INERIS
INSTITUT NATIONAL DE L'ENVIRONNEMENT INDUSTRIEL ET DES RISQUES

RAPPORT DE STAGE DE SPÉCIALITÉ
Département Informatique et Technologies de l'Information

APPLICATION DU DEEP LEARNING À LA SPECTROMÉTRIE DE MASSE

Stagiaire Hugo Tondenier

Tuteur Jean-Yves Chatelier
INERIS

Période 19 Mai — 14 Août 2025 : 12 semaines

2025

Sommaire du rapport

1 Remerciements

Je tiens à exprimer ma sincère gratitude à toutes les personnes qui ont contribué à la réussite de ce stage au sein de l'INERIS.

En premier lieu, je remercie chaleureusement mon tuteur de stage, M. Jean-Yves Chatelier, pour son encadrement, sa disponibilité constante, son investissement dans ce stage et ses conseils qui ont guidé mes travaux tout au long de ces trois mois. Sa vision très large et très à jour sur les techniques et modèles d'intelligence artificielle m'a permis de beaucoup apprendre pendant ce stage.

Je tiens aussi à remercier Peter Lin actuellement alternant à l'INERIS, qui a travaillé sur le projet "Empreinte environnementale" avec moi.

Mes remerciements s'adressent également à l'équipe ANAE notamment Azziz Assoumani et Nina Huynh pour leur accueil bienveillant et leur soutien technique, particulièrement lors de mes premiers pas dans la spectrométrie de masse, car ce domaine m'était inconnu et leurs efforts pédagogiques m'ont permis d'avancer et de comprendre le problème et les manières de le résoudre.

Enfin, ma reconnaissance va à l'INERIS pour m'avoir offert l'opportunité de contribuer à des recherches au service de la protection environnementale qui est un domaine passionnant et concret dans lequel il est très gratifiant de pouvoir appliquer mes connaissances théoriques et pratiques de Deep Learning.

2 Présentation de l'entreprise

Ce stage a été réalisé au sein de l'Institut National de l'Environnement Industriel et des Risques (INERIS). Pour appréhender la portée des travaux effectués, il est essentiel de comprendre la mission, l'organisation et les enjeux scientifiques de cet organisme.

Mission et statut de l'INERIS

L'INERIS est un Établissement Public à caractère Industriel et Commercial (EPIC) créé en 1990, placé sous la tutelle du Ministère de la Transition Écologique et de la Cohésion des Territoires. Sa mission principale est de développer et de mettre à disposition son expertise scientifique et technique pour la maîtrise des risques que les activités économiques, les substances chimiques et les procédés industriels peuvent engendrer pour la santé, la sécurité des personnes et des biens, ainsi que pour l'environnement.

L'Institut joue un rôle central d'expert public, apportant son appui aux autorités pour la définition des politiques de prévention, ainsi qu'aux entreprises pour les aider à répondre à leurs obligations réglementaires. Cette mission s'appuie sur une activité de recherche ambitieuse visant à toujours améliorer la compréhension des phénomènes complexes liés aux risques.

Domaines d'expertise et organisation interne

Pour mener à bien ses missions, l'INERIS couvre un large spectre de domaines d'expertise, incluant les risques accidentels (explosions, incendies), les risques chroniques (pollutions de l'air, de l'eau, des sols), la sécurité des produits et des équipements, ainsi que les risques liés au sous-sol.

Les valeurs au sein de l'INERIS sont les suivantes :

- **Sens du collectif** : Travailler ensemble dans un même but, l'excellence, en plaçant l'action individuelle dans la dynamique collective.
- **Intégrité** : La position d'expert reconnu et responsable de l'INERIS repose sur son indépendance de jugement et sur l'équité dans la conduite de ses missions, quel que soit le donneur d'ordre.
- **Ouverture** : Écouter pour comprendre la culture et les attentes, s'ouvrir aux autres et enrichir sa réflexion par la différence et la diversité.
- **Exigence** : L'exigence de chacun est le socle du professionnalisme et de la qualité des travaux de l'Institut. Elle nourrit l'image de l'INERIS et contribue à sa reconnaissance.

Missions et objectifs stratégiques

Dans son contrat d'objectifs et de performance pour 2021-2025 (contrat d'objectifs et de performance conclu entre l'État – ministère chargé de l'Environnement – et l'Institut national de l'environnement industriel et des risques), l'INERIS détaille ses orientations stratégiques dans le but de sécuriser la transition écologique et le renouveau de l'industrie :

- Maîtriser les risques liés à la transition énergétique et à l'économie circulaire
- Comprendre et maîtriser les risques à l'échelle d'un site industriel et d'un territoire
- Caractériser les dangers des substances et leurs impacts sur l'homme et la biodiversité
- Veille, ouverture & déontologie
- Renforcer le pilotage stratégique de l'institut et les synergies entre les activités de services aux entreprises, d'appui et de recherche

Engagement en développement durable et responsabilité sociétale

Si la prévention des risques n'est pas, en elle-même, un objectif de développement durable, elle en est un prérequis. Par sa mission, l'INERIS contribue plus particulièrement à l'atteinte des cibles suivantes :

- Réduire nettement le nombre de décès et de maladies dus à des substances chimiques dangereuses, à la pollution et à la contamination de l'air, de l'eau et du sol.
- Améliorer la qualité de l'eau en réduisant la pollution, en éliminant l'immersion de déchets et en réduisant au minimum les émissions de produits chimiques et de matières dangereuses et en augmentant considérablement à l'échelle mondiale le recyclage et la réutilisation sans danger de l'eau.
- Moderniser l'infrastructure et adapter les industries afin de les rendre durables, par une utilisation plus rationnelle des ressources et un recours accru aux technologies et procédés industriels propres et respectueux de l'environnement.
- Réduire considérablement le nombre de personnes tuées et le nombre de personnes touchées par les catastrophes.
- Incorporer des mesures relatives aux changements climatiques dans les politiques, les stratégies et la planification nationales.

Par son engagement dans une démarche de responsabilité sociétale et environnementale, l'INERIS participe en outre, à son niveau, à l'atteinte de ces objectifs dans ses modes de fonctionnement quotidiens.

3 Présentation du stage

Le stage s'est déroulé au sein du département des systèmes d'information (DSI), car c'est dans ce département que mon tuteur travaille. Pour autant, le projet « Empreinte Environnementale » concernait d'autres équipes comme celle du laboratoire de méthodes et développement en analyses pour l'environnement (ANAE), qui est l'expert métier dans le domaine de la spectrométrie de masse à l'INERIS. Elle rassemble les compétences et les moyens analytiques de pointe pour l'identification et la quantification de substances dans des matrices environnementales complexes. La spectrométrie de masse couplée à la chromatographie liquide (LC-MS) ou couplée à la chromatographie gazeuse (GC-MS) y sont des outils de travail quotidiens et essentiels, produisant les données spectrales qui sont au cœur de ce projet et donc du stage.

3.1 Problématique

L'une des missions du laboratoire ANAE est d'analyser des prélèvements d'eau de rivière pour déterminer la présence de polluants. Cette mission est essentielle, elle permet par exemple, de donner des indications aux stations d'épuration en analysant les échantillons d'eau en amont et en aval de la station, de faire évoluer les normes et la législation si les seuils de certains polluants sont trop hauts ou que de nouveaux polluants sont découverts. Pour réaliser cette analyse, ANAE utilise notamment la chromatographie en phase liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS).

Ces deux techniques couplées permettent de générer une quantité considérable d'informations sur la composition d'un échantillon. Cependant, l'interprétation de ces données représente une tâche complexe et extrêmement chronophage. Comme le soulignait l'offre de stage, si les appareils détectent un nombre croissant de signaux, la capacité à les caractériser humainement reste limitée par l'ampleur de la tâche.

Face à ce défi, l'exploration d'approches innovantes est devenue une nécessité. C'est dans ce contexte que le concept de modèle de fondation devient intéressant. Inspirés par les succès des grands modèles de langue en traitement du langage naturel (NLP), ces modèles pré-entraînés sur de vastes quantités de données non annotées (ici ce seraient des millions de spectres) peuvent ensuite être adaptés à des tâches spécifiques avec un effort et des données annotées moindres. Ils ne cherchent pas à répondre à une question, mais à apprendre la "grammaire" fondamentale des données. Comme les tâches à effectuer sur des données spectrométriques sont très nombreuses, le concept s'applique ici parfaitement.

L'enjeu de ce stage est donc de déterminer dans quelle mesure cette nouvelle approche peut répondre à la problématique de l'INERIS et si possible d'en faire une preuve de concept.

Question de recherche

Comment l'apprentissage profond, et en particulier l'approche par modèle de fondation, peut-il permettre d'automatiser et de fiabiliser la caractérisation de substances chimiques à partir de leurs spectres de masse, pour accélérer les missions de surveillance environnementale de l'INERIS ?

3.2 Objectifs du stage

Pour répondre à cette problématique, le stage a été structuré autour de plusieurs objectifs clairs, combinant recherche théorique, analyse de données et développement pratique. La finalité était de faire une preuve de concept d'un outil de Deep Learning performant et documenté, dont les développements pourraient être repris et poursuivis par l'INERIS s'il s'avérait que le concept était pertinent.

Les objectifs principaux étaient les suivants :

→ **Étude bibliographique approfondie** : Réaliser une analyse de l'état de l'art des modèles d'intelligence artificielle appliqués à la spectrométrie de masse, en se concentrant sur les architectures de type Transformer et les modèles de fondation existants.

→ **Proposition d'un modèle de fondation** : Mettre en œuvre et documenter un premier prototype, soit en adaptant un modèle open-source pertinent, soit en en définissant un "ex nihilo", et l'entraîner sur une base de données publique de référence telle que MassBank EU (Horai et al. 2010 et Oberacher et al. 2020), GNPS (Wang et al. 2016), puis l'adapter aux données spécifiques de l'INERIS.

→ **Évaluation et itération** : Analyser les succès et les échecs des différentes solutions testées afin de fournir des directions claires pour les travaux futurs au sein de l'Institut.

3.3 Annonce du plan

Afin de présenter de manière claire et structurée la démarche adoptée et les résultats obtenus, ce rapport s'articulera en trois grandes parties.

La **première partie** sera consacrée au contexte scientifique et technologique. Elle permettra d'introduire les principes fondamentaux de la spectrométrie de masse ainsi que les données qu'elle génère.

Le cœur de ce rapport, les **deuxième** et **troisième parties**, détailleront en profondeur la démarche expérimentale et les travaux menés. Tout d'abord, l'environnement technique mis en place, l'acquisition et le pré-traitement des données seront présentés, après cela, les différentes pistes envisagées seront expliquées. Ces parties retraceront nos explorations tout au long du stage.

Pour plus de clarté et pour éviter les nombreux allers-retours entre les différentes parties, le travail sera présenté sans suivre l'ordre chronologique sur la frise figure 1.

