâche de prédiction de la formule chimique au cours des époques.

Cette figure présente l'évolution de la précision pour la sortie de prédiction de la formule chimique dans le même modèle neuronal. À l'instar de la prédiction du nom, le modèle converge rapidement, atteignant une précision d'entraînement proche de 99 %, tandis que la précision en validation se stabilise autour de 94 %. L'évolution parallèle des courbes d'entraînement et de validation suggère un apprentissage efficace et une généralisation satisfaisante. Ces résultats soulignent la robustesse du modèle pour la prédiction simultanée de cibles structurellement liées.

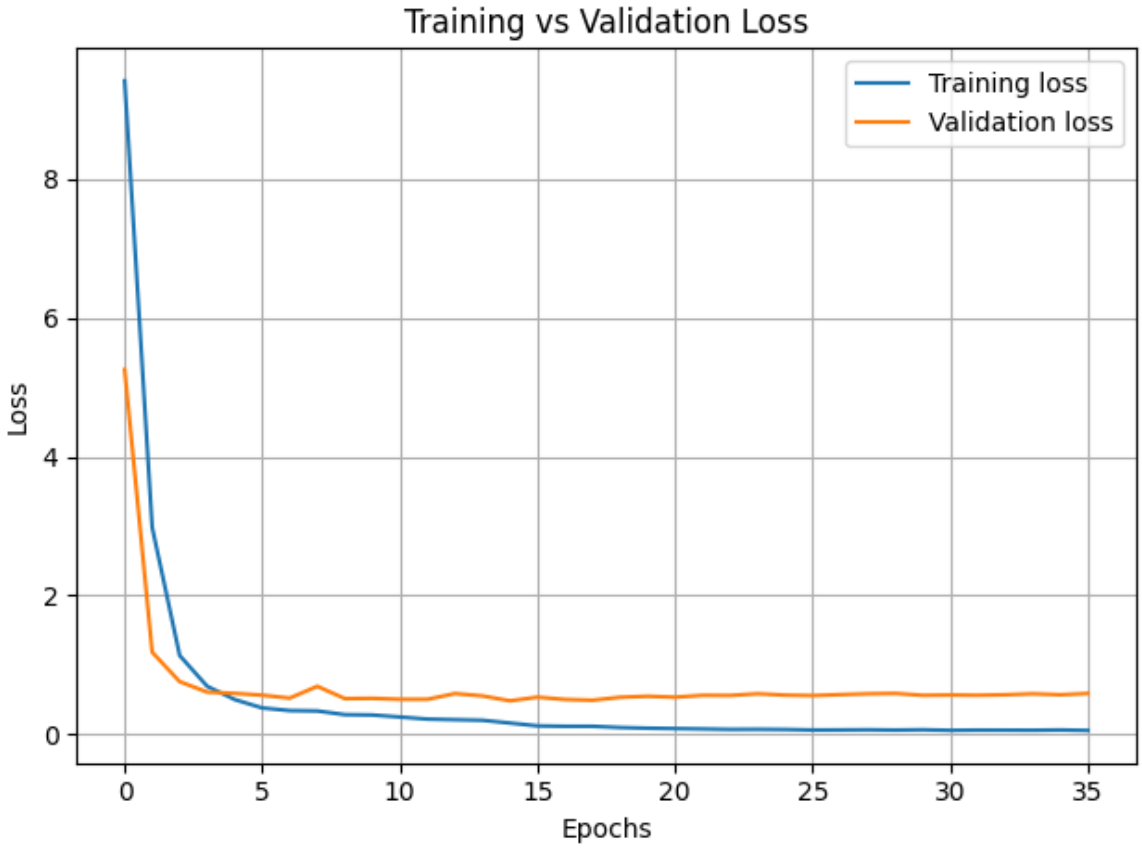

Figure IV.3. Évolution de la fonction de perte globale en entraînement et en validation au cours des époques.

Cette figure montre la dynamique de la fonction de perte globale du modèle au cours des différentes époques d'entraînement. Une diminution rapide est observée dès les premières itérations, indiquant une convergence initiale efficace. Par la suite, la perte en entraînement continue de décroître progressivement, atteignant des valeurs proches de zéro. La perte en validation se stabilise autour de 0,4 à partir de la dixième époque, suggérant un bon compromis entre apprentissage et généralisation. L'écart croissant mais modéré entre les deux courbes après stabilisation indique un surapprentissage limité, avec des performances de validation restant constantes sur les données non vues.