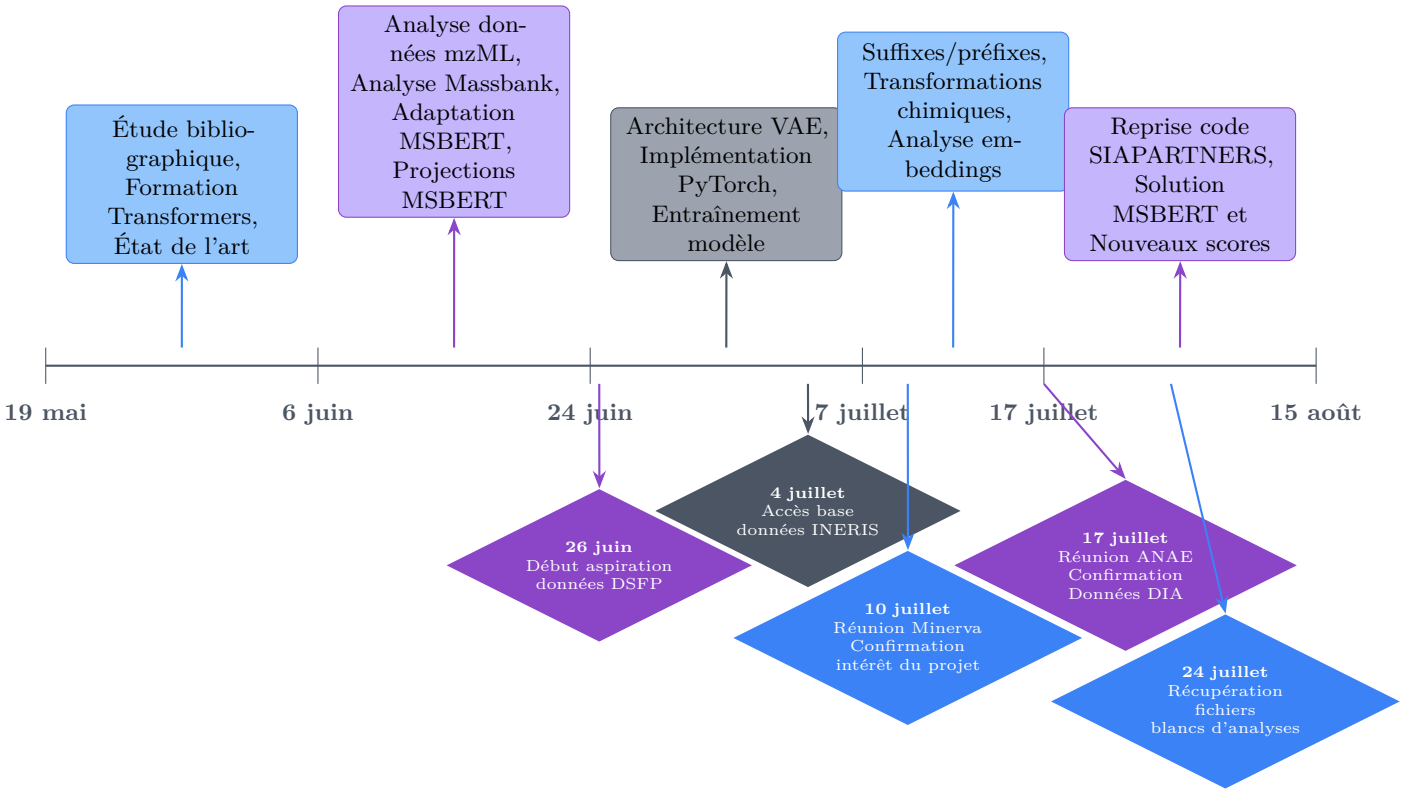


Figure 1 – Chronologie du stage

Enfin, la **dernière partie** dressera le bilan de ce stage. Nous y synthétiserons les résultats et les livrables concrets produits, avant de mener une analyse sur les difficultés surmontées et les apprentissages clés, tant sur le plan technique que méthodologique. En conclusion, nous ouvrirons sur des perspectives pour continuer la recherche sur ces travaux prometteurs au sein de l'INERIS.

4 Contexte scientifique et technologique

Comme prévu dans l'offre de stage, ce dernier a commencé par un travail de bibliographie de quelques semaines, en parallèle, pour comprendre les articles de recherche, un travail de formation et de montée en compétences pour comprendre la chimie nécessaire au projet, ainsi que certaines connaissances sur le Deep Learning qui n'auraient pas été étudiées en cours à l'INSA a été effectué. Notamment le fonctionnement des Transformers. Une ressource pertinente est la formation Fidle du CNRS.

Afin de saisir pleinement les enjeux et la portée des travaux réalisés au cours de ce stage, il est indispensable de poser les fondations scientifiques et techniques du projet. La première partie a pour vocation de présenter la technologie au cœur de nos analyses, la chromatographie en phase liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS), depuis son principe physique jusqu'à la nature des données qui ont été exploitées.

Nous présenterons les recherches issues d'un état de l'art des approches d'intelligence artificielle qui transforment actuellement ce domaine, ce qui permettra d'éclairer et de justifier les choix stratégiques qui ont guidé nos expérimentations. Bien que la méthode puisse sembler compliquée, il faut garder en tête pendant la lecture de ce rapport que l'objectif est simple, caractériser un échantillon prélevé dans le milieu naturel pour caractériser les substances qu'il contient car elles peuvent potentiellement présenter un risque pour l'environnement.

4.1 Le LC-MS/MS : séparer, peser et fragmenter

Au sein de l'INERIS, l'analyse de polluants dans des milieux complexes est une mission fondamentale. La technologie de choix pour cette tâche est le LC-MS/MS, une méthode hybride d'une grande précision qui combine la capacité de séparation de la chromatographie avec la finesse d'identification de la spectrométrie de masse.

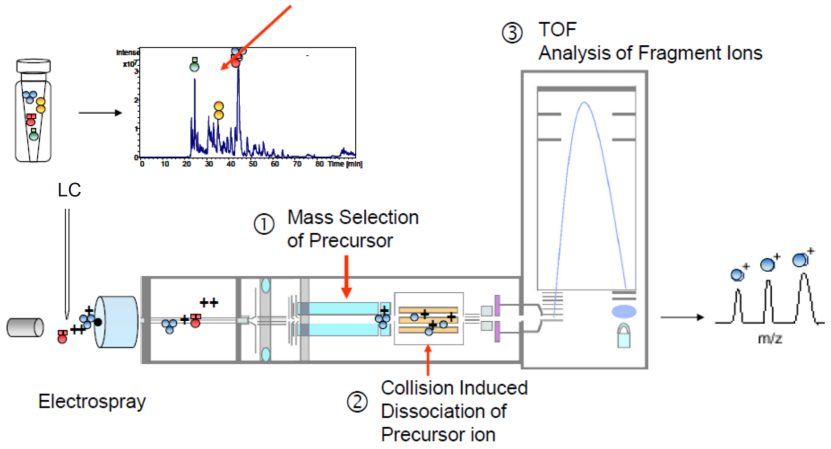


FIGURE 2 – Workflow d'analyse LC-MS/MS issue d'une présentation SIAPARTNERS

Analogie

Pour en saisir le principe de manière intuitive, on peut imaginer le processus en trois étapes :

→ **Démêler** : On démêle un écheveau de fils de toutes les couleurs (l'échantillon complexe) pour les présenter un par un

→ **Peser** : Chaque fil est pesé pour connaître sa masse globale

→ **Fragmenter** : On le casse en morceaux caractéristiques, qui sont pesés à leur tour pour obtenir une signature unique car chaque type de fil se casse d'une manière précise

Techniquement, cela se traduit par les étapes suivantes :

4.1.1 La séparation chromatographique (LC)

L'échantillon brut est d'abord injecté dans une colonne de chromatographie liquide. Cette colonne agit comme un filtre qui ralentit différemment chaque molécule en fonction de ses propriétés physico-chimiques (taille, polarité...). Les molécules sortent ainsi de la colonne en un flux temporellement séparé. Le moment précis où une molécule sort de la colonne est une information précieuse, nommée **temps de rétention**.

La figure 3 montre un chromatogramme, c'est une visualisation graphique des résultats obtenus par chromatographie. Il représente la variation d'un signal, lié à la concentration de chaque molécule séparée, en fonction du temps de rétention. Chaque pic sur le graphique correspond à une molécule distincte qui sort de la colonne à un temps de rétention spécifique (si on néglige les co-élutions où plusieurs molécules sortent en même temps). Dans des conditions d'analyse rigoureusement identiques (même colonne, même température, même débit, etc.), une molécule donnée sortira toujours au même moment. Cette nécessité d'avoir les mêmes conditions expérimentales pour pouvoir comparer les temps de rétention demande une certaine vigilance pour éviter de comparer des choses non comparables.

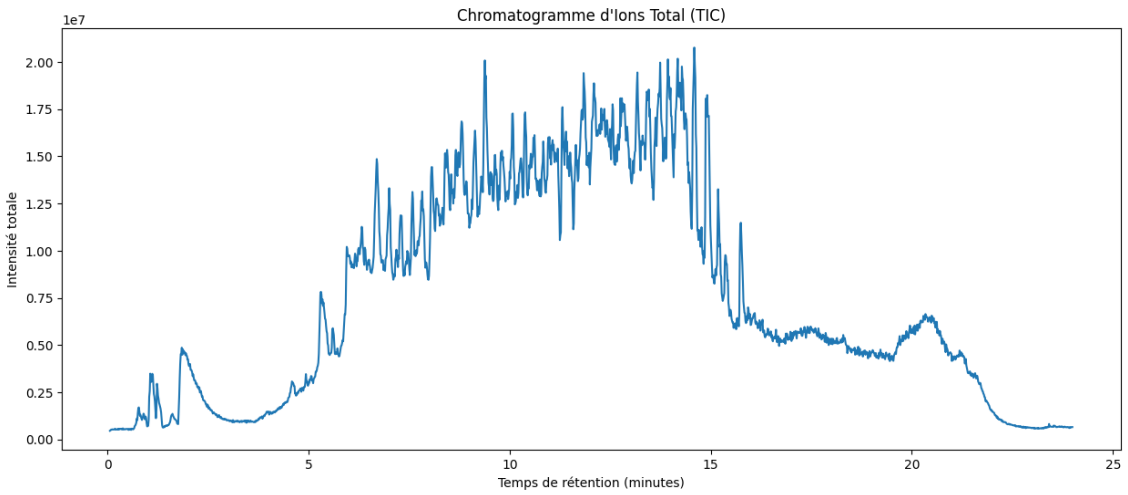


FIGURE 3 – Exemple de chromatogramme

4.1.2 L'Analyse par spectrométrie de masse en tandem (MS/MS)

À la sortie de la colonne, les molécules entrent une par une dans le spectromètre. Elles sont d'abord ionisées, pour qu'on puisse les manipuler avec des champs électromagnétiques. Le spectromètre effectue alors :

→ Mesure les masses de tout ce qui sort de la colonne chromatographique puis isole une raie/un ion (masse

de précurseur) qui va être analysé dans les étapes suivantes. Le spectre issu de cette première analyse est le MS1.

→ L'isolement de cet ion suivi d'une fragmentation contrôlée via un surplus d'énergie appelé **énergie de collision**.

→ La mesure des masses des fragments ou "ions produits" (analyse MS2).

Ce type d'analyse est appelé **DDA (Data-Dependent Acquisition)** car on sélectionne ce que l'on fragmente en analysant directement les données. Il existe aussi un autre mode d'analyse où le spectromètre fonctionne différemment, le **All-Ion Fragmentation** ou **Data-Independent Acquisition (DIA)** . À la différence de la DDA, la DIA fragmente simultanément tous les ions qui entrent dans le spectromètre à un instant t, sans sélection préalable. Le spectre MS2 qui en résulte est donc un mélange complexe des fragments de toutes les molécules co-éluées à ce moment-là.

Cette distinction est fondamentale : un spectre DDA est la signature d'un seul composé, tandis qu'un spectre DIA est la signature superposée de plusieurs composés, rendant son interprétation directe beaucoup plus ardue. La différence est visible sur la figure 4 (un spectre est retourné pour l'affichage, mais les intensités ne sont jamais négatives), on obtient beaucoup plus de raies en DIA car il y a beaucoup de substances différentes dans l'échantillon.

Structure des données spectrales

Le résultat final est un **spectre de masse**, qui représente la signature de fragmentation d'une molécule donnée à un temps de rétention spécifique. Visuellement, il se présente sous la forme d'un graphique où :

- L'axe des abscisses (x) représente le rapport masse/charge (m/z) des fragments régulièrement appelé "masse" par abus de langage
- L'axe des ordonnées (y) leur intensité relative
- Chaque pic correspond à un fragment de la molécule initiale

Ces données et leurs métadonnées (temps de rétention, masse du précurseur, énergie de collision...) sont stockées dans des fichiers au format standard `mzML` basé sur du XML.

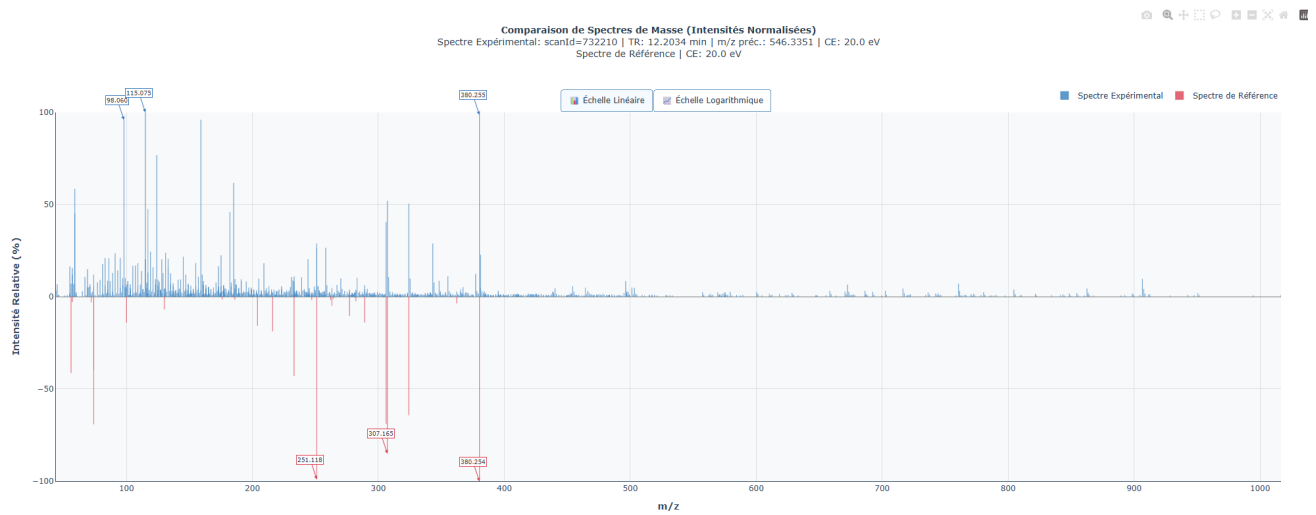


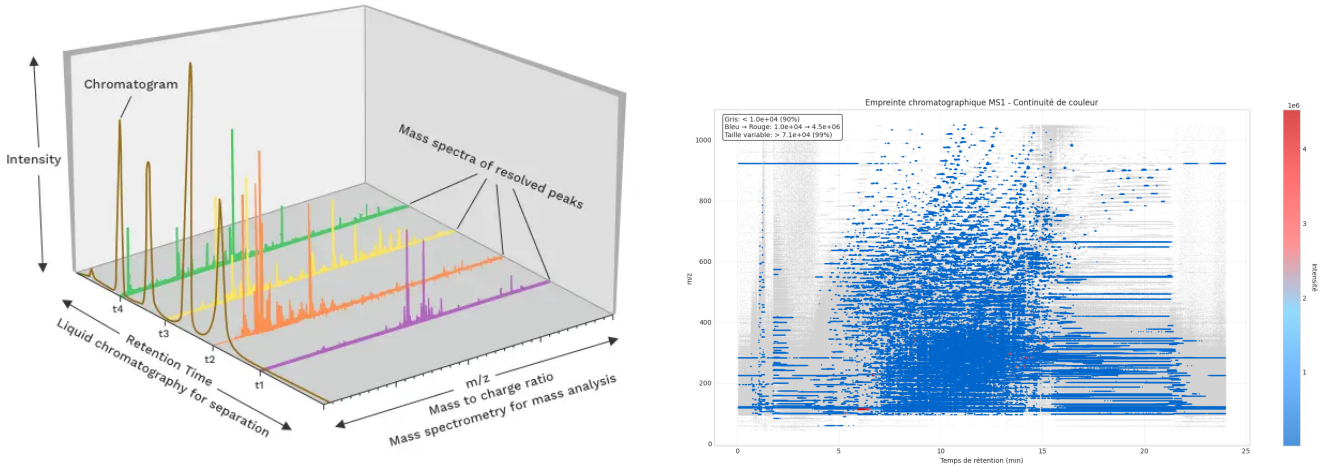
FIGURE 4 – Comparaison à 20 eV d'un spectre DIA (bleu) à un spectre DDA (rouge) de la base de données INERIS

Une erreur qui a été faite et qui devra être évitée dans les futurs travaux est la confusion entre les données DIA et DDA, pendant trop longtemps, il a été considéré que les données étaient exclusivement

issues d'acquisitions de type DDA.

Les tentatives pour comparer des spectres issus des deux modes d'acquisition n'étaient pas concluantes. La présence de plusieurs molécules a été confondue avec un bruit qu'il fallait caractériser dans les fichiers mais qui serait simple à filtrer avec une règle qui permettrait de supprimer les raies "en trop" par rapport au spectre de référence visible dans la figure 4.

On peut récapituler l'information extraite lors de l'analyse LC-MS dans un diagramme 3D qui permet de faire le lien entre le chromatogramme et les spectres de masse figure 5a. On peut aussi faire une empreinte avec les rapports masse sur charge (m/z) selon le temps de rétention en ajoutant l'intensité via la couleur figure 5b.



(a) Représentation 3D LC-MS (source : coursehero)

(b) Empreinte chromatographique

FIGURE 5 – Représentations des données LC-MS

4.2 Stratégies d'analyse : target, NTS, NTA

Pour identifier les composés chimiques dans un échantillon, la spectrométrie de masse s'appuie sur plusieurs stratégies distinctes :

- **L'analyse ciblée (Target Analysis)** : L'approche la plus spécifique, où l'on recherche activement la présence et la quantité d'un ou plusieurs composés connus. La question est : "La molécule X est-elle présente et à quelle concentration ?"
- **L'analyse de suspects (Non-Target Screening / NTS)** : Une approche plus large où l'on compare le spectre expérimental à une liste de composés "suspects" potentiellement présents, en se basant sur des bibliothèques spectrales de référence (MassBank, GNPS, etc.). C'est la tâche de *library matching*.
- **L'analyse non-ciblée (Non-Target Analysis / NTA)** : L'approche la plus exploratoire, visant à identifier des composés totalement inconnus, non répertoriés dans les bibliothèques. C'est elle qui permet de caractériser la "matière noire" chimique d'un échantillon. Elle peut par exemple consister à simuler une structure pour la molécule à partir de son spectre, à faire une régression de ses propriétés chimiques comme la toxicité sans connaître sa structure ou encore à faire du clustering pour observer les molécules "proches".

État des lieux à l'INERIS

Alors que l'analyse ciblée est une pratique maîtrisée à l'INERIS, l'analyse en mode suspects (NTS) fait l'objet d'un effort de recherche et développement continu. Une première solution SIAPARTNERS, basée sur un modèle de Random Forest, a été développée dans le cadre du projet "Empreintes environnementales" en 2021. Cependant, l'outil ne répond pas pleinement aux attentes du laboratoire ANAE en termes d'efficacité.

La NTS et la NTA sont donc les deux pistes que l'INERIS voudrait développer. Dans le cadre du stage, l'objectif principal est d'améliorer les méthodes de NTS mais le travail effectué a un fort potentiel de transfert vers la NTA car les modèles appris fonctionnent avec des données non annotées.

La raison pour laquelle on a autant de raies, c'est qu'on a beaucoup de substances différentes dans l'échantillon.

4.3 Recherches sur le Deep Learning en spectrométrie de masse

Plusieurs travaux de recherche ont permis de guider le projet, deux publications récentes illustrent parfaitement le potentiel des modèles de fondation pour la spectrométrie de masse. Les deux ont été publiées à la même période, et ont des approches similaires.

MSBERT - Analytical Chemistry - Zhang et al. 2024

Le modèle MSBERT propose une analogie directe : il traite un spectre de masse comme une "phrase" où chaque pic (m/z-intensité) est un "mot". Son pré-entraînement vise à ce qu'il apprenne les règles de la fragmentation chimique.

Stratégies d'apprentissage de MSBERT

- → **Apprentissage par masquage (mask learning)** : Le modèle doit deviner des pics qui ont été cachés
- → **Apprentissage contrastif** : S'assurer que des versions bruitées du même spectre restent sémantiquement proches

En apprenant ce "langage" des spectres, MSBERT génère des représentations vectorielles en 512 dimensions (*embeddings*) qui capturent des informations chimiques pertinentes, améliorant ainsi la recherche en bibliothèque. Il est le modèle à l'état de l'art en juin 2025 pour faire du library matching.

DreaMS - Nature Biotechnology - Bushuiev et al. 2025

L'approche DreaMS est la même : un pré-entraînement de modèle Transformer, mais ici sur 24 millions de spectres non annotés, là où MSBERT est entraîné sur seulement 164 000 spectres.

Stratégies de DreaMS

→ **Tâche auxiliaire plus originales** : Prédiction de l'ordre de rétention chromatographique - en présentant deux spectres issus de la même analyse LC-MS/MS, le modèle doit déterminer lequel est sorti de la colonne en premier

→ **Encodage haute résolution** : Utilisation de "Fourier features" pour représenter les masses avec une précision élevée.

Le résultat est un modèle de fondation extrêmement puissant, dont les représentations moléculaires se sont révélées performantes sur un large éventail de tâches. Le modèle pré-entraîné MSBERT se montre plus efficace pour les tâches de library matching.