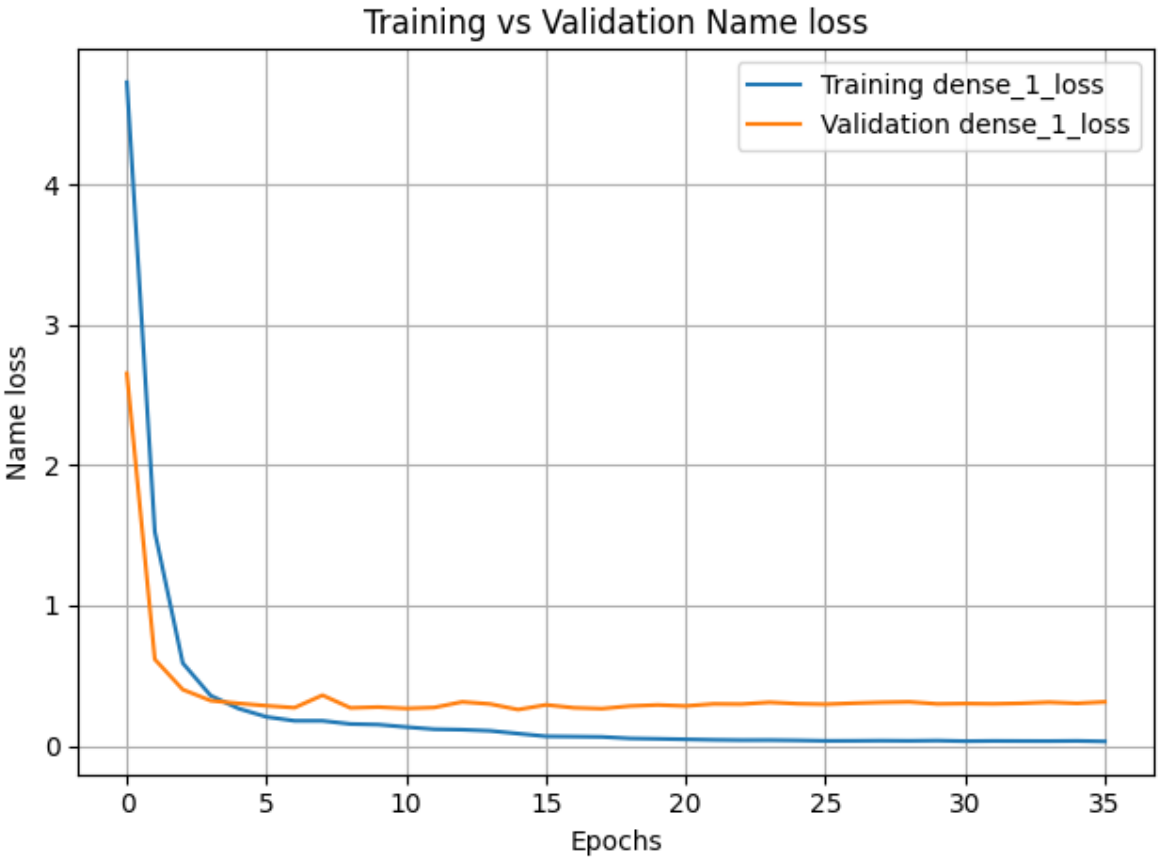

Figure IV.4. Évolution de la fonction de perte pour la prédiction du nom en entraînement et en validation.

Cette figure présente l'évolution de la fonction de perte associée à la tâche de prédiction du nom dans le modèle neuronal multi-sorties. La courbe d'apprentissage révèle une diminution rapide de la perte dès les premières itérations, indiquant une convergence initiale efficace du modèle. Par la suite, la perte en entraînement poursuit une décroissance régulière pour atteindre des valeurs proches de zéro. La perte en validation se stabilise précocement autour de 0,25, illustrant une bonne capacité de généralisation du modèle. L'écart faible et constant entre les deux courbes suggère un surapprentissage négligeable et une robustesse accrue du modèle face aux données de validation.

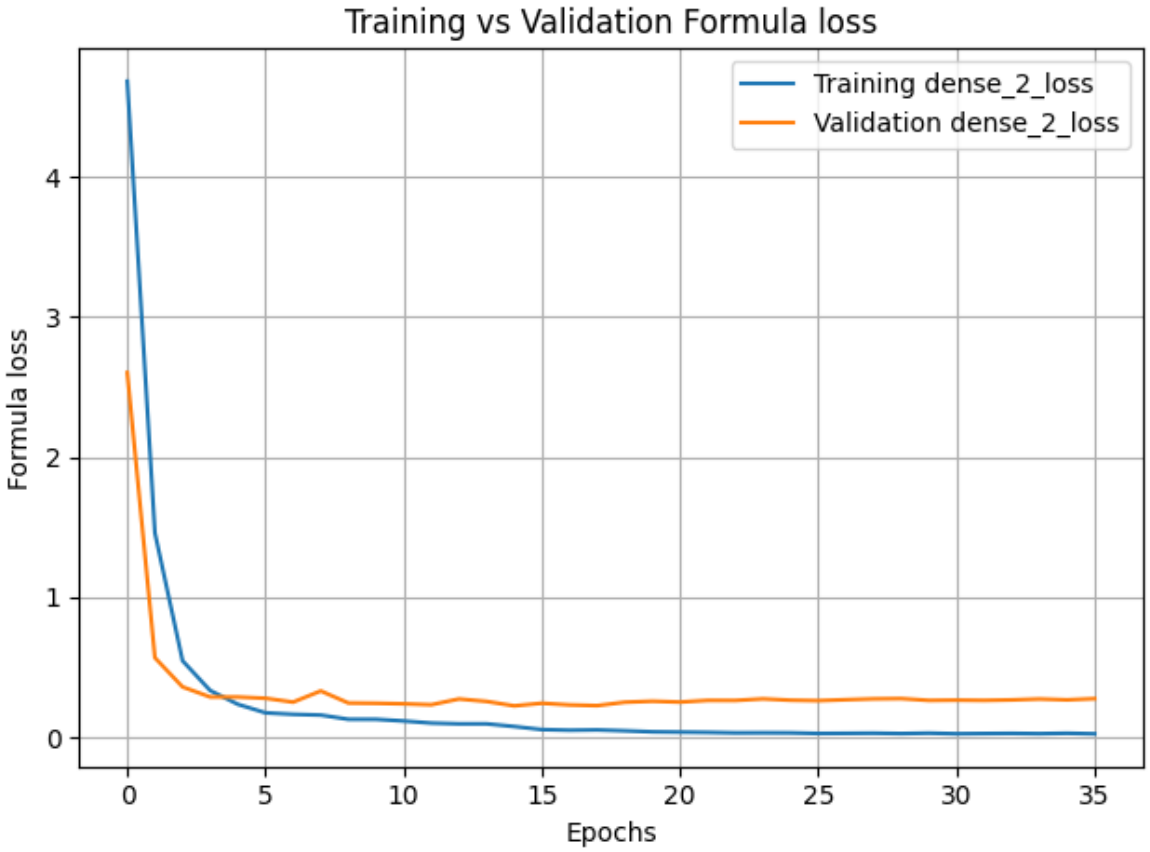

Figure IV.5. Évolution de la fonction de perte pour la prédiction de la formule chimique en entraînement et en validation.

Cette figure illustre l'évolution de la fonction de perte associée à la prédiction de la formule chimique dans le modèle neuronal multi-sorties. Une décroissance rapide de la perte est observée dès les premières itérations, traduisant une convergence efficace de l'apprentissage. Par la suite, la perte en entraînement continue de diminuer pour atteindre des valeurs proches de zéro, tandis que la perte en validation se stabilise autour de 0,2 après quelques itérations seulement. Le faible écart constant entre les courbes d'entraînement et de validation témoigne d'une généralisation robuste du modèle et d'une absence notable de surapprentissage.

IV.2. Discussion

L'étude de Pomyen *et al.* (2020) confirme que le choix d'un modèle CNN, comme celui adopté dans notre travail, est pertinent pour la métabolomique. Leur revue met en lumière les applications croissantes de l'apprentissage profond, notamment les réseaux neuronaux convolutifs, qui ont démontré un fort potentiel dans le prétraitement des données et l'identification des structures.

Cependant, Pomyen *et al.* (2020) soulignent aussi plusieurs défis à relever pour optimiser l'application de l'apprentissage profond en métabolomique, tels que le développement d'architectures spécifiques au métabolome, la gestion de la haute dimensionnalité des données, et la mise en place de protocoles robustes d'évaluation des modèles.

Bonetta *et al.* (2023) ont démontré l'utilité de l'apprentissage profond pour la prédiction multi-étiquette des classes de voies métabolomiques à partir des données KEGG, en insistant particulièrement sur l'ingénierie des caractéristiques et l'évaluation des modèles. Leur approche concorde avec notre utilisation des réseaux neuronaux convolutifs (CNN) pour l'annotation spectrale, renforçant ainsi l'efficacité de l'apprentissage profond en métabolomique. L'intégration de la prédiction des voies métabolomiques avec l'analyse spectrale pourrait par ailleurs améliorer davantage l'identification des métabolites et l'interprétation biologique.