Impact sur notre approche

Ces deux approches de pointe ont non seulement validé la pertinence de la démarche envisagée pour ce stage, mais ont aussi fourni un socle théorique et pratique solide pour nos propres expérimentations, qui visaient à explorer et à adapter ces concepts afin de démontrer qu'ils pouvaient s'adapter aux besoins spécifiques de l'INERIS et améliorer les capacités d'analyse du laboratoire ANAE.
Étant donné que MSBERT déclare surperformer DreaMS sur la tâche de library matching, la plupart de nos expérimentations se basent sur ce modèle. Il reste cependant nécessaire de départager les approches de ces deux modèles pour chacun des nombreux cas d'utilisation envisageables.

Il est aussi important de citer les travaux qui ont permis de faire des choix pour guider le projet ou simplement de confirmer que la direction prise était la bonne.

- Jin et al. 2025, Hupatz et al. 2025 Beck et al. 2024, ont permis de comprendre les enjeux du sujet et de mettre les recherches actuelles dans leur cadre global.
- Jonge, Hooft et Probst 2025 a permis de déterminer une manière efficace d'utiliser les spectres de masse pour apprendre un modèle ainsi que la précision à utiliser pour les représenter. De plus, il a contribué à illustrer l'efficacité des architectures transformer en termes de performances par rapport au nombre de paramètres.
- Degnan et al. 2023 a permis de trouver les scores non basés sur le Deep Learning les plus efficaces pour comparer des spectres et de comprendre qu'il n'y a aucun consensus clair à ce niveau, d'où la nécessité d'avoir des modèles comme MSBERT et DreaMS.

D'autres articles avec des approches intéressantes n'ont pas été explorés mais mériteraient de l'attention comme Bui-Thi et al. 2024, Liu et al. 2025.

5 Outils de développement et données

Afin de permettre la meilleure reproductibilité possible, les outils utilisés pour le développement ainsi que les données qui ont permis d'obtenir les résultats seront détaillés dans cette partie du rapport.

5.1 Outils de développement

Le développement a été réalisé sur la plateforme Onyxia de sspcloud qui permet d'accéder à des GPUs A100 qui ont offert la puissance de calcul nécessaire pour les différentes expérimentations, entraînements et inférences de modèles de Deep Learning. Le code est sourcé à l'aide du gitlab privé de l'INERIS. Un espace de stockage S3 est disponible pour stocker les données.

Que ce soit pour le développement d'algorithmes, de modèles Keras et PyTorch ou de fonctions utilitaires, le code a été écrit dans des notebooks Python permettant de présenter rapidement et de manière visuelle les résultats obtenus tout au long du stage. Cela a permis à mon tuteur de valider ou non les résultats obtenus de faire participer à la réflexion les experts métiers du laboratoire ANAE. De plus les notebooks permettent une meilleure reproductibilité de nos résultats et une reprise plus facile du projet à l'avenir.

Les fonctions les plus utiles ont progressivement été regroupées dans un module Python facilement installable pour permettre des imports de fonctions rapides dans tous les notebooks.

librairies Python utilisées : Trois librairies Python ont été explorées pour récupérer les données des mzML, PymzML, Pyteomics et Matchms, il a été décidé d'utiliser Pyteomics pour récupérer les données dans les fichiers et conjointement Matchms pour certaines fonctions de traitement des données. Ces librairies permettent de parcourir les données XML des mzML, pour récupérer les spectres et métadonnées qu'ils contiennent. Pour l'analyse de données et les traitements de Machine Learning, les librairies Numpy, Scikit-learn, pandas et scipy ont été utilisées. Matplotlib et Plotly ont permis de réaliser la plupart des visualisations de données, notamment celles présentes dans ce rapport. Pour le Deep Learning, les librairies Keras 3 et PyTorch ont été utilisées.

5.2 Collecte et analyse des données

La constitution d'un jeu de données pertinent et de qualité est une base sans laquelle un projet a beaucoup de chances de partir dans la mauvaise direction. Cette étape, qui s'est étendue sur tout le stage, a été un processus itératif riche en apprentissages sur la nature des données spectrales.

Il a été nécessaire de questionner les experts métier pour savoir si certaines sources de données existaient et étaient mobilisables. Le défi est le même que dans tous les projets de Machine Learning : être en capacité de mobiliser rapidement des données de qualité pour l'entraînement et la validation des modèles. Cela a mené à l'utilisation de plusieurs sources de données, la moitié étant interne à l'INERIS, l'autre publiquement accessible (tableau 1). Tout au long du stage, chaque source de données a été analysée pour comprendre ses spécificités.

Au démarrage du projet, deux sources de données principales étaient à disposition :

- La bibliothèque publique de référence **MassBank EU** alimentée par plusieurs institutions européennes.
- 4 échantillons d'eau de quatre rivières analysés par l'INERIS via un spectromètre Agilent.

Le besoin de données plus fournies et plus représentatives des analyses menées à l'INERIS s'est fait sentir. Il a aussi été décidé d'utiliser

- La base de données interne de l'institut, contenant des analyses d'étalons, c'est-à-dire des produits chimiques purs analysés dans des conditions contrôlées pour obtenir des spectres de référence de très haute qualité.
- Une partie des données de la **Digital Sample Freezing Platform (DSFP)** a été collectée via l'API REST. DSFP est une initiative du réseau NORMAN (network of reference laboratories, research centres and related organisations for monitoring of emerging environmental substances), dont l'INERIS

fait partie, visant à créer une "bibliothèque d'échantillon digitale" en archivant des données brutes de spectrométrie de masse pour permettre des analyses rétrospectives.

TABLE 1 – Sources de données utilisées dans le projet

Source	Format	Nombre de fichiers	Nombre de spectres	Format stockage	Annoté
MassBank EU	Texte standardisé	-	123 000 MS2	DataFrame / Parquet	Oui
Analyses de 4 rivières	mzML	4	3700 MS2 et 1800 MS1 par fichier	DataFrame / Parquet	Non
Base interne INERIS + Agilent	Spectres MS2	-	257 (INERIS) + 1 042 (Agilent) en MS2	DataFrame / Parquet	Oui
Digital Sample Freezing Platform (DSFP)	mzML	12 700	moyenne à 3500 par fichier avec un tiers de MS1	DataFrame / Parquet	Non

Parmi toutes ces données, une distinction doit être faite entre les données annotées qui sont des spectres associés à des informations permettant d'identifier la molécule (codes SMILES, CAS, InChIKey, ...) comme Massbank et la base INERIS et les données non annotées qui se limitent à des fichiers mzML contenant des spectres avec leur énergie de collision, leur temps de rétention et d'autres métadonnées (type d'ionisation par exemple).

L'agrégation de ces différentes sources a mis en évidence un défi majeur : l'extrême **hétérogénéité des données spectrales**. Les données pour une même molécule peuvent varier drastiquement en fonction des conditions expérimentales :

- **La polarité de l'ionisation** : En mode positif ou négatif, la molécule ne se fragmente pas de la même manière.
- **Les énergies de collision** : Une énergie plus élevée provoque une fragmentation plus intense, générant potentiellement plus de fragments et donc un spectre avec plus de raies.
- **Les conditions chromatographiques** : Comme mentionné précédemment, le temps de rétention est très difficilement comparable entre des expériences menées avec des colonnes, des solvants ou des températures distinctes. Par contre, pris dans les mêmes conditions expérimentales, il devient l'un des moyens d'identification avec le plus haut taux de confiance.

Ces 4 sources de données ont été intégrées dans des dataframes pandas (des tableaux de données) et sauvegardés en fichiers parquet ou CSV après transformation.

m/z array	intensity array	precursor_mz	collision_energy	retention_time	spectrum_id	file	ms_level	total_ion_current
[44.01211696875716, 44.0479897698135, 46.06458...	[1138.913818359375, 700.84759552148438, 1006.16...	484.014177	20.0	0.052300	scanId=3146	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	2	419612.188
[44.012222971131415, 44.04851474212662, 44.060...	[3369.95703125, 1033.0382080078125, 354.809509...	484.013606	40.0	0.056583	scanId=3402	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	2	331726.500
[59.05988313686875, 76.93323481526002, 83.9525...	[1403.2823486328125, 215.69444274902344, 277.2...	0.000000	NaN	0.060867	scanId=3659	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	1	464521.906
[44.01181844118489, 44.019141036533366, 44.0476...	[917.2074584960938, 466.84951782226562, 425.474...	540.294086	20.0	0.065167	scanId=3918	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	2	432259.406
[43.95397123010809, 44.011752676139998, 44.0482...	[251.80372619628906, 2088.841064453125, 1444.2...	536.890062	40.0	0.069450	scanId=4174	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	2	328888.313
[54.984739338375554, 59.059601796593036, 60.06...	[268.29132080078125, 1493.31689453125, 219.992...	0.000000	NaN	0.073733	scanId=4431	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	1	485365.719
[44.01261881060277, 44.047708849035644, 45.032...	[798.241943359375, 476.2348937988281, 568.2785...	501.986291	20.0	0.078017	scanId=4689	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	2	434628.000
[44.0121803609477, 44.048826414378354, 44.9783...	[1750.5101318359375, 1665.1868896484375, 326.2...	503.604307	40.0	0.082300	scanId=4946	06-09_18-6-014-R-escaut-POCIS-allion-pos-1.mzML	2	323441.656

FIGURE 6 – exemple de dataframe généré par notre fonction mzml_to_dataframe

Ensuite nous avons défini une structure de données pour uniformiser le tout et pouvoir commencer à travailler sur les données globales. Ce traitement n'a pas pu être finalisé pendant le stage faute de temps, au moment où les dernières données ont été reçues, d'autres tâches étaient prioritaires, les 4 sources de données ont été mises sous cette forme mais un fichier contenant toutes les données n'existe pas. Cependant

cette structure est ce qui a été jugé pertinent de rassembler et d'avoir pour faire un travail de Deep Learning à partir des données des sources présentées.

TABLE 2 – Schéma de données unifiée pour les spectres de masse.

Nom Colonne	Description	Type de Donnée	Disponibilité
ID_Spectre	Identifiant unique pour chaque spectre	Entier ou Chaîne	Toutes sources
Source	Origine de la donnée (Massbank, INERIS, etc.)	Catégorie/Chaîne	Toutes sources
File_name	Nom du fichier d'origine pour les mzml	Chaîne	Toutes sources
Ion_polarity	Polarité de l'ionisation	Enumération [Positif, Négatif]	Toutes sources
Ion_mode	ion utilisé (ex : [M+H]+)	Chaîne	Données annotées
Precursor_mass	Masse m/z du précurseur	Flottant (32-bit)	Toutes sources
Retention_time	Temps de rétention standardisé en minutes	Flottant (32-bit)	Toutes sources
Collision_Energy	Énergie de collision normalisée en eV (arrondie)	Entier	Toutes sources
m/z_array	Liste des valeurs m/z des fragments	Objet (Array de flottants)	Toutes sources
Intensity_array	Liste des intensités relatives des fragments	Objet (Array de flottants)	Toutes sources
Annotated	Indique si la molécule est identifiée	Booléen	Toutes sources
smiles	Représentation textuelle de la structure moléculaire	Chaîne	Données annotées
inchi / inchikey	Autres identifiants structuraux	Chaîne	Données annotées
Instrument_type	Type d'instrument utilisé pour l'acquisition	Catégorie/Chaîne	Toutes sources

5.3 Préparation et transformation des données

Les données analysées provenant de plusieurs sources, chaque source a été analysée individuellement pour trouver les potentiels biais statistiques. Ces analyses ont principalement mené à se concentrer uniquement sur les spectres acquis en mode positif et à effectuer des travaux de normalisation et d'équivalence comme pour les énergies de collision qui étaient très variables comme on le voit figure 7.