Le travail de Li *et al.* (2019), intitulé : ‘‘Identification des métabolites à partir des spectres de masse en tandem grâce à une approche d'apprentissage automatique utilisant des caractéristiques structurales’’, valide le potentiel de l'apprentissage profond en métabolomique. Ils ont proposé une méthode d'apprentissage profond exploitant les schémas de fragmentation moléculaire pour l'identification des métabolites, atteignant une grande précision dans la prédiction structurale. Notre modèle multitâche CNN s'appuie sur cette approche en prédisant simultanément les noms des composés et leurs formules moléculaires directement à partir des spectres MS/MS, améliorant ainsi la complétude et la fiabilité de l'annotation spectrale.

Nguyen *et al.* (2021), dans leur chapitre Machine Learning for Metabolic Identification, mettent en évidence le rôle central de l'apprentissage automatique, notamment des CNNs, pour l'identification métabolique à partir de spectres MS. Leur approche, axée sur l'ingénierie des caractéristiques et les algorithmes supervisés, valide l'usage de modèles profonds adaptés. Notre architecture multitâche basée sur un CNN prolonge ces travaux en améliorant à la fois la précision de l'identification et l'interprétabilité du modèle.

La revue de Liebal *et al.* (2020) souligne l'intérêt croissant de l'apprentissage automatique supervisé pour l'analyse des spectres de masse en métabolomique, notamment pour des tâches

telles que le peak picking, la normalisation et l'imputation des données manquantes. Leur analyse confirme que des algorithmes comme les réseaux de neurones ou les forêts aléatoires peuvent améliorer la détection de biomarqueurs et l'identification de voies métaboliques. Ces conclusions soutiennent notre choix d'un modèle CNN multitâche.

La revue de Galal *et al.* (2022) met en évidence l'intérêt croissant pour l'application des méthodes d'apprentissage automatique dans l'analyse des données métabolomiques, notamment pour la classification, la régression et le regroupement de données complexes. Les auteurs soulignent que les techniques telles que les SVM, arbres de décision, forêts aléatoires, réseaux neuronaux et apprentissage profond permettent d'améliorer la modélisation des maladies et le diagnostic via un profilage métabolomique approfondi. Notre approche par CNN s'inscrit dans cette dynamique.

Chau *et al.* (2025) ont proposé MetFID, un modèle d'apprentissage profond basé sur des réseaux de neurones convolutifs (CNN) pour la prédiction des empreintes moléculaires à partir de spectres LC-MS/MS. En évaluant leur approche sur les jeux de données de référence CASMI 2016 et CASMI 2022, ils ont obtenu des scores de similarité de Tanimoto de 46 % et 20 %, respectivement, avec des scores F1 correspondants de 61 % et 32 %. Ces performances, comparables à celles de CSI:FingerID, soulignent la capacité des modèles CNN à extraire des représentations pertinentes pour l'annotation métabolomique.

IV.3. Limitations de l'étude

- Diversité limitée des données : Le modèle a été entraîné sur une seule base de données spectrales (MassBank-NIST 2024.11), ce qui peut restreindre sa capacité de généralisation à d'autres instruments, modes d'ionisation ou conditions expérimentales.
- Biais en faveur des composés fréquents : L'exclusion des molécules possédant moins de 50 spectres introduit un biais vers les composés les plus représentés, réduisant ainsi la capacité du modèle à identifier des métabolites rares ou nouveaux.
- Résolution structurale limitée : Le modèle CNN peut rencontrer des difficultés à distinguer des isomères structurellement proches à partir de spectres binned, ce qui affecte la spécificité de l'identification.
- Perte d'information due au binning : Le regroupement des spectres en intervalles (binning), bien qu'il simplifie les entrées, peut masquer des détails spectraux fins cruciaux pour une classification précise.
- Absence de données orthogonales : L'absence d'informations complémentaires telles que le temps de rétention ou l'énergie de collision réduit la fiabilité et la précision de l'annotation des composés.

Conclusion

Conclusion

Cette étude démontre l'efficacité d'un modèle multitâche 1D-CNN pour la prédiction simultanée des noms de composés et de leurs formules moléculaires à partir des spectres MS. Grâce à un prétraitement rigoureux des données et à l'utilisation de techniques avancées d'apprentissage profond, le modèle a atteint une haute précision avec un minimum de surapprentissage sur un jeu de données métabolomiques rigoureusement sélectionné. Ces résultats confirment le potentiel de l'apprentissage profond pour améliorer l'annotation spectrale et accélérer l'identification des composés. Les travaux futurs viseront à étendre l'applicabilité du modèle à des bibliothèques spectrales plus larges et diversifiées, ainsi qu'à intégrer des descripteurs moléculaires supplémentaires pour renforcer la robustesse et la généralisation des prédictions.

Perspective

L'intégration de bases de données spectrales à grande échelle telles que Wiley (950 200 composés uniques) et NIST (Electron Ionization : 347 100 ; MS/MS : 51 500) dans des cadres d'apprentissage profond ouvre des perspectives prometteuses pour l'identification automatisée des composés. La diversité chimique étendue et le volume important de spectres validés présents dans ces bibliothèques peuvent considérablement améliorer la généralisation des modèles, accroître la précision des prédictions et favoriser le développement d'outils universels d'annotation spectrale en métabolomique et chimio-informatique. Toutefois, la gestion de l'hétérogénéité des données et des exigences computationnelles demeure un défi technique majeur.

Références bibliographiques

Références bibliographiques

Aggarwal, K., Mijwil, M., Garg, S., Al-Mistarehi, A.-H., Alomari, S., Gök, M., Zein Alaabdin, A., & Abdul Rahman, S. (2022). Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning. *Iraqi Journal for Computer Science and Mathematics*, *3*, 115–123. https://doi.org/10.52866/ijcsm.2022.01.01.013

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *7*, e7702. https://doi.org/10.7717/peerj.7702

Alharbi, H., Sampedro, G. A., Juanatas, R. A., & Lim, S.-j. (2024). Enhanced skin cancer diagnosis: a deep feature extraction-based framework for the multi-classification of skin cancer utilizing dermoscopy images [Original Research]. *Frontiers in Medicine*, *Volume 11 - 2024*. https://doi.org/10.3389/fmed.2024.1495576

Ali, A. (2022). High-Performance Liquid Chromatography (HPLC): A review. *Annals of Advances in Chemistry*, *6*, 010–020. https://doi.org/10.29328/journal.aac.1001026

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., C. Ewald, J., Fraser, P. D., Giavalisco, P., Hall, R. D., Heinemann, M., Link, H., Luo, J., Neumann, S., Nielsen, J., Perez de Souza, L., Saito, K., Sauer, U., Schroeder, F. C.,…Fernie, A. R. (2021). Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature Methods*, *18*(7), 747–756. https://doi.org/10.1038/s41592-021-01197-1

Auvil, N. C., & Bier, M. E. (2024). Nanoelectrode Atmospheric Pressure Chemical Ionization Mass Spectrometry. *Journal of the American Society for Mass Spectrometry*, *35*(10), 2288–2296. https://doi.org/10.1021/jasms.4c00117

Badmus, O., Rajput, S. A., Arogundade, J. B., & Williams, M. (2024). AI-driven business analytics and decision making. *World Journal of Advanced Research and Reviews*, *24*(1), 616–633.

Bahi, M., & Batouche, M. (2018). Drug-Target Interaction Prediction in Drug Repositioning Based on Deep Semi-Supervised Learning. In A. Amine, M. Mouhoub, O. Ait Mohamed, & B. Djebbar, *Computational Intelligence and Its Applications* Cham.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Corrected at 8th printing 2009 ed.). Springer Science + Business Media, New York, 738 p.

Blaženović, I., Kind, T., Ji, J., & Fiehn, O. (2018). Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites*, *8*(2). https://doi.org/10.3390/metabo8020031

Blum, F. (2014). High performance liquid chromatography. *British Journal of Hospital Medicine*, *75*(Sup2), C18–C21. https://doi.org/10.12968/hmed.2014.75.Sup2.C18

Bohr, A., & Memarzadeh, K. (2020). Chapter 2 - The rise of artificial intelligence in healthcare applications. In A. Bohr & K. Memarzadeh (Eds.), *Artificial Intelligence in Healthcare* (pp. 25–60). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-818438-7.00002-2

Bonetta, V., Rosalin, Ebejer, J.-P., & Valentino, G. (2023). Machine Learning Using Neural Networks for Metabolomic Pathway Analyses. In K. Selvarajoo (Ed.), *Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology* (pp. 395–415). Springer US. https://doi.org/10.1007/978-1-0716-2617-7_17