Pour la suite du travail, les temps de rétention ont toujours été utilisés en minutes, les énergies de collision en eV et arrondis à l'unité si les valeurs étaient flottantes. Des traitements existent pour ramener les énergies de collision nominales en eV si on a la connaissance du base peak du spectre (le pic avec l'intensité la plus forte dans un spectre), pour d'autres formats la conversion en eV n'a pas été possible. Après ce travail de préparation, certaines données n'ont finalement pas été exploitées. La solution de NTS développée n'a exploité que les échantillons de rivière INERIS et la Base INERIS mais beaucoup d'autres expérimentations ont été réalisées sur Massbank. Les visualisations de données indiqueront la source des données exploitées.

Tous les spectres ont été tronqués en supprimant la partie supérieure à 1000 m/z car le modèle MSBERT pré-entrainé n'a appris que sur des données allant de 10 à 1000 m/z, les masses hors de cette plage créent des tokens qui n'ont pas de sens pour le modèle. De plus, l'analyse de polluants cible souvent des petites

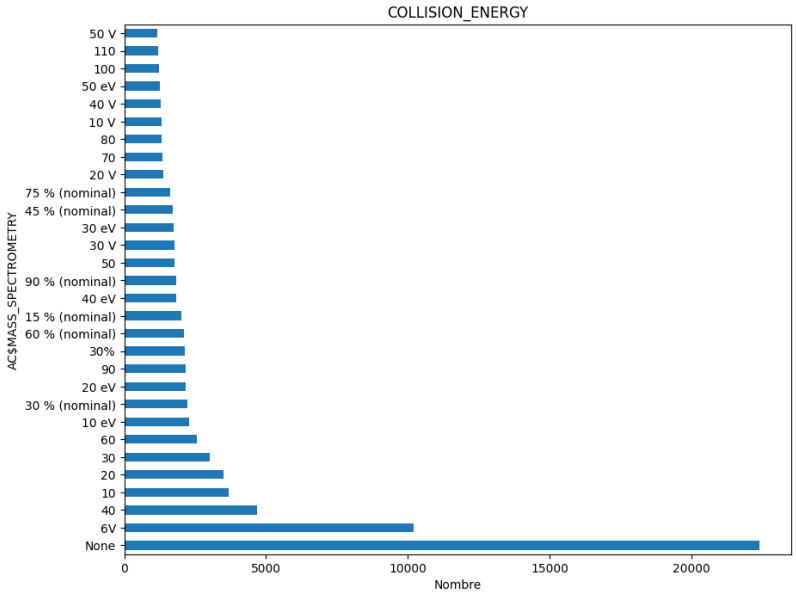


FIGURE 7 – Fréquences des différentes énergies de collision de Massbank

molécules avec une masse faible et l'expert métier nous a confirmé que l'on travaille principalement dans la plage 40 à 500 m/z.

Beaucoup de raies étant régulièrement en dehors de la plage, cela a renforcé les soupçons sur la présence de bruits dans les spectres, certains bruits ont en effet pu être découverts dans les données. Une analyse de la fréquence d'apparition des masses dans les spectres des 4 échantillons a révélé des masses suspectes. Certaines présentes dans plus de 90% des spectres ne pouvaient pas être "normales" car une masse est censée caractériser des molécules ou des fragments et donc sortir du chromatographe à un instant donné, pas tout au long de l'analyse, et ce sur 4 analyses indépendantes.

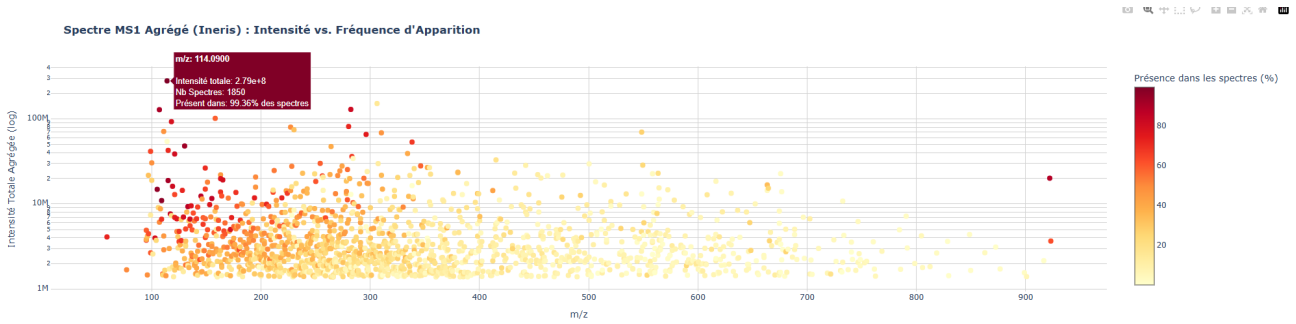


FIGURE 8 – Analyse des masses par leur intensité et leur fréquence sur l'échantillon Escaut

Après recherche, il s'est avéré que certaines masses atypiques correspondaient à des substances de référence utilisées dans le protocole de LC-MS Agilent, elles ont pu être retrouvées dans la documentation officielle du constructeur et pouvaient donc être supprimées systématiquement des spectres. Cependant, la majorité des masses avec des fréquences élevées n'ont pas été identifiées ni supprimées des échantillons au risque de supprimer des informations pertinentes.
Ces premières recherches de bruit concluantes ont soulevé d'autres questions, face à ces interrogations, l'experte métier a pu fournir différents fichiers de blancs, notamment des blancs terrains POCIS (outil de prélèvement utilisé pour les échantillons environnementaux réels) et des blancs en laboratoire. Faute de temps, les blancs en laboratoire n'ont pas pu être analysés et le protocole pour les créer n'a pas été abordé, cependant ils sont probablement une source d'information non négligeable pour caractériser le bruit dans l'échantillon.

Importance du traitement des données

L'objectif du projet étant d'utiliser des modèles capables de créer des représentations de ces signaux spectraux via un apprentissage profond, il est inenvisageable d'obtenir une représentation qui ait du sens sans nettoyer les échantillons très bruités.

L'analyse des blancs terrains est surprenante car ils semblent très similaires à nos échantillons d'eau de rivière, ce qui fait penser que le ratio signal sur bruit n'est pas satisfaisant (voir figure 9) . D'autant plus que beaucoup des masses très fréquemment présentes dans l'échantillon le sont aussi dans le bruit, d'où la similarité entre la figure 8 et la figure 10.

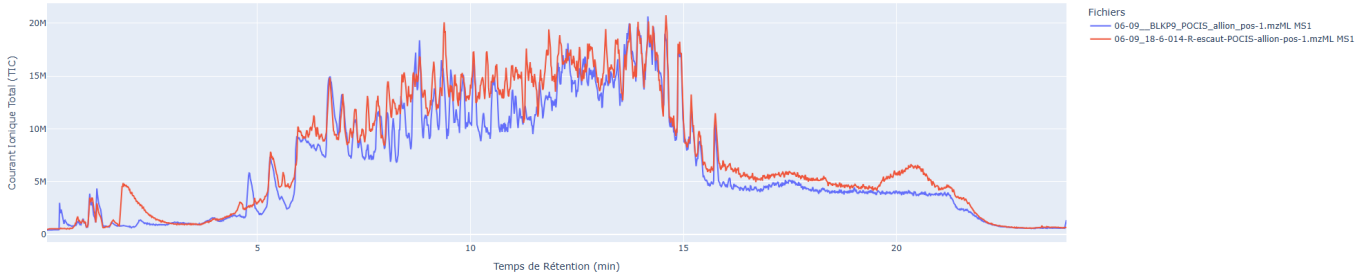


FIGURE 9 – Comparaison des chromatogrammes de l'échantillon et du blanc terrain Escaut

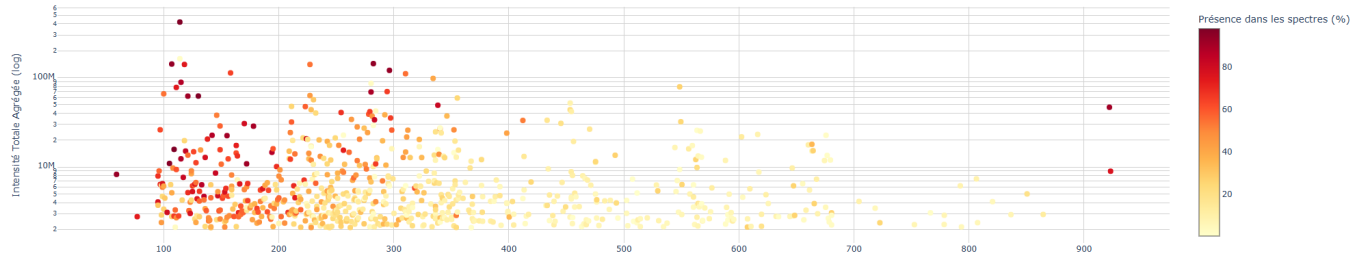


FIGURE 10 – Analyse des masses par leur intensité et leur fréquence sur le blanc terrain Escaut

Aucun blanc n'a encore été exploité dans le développement de solutions de NTS mais le fait de pouvoir nettoyer les échantillons DIA qui contiennent beaucoup de substances mais donc aussi beaucoup de bruit expérimental, est une piste privilégiée pour réaliser une solution de NTS plus efficace et une piste obligatoire pour réaliser une solution de NTA pertinente.

6 Développement et expérimentation

Les dernières recherches en spectrométrie de masse suggèrent que seulement 2 % des spectres peuvent être annotés à l'aide des librairies spectrales existantes. Cette idée évoquée dans l'introduction des articles MSBERT et DreaMS, oriente la recherche vers des architectures de modèles capables d'apprendre sur des données non annotées. L'avantage de ces architectures sera pour nous que les modèles pourront facilement être fine-tunés sur des données spécifiques à un domaine sans nécessiter d'annotations massives.

Parmi les modèles testés, l'architecture qui a retenu notre attention est celle de MSBERT, qui est open source et open weights, un travail a été réalisé pour identifier les forces et points faibles du modèle, les potentielles modifications nécessaires de son architecture, à son fine-tuning ou encore son intégration dans un modèle en tant que squelette pré-entraîné, bien que certaines n'aient pas été retenues pour le moment, ces pistes restent toutes très pertinentes.

Mon tuteur souhaitait aussi explorer une architecture de Variational Auto Encoder (VAE) dont le pouvoir de clustering peut se révéler meilleur qu'un modèle BERT. Nous avons testé sur nos données les modèles qui étaient ressortis de l'étude bibliographique initiale, MSBERT et DreaMS, et deux premières architectures de VAE et une de conditional VAE (CVAE) ont été entraînées mais n'ont pas encore abouti à des résultats concluants.

6.1 Essais d'entraînements de modèles

La piste des VAE a été explorée brièvement et pas suffisamment pour tirer des conclusions quant à leur utilisation dans les applications concrètes du laboratoire ANAE. Les résultats avec MSBERT étant plus concluants plus rapidement et au vu de la courte durée du stage, il a été décidé que cette piste ne serait pas poursuivie durant le stage. Pour autant, cette piste n'est pas inintéressante et mérite d'être présentée.

L'intérêt d'une architecture VAE réside dans le fait que l'espace latent formé par le modèle est continu. Il est donc possible de se déplacer linéairement dans l'espace et de faire des interpolations entre les points associés aux données pour tenter d'interpréter les critères sur lesquels le modèle a espacé ces points. Comme l'espace est continu, peu importe où l'on se trouve, pour peu qu'on ne s'éloigne pas trop des zones de l'espace contenant des données d'apprentissage, il est possible de générer des données cohérentes.

Pour réussir à faire cela on apprend un modèle qui minimise cette fonction de perte (loss) composite.

Fonction de loss VAE

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathcal{L}_{\text{reconstruction}}(X, \hat{X})}_{\text{Fidélité}} + \underbrace{\beta \cdot D_{KL}(q(z|X) \| p(z))}_{\text{Régularisation}}$$

- $\mathcal{L}_{\text{reconstruction}}$: Qualité de la reconstruction (MSE, BCE)
- D_{KL} : Divergence KL vers distribution a priori $\mathcal{N}(0, 1)$
- β : Poids d'équilibrage des loss

L'une des grandes difficultés est de régler le paramètre β car il est essentiel pour obtenir une bonne reconstruction avec un espace latent bien structuré, s'il est trop faible on n'observera aucune structure dans l'espace et s'il est trop grand, la reconstruction sera mauvaise.

Trop peu d'essais ont été effectués et aucun des entraînements n'a fourni de résultats satisfaisants. Cependant, le point positif est que les problèmes rencontrés ont des solutions bien qu'il ait été décidé de ne pas poursuivre ces travaux car les autres pistes semblaient plus raisonnables. Deux principaux problèmes sont que les architectures de VAE testées n'étaient peut-être pas assez complexes (uniquement des couches LSTM et des couches denses), et que le coefficient d'équilibrage des loss n'a pas été réglé. Les courbes de loss fournies en annexe 9.3 permettent de visualiser la non convergence du dernier entraînement de VAE réalisé.

En tout, 4 entraînements durant entre 2 et 7 heures sur des GPU A100 auront été lancés, deux sur un modèle de VAE "classique" et deux sur un conditional VAE (CVAE) qui prenait en entrée les énergies de collision comme condition. L'idée du CVAE était de pouvoir apprendre à générer un spectre à une énergie de collision de 40eV à partir par exemple du 20eV ou du 0eV.

Lors d'une réunion avec le consortium MINERVA (un projet européen visant à supporter le développement de l'IA). Nos différents projets et essais ont été présentés, il a été conseillé de s'intéresser en priorité aux architectures transformer comme par exemple MSBERT ou bien aux Graph Neural Networks (GNN) qui seraient les plus prometteuses face aux données spectrales telles que nous leur avons présentées. Cela a aussi contribué à ne pas continuer la recherche sur les VAE qui pourraient être plus difficiles à entraîner pour obtenir de bonnes performances en généralisation car la reconstruction exacte de spectres est une tâche complexe à optimiser

La piste des VAE reste donc ouverte et à creuser dans de futurs travaux de l'INERIS, une veille des

articles de recherche sur ces architectures de modèles dans le domaine de la spectrométrie de masse serait pertinente.

6.1.1 Évaluation des modèles

Une liste de métriques permettant de définir un "bon modèle" et de les comparer a été réalisée. Elle combine deux types de métriques :

→ Celles pour le library matching qui sont similaires à des métriques de RAG en NLP puisque le but est de voir si le modèle récupère bien les bonnes informations à partir d'une base de connaissances.

→ Celles pour le clustering dans un espace latent structuré. Les propriétés de regroupement sont intéressantes sur le plan chimique par exemple pour faire un clustering de molécules aux propriétés chimiques similaires ou de la classification.

Modèle	Library Matching		Clustering			
	Recall@k	MRR	ARI	Homogénéité	Complétude	Silhouette Score
VAE	x	x	x	x	x	x
MS-BERT	x	x	x	x	x	x
DreaMS	x	x	x	x	x	x

Table 3 – Tableau illustratif des métriques

Le tableau 3 n'a finalement pas été rempli durant le stage mais constituer un jeu de données pour réaliser un benchmark capable d'évaluer les modèles sur les capacités attendues serait intéressant pour quantifier les éventuelles avancées futures dans le projet "Empreinte environnementale".

6.2 MSBERT et adaptations

Le modèle MSBERT est un modèle qui s'inspire des modèles de NLP ici plus précisément de BERT (Bidirectional Encoder Representations from Transformers). L'approche transpose les concepts de l'analyse textuelle au domaine de la spectrométrie de masse. L'idée fondamentale est la suivante : si BERT peut apprendre le sens d'un mot en analysant son contexte dans une phrase, MSBERT peut apprendre la signification d'un pic m/z en analysant son contexte au sein d'un spectre complet et créer une représentation de ce spectre dans un espace latent.

Contrairement à l'architecture complète des Transformers d'origine (à la fois encodeur et décodeur, utilisée par exemple en traduction automatique pour encoder le sens d'une phrase en anglais et le décoder en français), BERT se limite à la seule partie encodeur. Cela se reflète visuellement dans la figure 11, où l'on voit que BERT reprend uniquement l'empilement de couches d'attention multi-tête suivies de couches feed-forward, sans inclure de décodeur ou d'attention masquée (9.1.2).

Cette architecture encodeur uniquement permet à BERT de traiter une séquence dans sa globalité, en appliquant un mécanisme d'attention bidirectionnelle : chaque token (ou ici, chaque pic m/z) peut intégrer l'information de tous les autres, peu importe leur position. Cela diffère fondamentalement des modèles de type GPT, qui n'utilisent qu'un décodeur avec attention causale (masquée), ce qui les contraint à ne prendre en compte que le passé, et non le futur dans la séquence.

Dans le cas de MSBERT, cette structure encodeur est particulièrement adaptée, car l'objectif n'est pas de générer une séquence de masse ou d'intensités, mais de produire une représentation riche et contextuelle du spectre dans son ensemble. Chaque pic est contextualisé à l'intérieur du spectre par la couche d'attention bidirectionnelle, de manière analogue à la contextualisation d'un mot dans une phrase. Cette représentation

contextuelle peut ensuite être utilisée pour des tâches comme la mise en correspondance (matching), la classification ou l'annotation.

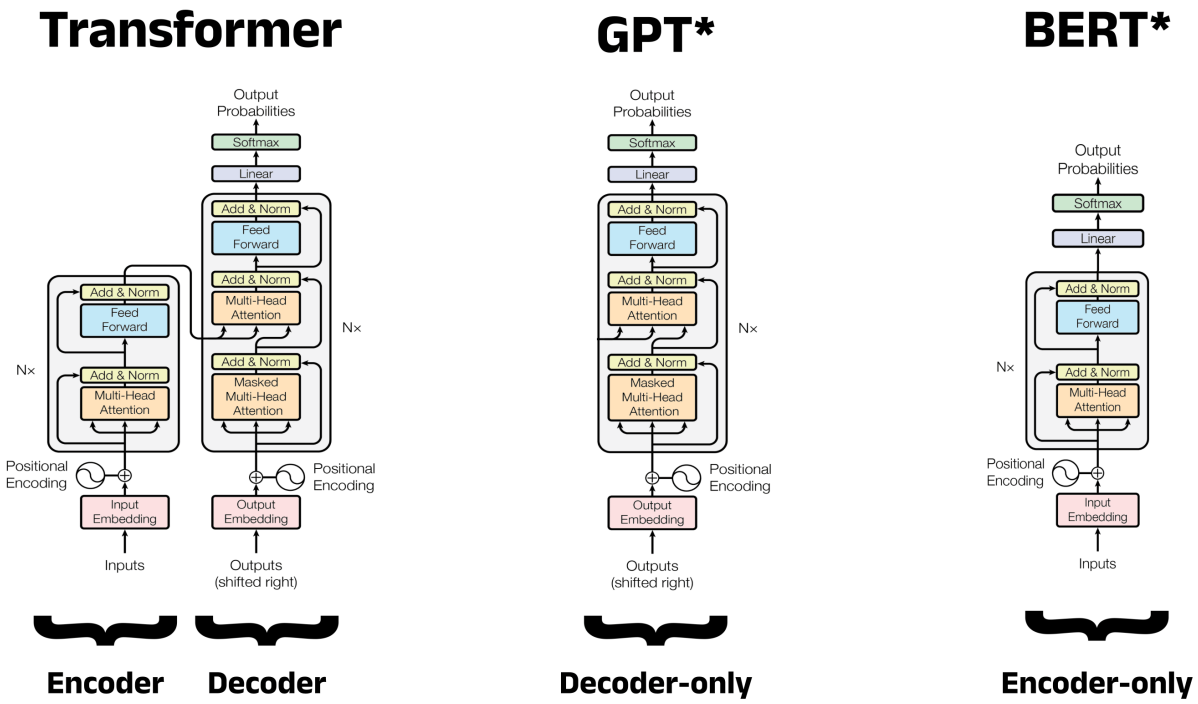


FIGURE 11 – Schéma comparatif des architectures Transformer et BERT (Source towardsdatascience)

Pour adapter l'architecture BERT à la nature unique des données spectrales, plusieurs modifications sont apportées (analogie figure 12) :

- **Du spectre à la séquence :** Un spectre MS/MS est traité comme une "phrase" où les "mots" sont les pics de fragments. Chaque pic est défini par son m/z et son intensité. Contrairement à une phrase, l'ordre des pics dans un spectre n'a pas de signification intrinsèque. Par conséquent, l'étape d'*encodage positionnel*, fondamentale dans les modèles BERT classiques pour encoder l'ordre des mots, est ici supprimée. La notion "d'ordre" est ici intrinsèque à la définition des tokens.

$$[(86.096,\ 999),\ (114.52,\ 850),\ (328.0485,\ 120)] \equiv [(114.52, 850), (86.096, 999), (328.0485, 120)$$

- **L'intensité comme information clé :** Alors que le m/z indique "quel" fragment est présent, l'intensité indique "en quelle proportion". MSBERT conserve cette information capitale en l'utilisant pour pondérer les représentations vectorielles (embeddings) de chaque pic m/z. Ainsi, les pics les plus intenses, qui sont souvent les plus informatifs et les moins bruités, ont plus de poids dans la représentation finale du spectre.
- MSBERT prend aussi une autre information clé en entrée : la masse du précurseur. Elle permet au modèle de situer la masse qui a été fragmentée dans le MS1 et qui donne le spectre MS2 que l'on veut représenter. Cette masse est concaténée à la liste de token, mais là où toutes les autres intensités sont normalisées entre 0 et 1 celle-ci reçoit une valeur à 2 pour aider le modèle à comprendre son importance. Cela est négligé dans le schéma figure 12 par souci de simplification.