Bunod, R., Augstburger, E., Brasnu, E., Labbe, A., & Baudouin, C. (2022). Intelligence artificielle et glaucome : une revue de la littérature. *Journal Français d'Ophtalmologie*, *45*(2), 216–232. https://doi.org/https://doi.org/10.1016/j.jfo.2021.11.002

Burlingame, A. L., Boyd, R. K., & Gaskell, S. J. (1996). Mass Spectrometry. *Analytical Chemistry*, *68*(12), 599–652. https://doi.org/10.1021/a1960021u

Capellades, J., Junza, A., Samino, S., Brunner, J. S., Schabbauer, G., Vinaixa, M., & Yanes, O. (2021). Exploring the Use of Gas Chromatography Coupled to Chemical Ionization Mass Spectrometry (GC-CI-MS) for Stable Isotope Labeling in Metabolomics. *Analytical Chemistry*, *93*(3), 1242–1248. https://doi.org/10.1021/acs.analchem.0c02998

Carr, S. A., Abbatiello, S. E., Ackermann, B. L., Borchers, C., Domon, B., Deutsch, E. W., Grant, R. P., Hoofnagle, A. N., Hüttenhain, R., Koomen, J. M., Liebler, D. C., Liu, T., MacLean, B., Mani, D. R., Mansfield, E., Neubert, H., Paulovich, A. G., Reiter, L., Vitek, O.,…Weintraub, S. (2014). Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-

Purpose Approach *. *Molecular & Cellular Proteomics*, *13*(3), 907–917. https://doi.org/10.1074/mcp.M113.036095

Chang, W. C., Huang, L. C. L., Wang, Y.-S., Peng, W.-P., Chang, H. C., Hsu, N. Y., Yang, W. B., & Chen, C. H. (2007). Matrix-assisted laser desorption/ionization (MALDI) mechanism revisited. *Analytica Chimica Acta*, *582*(1), 1–9. https://doi.org/https://doi.org/10.1016/j.aca.2006.08.062

Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C. J., Kadoury, S., & Tang, A. (2017). Deep Learning: A Primer for Radiologists. *Radiographics*, *37*(7), 2113–2131. https://doi.org/10.1148/rg.2017170077

Chau, H. Y. K., Zhang, X., & Ressom, H. W. (2025). Deep Learning-Based Molecular Fingerprint Prediction for Metabolite Annotation. *Metabolites*, *15*(2). https://doi.org/10.3390/metabo15020132

Chernushevich, I. V., Loboda, A. V., & Thomson, B. A. (2001). An introduction to quadrupole–time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, *36*(8), 849–865. https://doi.org/https://doi.org/10.1002/jms.207

Chipuk, J. E., & Brodbelt, J. S. (2008). Transmission mode desorption electrospray ionization. *Journal of the American Society for Mass Spectrometry*, *19*(11), 1612–1620. https://doi.org/10.1016/j.jasms.2008.07.002

Choudhary, L., & Choudhary, J. S. (2024). Deep Learning Meets Machine Learning: A Synergistic Approach towards Artificial Intelligence. *Journal of Scientific Research and Reports*, *30*(11), 865–875. https://doi.org/10.9734/jsrr/2024/v30i112614

Colliot, O. (2023). *Machine Learning for Brain Disorders* (C. Olivier, Ed. Vol. 197). Springer. https://doi.org/10.1007/978-1-0716-3195-9

Cordero, J., Menkovski, V., & Allmer, J. (2020). Detection of pre-microRNA with Convolutional Neural Networks. *bioRxiv*, 840579. https://doi.org/10.1101/840579

Crozier, A., Yokota, T., Jaganath, I. B., Marks, S., Saltmarsh, M., & Clifford, M. N. (2006). Secondary Metabolites in Fruits, Vegetables, Beverages and Other Plant-based Dietary Components. In *Plant Secondary Metabolites* (pp. 208–302). https://doi.org/https://doi.org/10.1002/9780470988558.ch7

Da Silva, R. R., Dorrestein, P. C., & Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A*, *112*(41), 12549–12550. https://doi.org/10.1073/pnas.1516878112

Datta, S., Li, Y., Ruppert, M. M., Ren, Y., Shickel, B., Ozrazgat-Baslanti, T., Rashidi, P., & Bihorac, A. (2021). Reinforcement learning in surgery. *Surgery*, *170*(1), 329–332. https://doi.org/10.1016/j.surg.2020.11.040

Degano, I. (2019). Liquid chromatography: Current applications in Heritage Science and recent developments. *Physical Sciences Reviews*, *4*(5). https://doi.org/doi:10.1515/psr-2018-0009