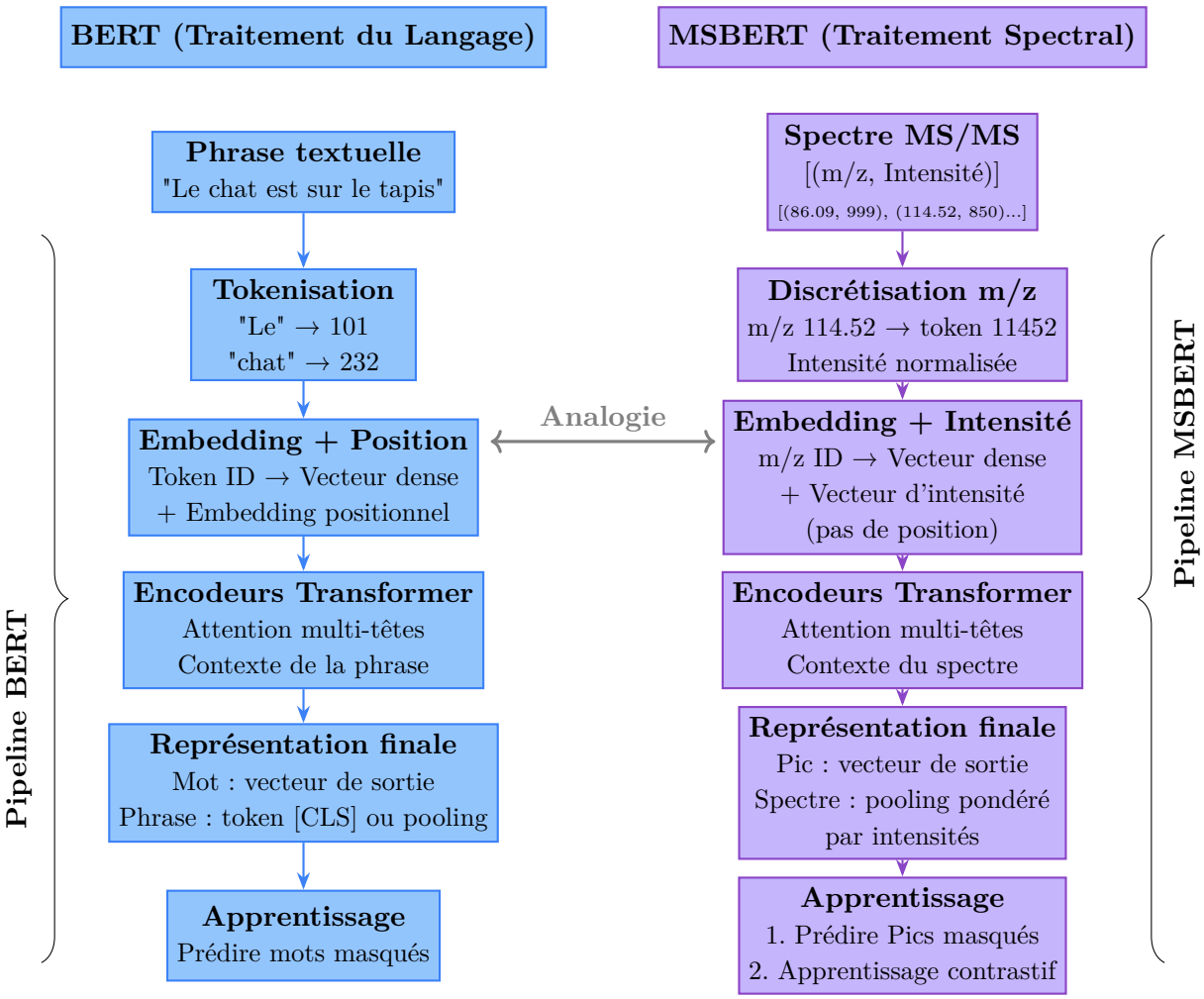


FIGURE 12 – Schéma comparatif des architectures BERT et MSBERT

1. L'apprentissage par masquage de fragments (*Masked Fragment Modeling*) : Cette tâche est l'analogue du *Masked Language Modeling* de BERT. Le modèle masque aléatoirement deux pics parmi les cinq plus intenses d'un spectre et son objectif est de prédire les valeurs m/z de ces pics masqués en se basant sur les pics environnants (le "contexte"). Pour réussir cette tâche, le modèle n'a d'autre choix que d'apprendre les relations chimiques sous-jacentes : quelles familles de fragments apparaissent systématiquement ensemble ? Quelle est la probabilité d'observer un fragment X si le fragment Y est présent ? Il apprend ainsi la logique de fragmentation des molécules. La fonction de loss utilisée pour apprendre ce modèle est une Cross Entropy.

2. L'apprentissage contrastif (*Contrastive Learning*) : Cette seconde tâche vise à garantir que l'embedding généré soit une signature robuste et spécifique de la molécule d'origine. Le principe est le suivant :

- **Échantillons positifs :** On prend un même spectre et on y applique deux masques différents. Bien que les deux versions masquées soient légèrement différentes, elles proviennent de la même molécule. Le modèle est donc entraîné à rendre leur embedding aussi similaire que possible (maximiser la similarité).
- **Échantillons négatifs :** On prend des spectres issus de molécules différentes au sein d'un même batch d'entraînement. Le modèle est entraîné à rendre leur embedding aussi dissemblable que possible (minimiser la similarité).

En combinant ces deux objectifs, l'apprentissage contrastif pousse le modèle à créer un espace latent où

les variations intra-molécule (dues au bruit ou au masquage) sont minimisées, tandis que les différences inter-molécules sont maximisées. La loss est celle classiquement utilisée pour réaliser des apprentissages contrastifs : InfoNCE.

Pour récapituler, l'architecture MSBERT présentée dans la figure 12 est entraînée avec une fonction de loss composite qui combine les deux objectifs d'apprentissage. La loss totale s'exprime comme : $L_{\text{totale}} = L_{\text{CrossEntropy}} + L_{\text{InfoNCE}}$, où la loss Cross Entropy pénalise les mauvaises prédictions des pics masqués tandis que la loss InfoNCE rapproche les représentations des deux versions masquées du même spectre tout en éloignant celles de spectres différents. Cette approche d'entraînement unifié permet au modèle d'apprendre à créer ainsi des représentations spectrales riches et contextualisées dans l'espace latent.

Dans l'espace latent du modèle, les spectres peuvent être comparés entre eux, comparés à des bases de référence (NTS), clusterisés pour rapprocher des substances inconnues et jamais référencées avec d'autres substances, ou encore servir de représentation de base pour faire de la génération de structure moléculaire "de novo". Énormément de tâches peuvent être imaginées, c'est là tout l'intérêt des modèles de fondation.

Impact sur notre approche

L'adoption de l'architecture MSBERT représente un changement de paradigme. Au lieu de comparer directement les spectres bruts, qui sont des objets de taille variable et bruités, nous cherchons à les projeter dans un espace vectoriel de taille fixe, "chimiquement rationnel". Dans cet espace, une simple mesure de similarité cosinus entre deux vecteurs devient une approximation plus efficace de la similarité spectrale que les scores algorithmiques utilisés traditionnellement.

6.2.1 Exploration de la rationalité de l'espace latent de MSBERT

Pour vérifier la rationalité de l'espace latent de MSBERT, la plupart des tests ont été faits sur les données Massbank, étant donné qu'elles étaient les seules données annotées disponibles dès le début du stage. Après quelques modifications de la pipeline de pré-traitement de MSBERT, une projection de 35 000 spectres sur les 123 000 de Massbank a été réalisée. Grâce aux annotations des spectres, le regroupement des molécules dans l'espace latent a pu être confirmé visuellement en réduisant la dimension des vecteurs d'embeddings des spectres de 512 à 2 ou 3 dimensions avec des algorithmes comme UMAP, T-SNE ou une PCA. Les figure 13 et figure 14 montrent des exemples de ces projections sur lesquelles un regroupement est visible, les points (spectres) sont colorés par SMILES, c'est-à-dire par molécules.

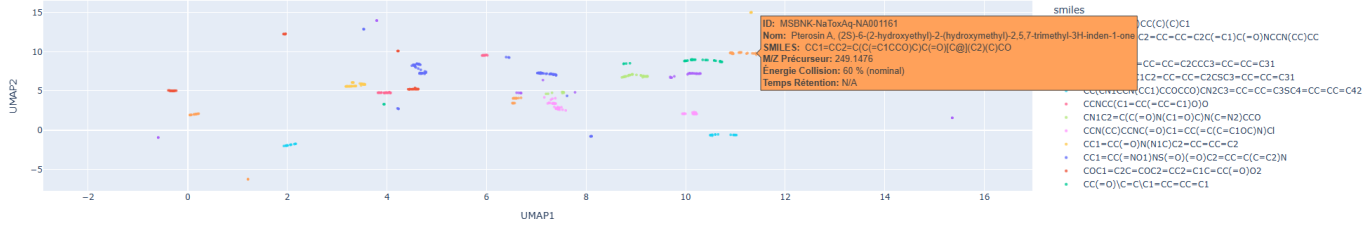


FIGURE 13 – UMAP en 2D de projections de spectres Massbank par MSBERT colorées par SMILES

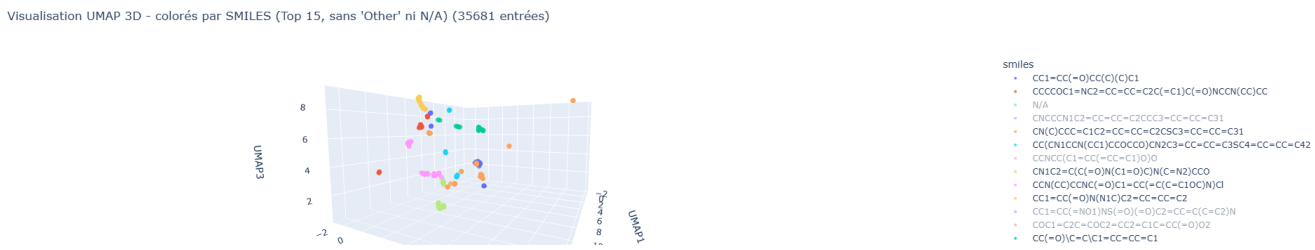


FIGURE 14 – UMAP en 3D de projections de spectres Massbank par MSBERT colorées par SMILES

Deux autres modalités issues des annotations n'ont pas montré de regroupement à l'échelle de ces 35 000 points après réduction de dimension. Au niveau des données Massbank, nous n'avons pas été en mesure de montrer un regroupement sur des informations ne faisant pas partie des données d'apprentissage comme le temps de rétention et l'énergie de collision. Que ce soit avec UMAP, T-SNE ou la PCA, aucune distribution sous-jacente basée sur ces variables ne semblait présente. (figure 15).

(a) Colorié par temps de rétention

(b) Colorié par énergie de collision

FIGURE 15 – UMAP 3D de projections de spectres Massbank par MSBERT

Cela pourrait notamment s'expliquer par les différentes conditions expérimentales d'obtention des spectres de la base de données. En effet, chaque institution doit avoir ses propres protocoles expérimentaux de mesure des spectres, ce qui biaise la donnée et la rend plus difficilement comparable, contrairement aux données INERIS.