Domon, B., & Aebersold, R. (2006). Mass Spectrometry and Protein Analysis. *Science*, *312*(5771), 212–217. https://doi.org/doi:10.1126/science.1124619

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

Dührkop, K., Nothias, L. F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., Petras, D., Gerwick, W. H., Rousu, J., Dorrestein, P. C., & Böcker, S. (2021). Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*, *39*(4), 462–471. https://doi.org/10.1038/s41587-020-0740-8

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *Radiographics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

Faktor, J., Dvorakova, M., Maryas, J., Struharova, I., & Bouchal, P. (2012). Identification and characterisation of pro-metastatic targets, pathways and molecular complexes using a toolbox of proteomic technologies. *Klin Onkol*, *25 Suppl 2*, 2s70–77.

Fernie, A. R., & Pichersky, E. (2015). Focus Issue on Metabolism: Metabolites, Metabolites Everywhere. *Plant Physiol*, *169*(3), 1421–1423. https://doi.org/10.1104/pp.15.01499

Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. *Front Genet*, *13*, 1017340. https://doi.org/10.3389/fgene.2022.1017340

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hameed, K., Khan, M. S., Fatima, A., Shah, S. M., & Abdullah, M. A. (2023). Exploring the Word of Thin-Layer Chromatography: A Review. *Asian Journal of Applied Chemistry Research*, *14*(3), 23–38. https://doi.org/10.9734/ajacr/2023/v14i3268

Harwood, L. M., & Claridge, T. D. (1997). Introduction to organic spectroscopy. *1st publ., Oxford University Press*, 91 p. https://doi.org/http://dl.iranchembook.ir/ebook/organic-chemistry-2747.pdf

Helfer, A. G., Michely, J. A., Weber, A. A., Meyer, M. R., & Maurer, H. H. (2015). Orbitrap technology for comprehensive metabolite-based liquid chromatographic–high resolution-tandem mass spectrometric urine drug screening – Exemplified for cardiovascular drugs. *Analytica Chimica Acta*, *891*, 221–233. https://doi.org/https://doi.org/10.1016/j.aca.2015.08.018

Holley, K., Pennington, M., & Phillips, P. (1995). Gas chromatography in food analysis: an introduction. *Nutrition & Food Science*, *95*(5), 10–12. https://doi.org/10.1108/00346659510093973

Howard, A., Zhang, C., & Horvitz, E. (2017, 8–10 March 2017). Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO),

Ikonomou, M. G., & Rayne, S. (2002). Chromatographic and Ionization Properties of Polybrominated Diphenyl Ethers Using GC/High-Resolution MS with Metastable Atom Bombardment and Electron Impact Ionization. *Analytical Chemistry*, *74*(20), 5263–5272. https://doi.org/10.1021/ac020191j

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666. https://doi.org/https://doi.org/10.1016/j.eswa.2023.122666

Khalifea, H., & Ali, N. (2025). Exploring the Principles of GC-MS: Techniques and Applications. *Physical Sciences, Life Science and Engineering*, *2*(3), 10. https://doi.org/10.47134/pslse.v2i3.388

Korfmacher, W. A. (2005). Foundation review: Principles and applications of LC-MS in new drug discovery. *Drug Discovery Today*, *10*(20), 1357–1367. https://doi.org/https://doi.org/10.1016/S1359-6446(05)03620-2

Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification Techniques* Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies,

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kuril, A. K. (2024). Navigating mass spectrometry: a comprehensive guide to basic concepts and techniques. *Available at SSRN 4879107*.

Li, Y., Kuhn, M., Gavin, A.-C., & Bork, P. (2019). Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. *Bioinformatics*, *36*(4), 1213–1218. https://doi.org/10.1093/bioinformatics/btz736

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., & Blank, L. M. (2020). Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites*, *10*(6). https://doi.org/10.3390/metabo10060243

Liebler, D. C., & Zimmerman, L. J. (2013). Targeted Quantitation of Proteins by Mass Spectrometry. *Biochemistry*, *52*(22), 3797–3806. https://doi.org/10.1021/bi400110b

Lindley, S. E., Lu, Y., & Shukla, D. (2024). The Experimentalist's Guide to Machine Learning for Small Molecule Design. *ACS Applied Bio Materials*, *7*(2), 657–684. https://doi.org/10.1021/acsabm.3c00054

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/https://doi.org/10.1016/j.media.2017.07.005

Masucci, J. A., & Caldwell, G. W. (2004). Techniques for Gas Chromatography/Mass Spectrometry. In *Modern Practice of Gas Chromatography* (pp. 339–401). https://doi.org/https://doi.org/10.1002/0471651141.ch7

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), 517.

Naeem, M., Rizvi, S. T. H., & Coronato, A. (2020). A Gentle Introduction to Reinforcement Learning and its Application in Different Fields. *IEEE Access*, *8*, 209320–209344. https://doi.org/10.1109/ACCESS.2020.3038605

Nguyen, D. H., Nguyen, C. H., & Mamitsuka, H. (2019). Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief Bioinform*, *20*(6), 2028–2043. https://doi.org/10.1093/bib/bby066

Nguyen, D. H., Nguyen, C. H., & Mamitsuka, H. (2021). Machine Learning for Metabolic Identification. In K. Nishimura, M. Murase, & K. Yoshimura (Eds.), *Creative Complex Systems* (pp. 329–350). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-4457-3_20

Nolting, D., Malek, R., & Makarov, A. (2019). Ion traps in modern mass spectrometry. *Mass Spectrometry Reviews*, *38*(2), 150–168. https://doi.org/https://doi.org/10.1002/mas.21549

Patel, M. (2018). Review Article: Chromatography Principle and Applications. *Ijppr*, *13*(4). https://doi.org/https://ijppr.humanjournals.com/wp-content/uploads/2018/12/26.Mimansha-Patel.pdf

Patel, M. K., Kumar, M., Li, W., Luo, Y., Burritt, D. J., Alkan, N., & Tran, L. P. (2020). Enhancing Salt Tolerance of Plants: From Metabolic Reprogramming to Exogenous Chemical Treatments and Molecular Approaches. *Cells*, *9*(11). https://doi.org/10.3390/cells9112492

Pesapane, F., Codari, M., & Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, *2*(1), 35. https://doi.org/10.1186/s41747-018-0061-6

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., & Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. *Computational and Structural Biotechnology Journal*, *18*, 2818–2825. https://doi.org/https://doi.org/10.1016/j.csbj.2020.09.033

Prakash, C., Kumar, R., & Mittal, N. (2018). Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. *Artificial Intelligence Review*, *49*(1), 1–40. https://doi.org/10.1007/s10462-016-9514-6

Rahman, A., Debnath, T., Kundu, D., Khan, M. S. I., Aishi, A. A., Sazzad, S., Sayduzzaman, M., & Band, S. S. (2024). Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health*, *11*(1), 58–109. https://doi.org/10.3934/publichealth.2024004

Rajawat, J., & Jhingan, G. (2019). Chapter 1 - Mass spectroscopy. In G. Misra (Ed.), *Data Processing Handbook for Complex Biological Data Sources* (pp. 1–20). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-816548-5.00001-0

Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: a brief review. *International journal of engineering and technology*, *7*, 81.

Richer, J., Spencer, J., & Baird, M. (2006). Identification of Glue Vapors Using Electron Impact and Chemical Ionization Modes in GC–MS. *Journal of Chemical Education*, *83*(8), 1196. https://doi.org/10.1021/ed083p1196

Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. *J Cheminform*, *8*, 3. https://doi.org/10.1186/s13321-016-0115-9

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003

Strathmann, F. G., & Hoofnagle, A. N. (2011). Current and Future Applications of Mass Spectrometry to the Clinical Laboratory. *American Journal of Clinical Pathology*, *136*(4), 609–616. https://doi.org/10.1309/ajcpw0ta8obbngck

Taylor, L. T. (2009). Supercritical fluid chromatography for the 21st century. *The Journal of Supercritical Fluids*, *47*(3), 566–573. https://doi.org/https://doi.org/10.1016/j.supflu.2008.09.012

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Vidova, V., & Spacil, Z. (2017). A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Analytica Chimica Acta*, *964*, 7–23. https://doi.org/https://doi.org/10.1016/j.aca.2017.01.059

Vij, I., & Pathania, A. (2023). An Overview- Advances in Chromatographic Techniques in Phytochemistry. *E3S Web Conf.*, *391*, 01038.

Wellen, K. E., & Thompson, C. B. (2012). A two-way street: reciprocal regulation of metabolism and signalling. *Nat Rev Mol Cell Biol*, *13*(4), 270–276. https://doi.org/10.1038/nrm3305

Woschank, M., Rauch, E., & Zsifkovits, H. (2020). A Review of Further Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics. *Sustainability*, *12*(9), 3760.

Zaraté, P. (2021). L'intelligence artificielle d'hier à aujourd'hui.

Zhou, Y. (2024). Advances in Metabolomics based on Mass Spectrometry. *Theoretical and Natural Science*, *66*, 50–54.