Au niveau des données des 4 mzML de l'INERIS, les prélèvements en rivière, les métadonnées comprennent l'énergie de collision et le temps de rétention, ici une structure sous-jacente commune aux 4 fichiers a pu être identifiée figure 16, plusieurs points sont à noter :

- Les spectres de 4 mzML semblent se superposer à quelques différences près (visible lorsqu'on sélectionne ou désélectionne les fichiers avec plotly). Les 4 échantillons proviennent d'eaux de rivière éloignées les unes des autres en France. Ils ont été analysés dans les mêmes conditions expérimentales.
- Les spectres avec une énergie de collision à 20 et à 40 semblent quasiment créer une symétrie.

- Les projections semblent suivre un ordre lié au temps de rétention, qui finalement est probablement lié au fait que les données soient prises en DIA et pas en DDA. Cela donne aux données une continuité par rapport aux molécules qui sortent du chromatographe au fil du temps.

(a) Coloré par temps de rétention

(b) Coloré par énergie de collision

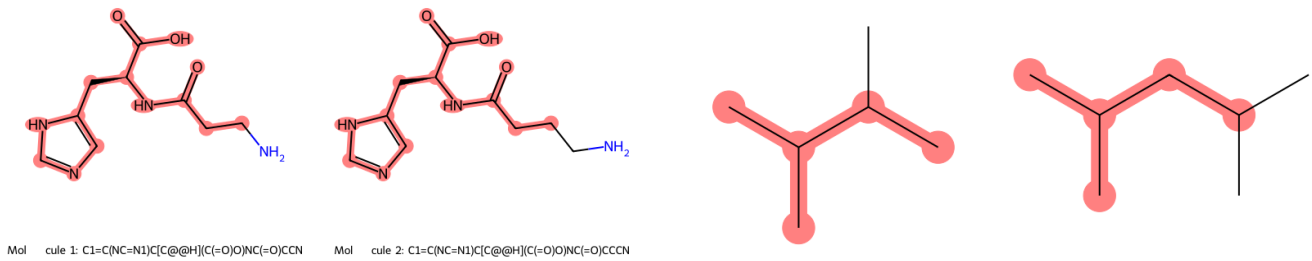
FIGURE 16 – UMAP 3D de projections de spectres de mzML INERIS (eau de rivière) par MSBERT

6.2.2 Algèbre dans l'espace latent

En plus de pouvoir améliorer l'analyse NTS, MSBERT porte une autre promesse, celle d'avoir un espace latent rationnel dans lequel certaines opérations vectorielles pourraient avoir du sens. Dans l'article, la substitution d'un groupement sur une molécule est abordée, le but est de vérifier si pour plusieurs molécules le vecteur associé à une transformation chimique est similaire. Quelques tests ont été réalisés durant le stage mais cette hypothèse nécessite une recherche plus approfondie. Deux expériences ont été réalisées :

Modification de la chaîne carbonée :

483 paires de spectres (x, y) ont été sélectionnées dans Massbank telles que y soit la même molécule que x mais avec un atome de carbone en plus n'importe où dans la chaîne.



(a) (b)

FIGURE 17 – Exemples de sous paires de molécules pour l'expérience 1

On commence par calculer, pour chaque paire de molécules, la différence entre l'embedding de la molécule avec un carbone supplémentaire et celui de la molécule sans ce carbone. La moyenne de ces vecteurs de différence constitue le vecteur moyen "+C". Ce vecteur est ensuite ajouté à l'embedding de chaque molécule "sans C". On mesure ensuite, pour chaque paire, comment le score de similarité cosinus avec la molécule "+C" correspondante évolue. L'histogramme obtenu représente la distribution de ces variations de similarité : un décalage vers la droite indique que la transformation rapproche les embeddings, tandis qu'un décalage vers la gauche indique qu'elle les éloigne.

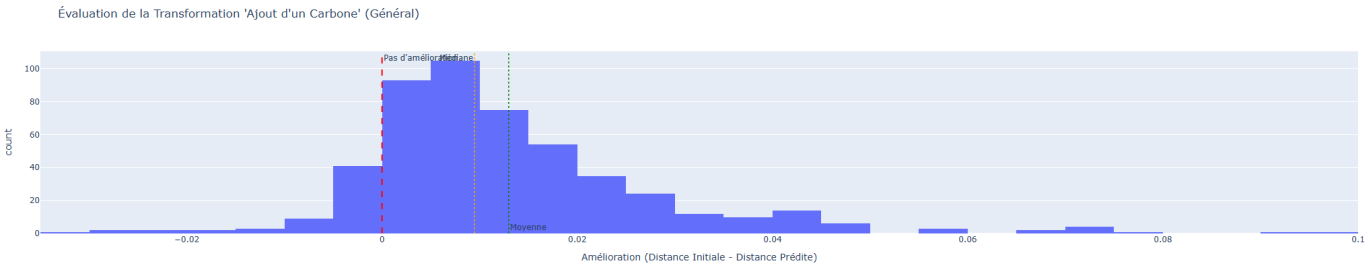
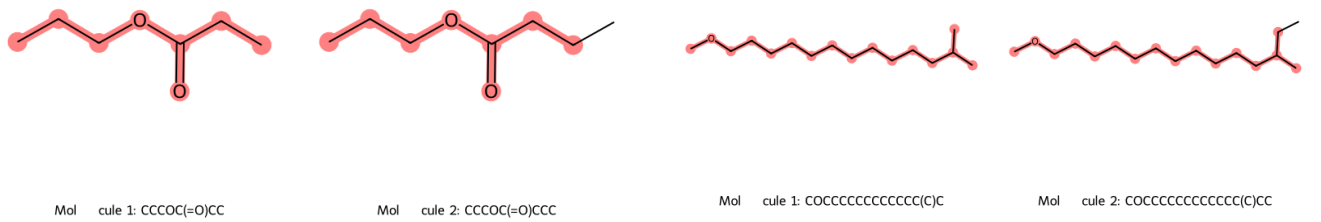


FIGURE 18 – Histogramme de la réduction de la distance cosinus pour l'expérience 1

La deuxième expérience est la même mais le carbone est uniquement positionné en suffixe de la chaîne carbonée de la molécule. L'idée était d'avoir moins de variations dans la nature chimique des transformations.



(a) (b)

FIGURE 19 – Exemples de paires de molécules pour l'expérience 2

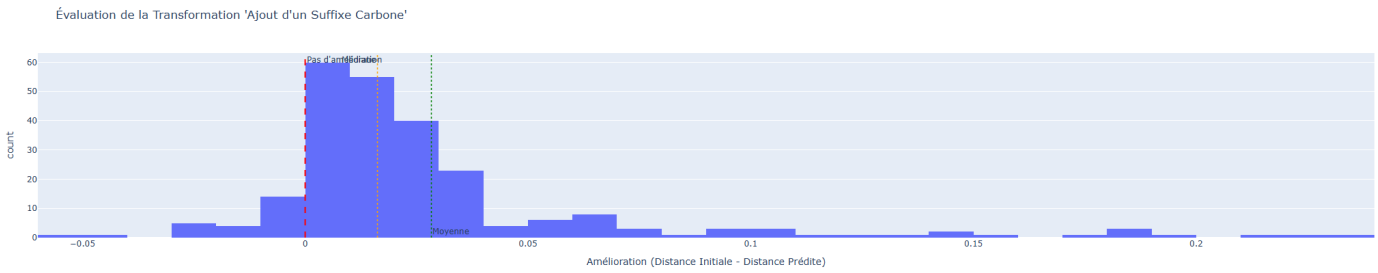


FIGURE 20 – Histogramme de la réduction de la distance cosinus pour l'expérience 2

voici une illustration concrète (figure 21) de ce que l'on essaie de faire, la molécule sans carbone supplémentaire est représentée par un losange et celle avec un carbone supplémentaire par un carré. Un vecteur carbone moyen parfait atteindrait à chaque fois le carré. Cette représentation permet de comprendre mais seuls les résultats des histogrammes sont significatifs, car même si on a utilisé la PCA pour garder la linéarité et avoir des vecteurs identiques qui ont la même direction, une comparaison fiable ne peut se faire que dans l'espace latent à 512 dimensions sans reprojections.

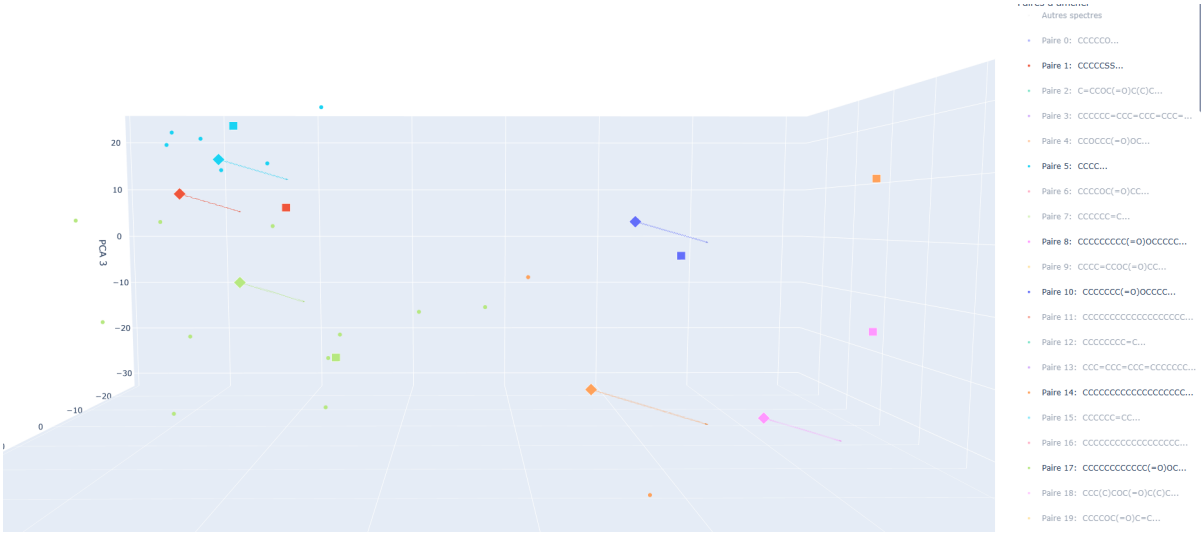


FIGURE 21 – Visualisation PCA d'un vecteur suffixe carbone moyen calculé sur 483 spectres Massbank

Les deux expériences révèlent une amélioration modeste mais avec une tendance positive. Il serait pertinent de concevoir des expériences plus spécifiques et chimiquement plus significatives que celles présentées pour obtenir des résultats plus probants. Dans la poursuite de ces travaux, il conviendrait de consulter l'expert métier afin de définir ses attentes précises sur cette question et de concevoir des tests plus ciblés. Cette approche pourrait permettre d'identifier des propriétés émergentes du modèle, notamment une structuration efficace de l'espace latent pour certains types de transformations chimiques spécifiques.

6.3 Application à la DIA

MSBERT est un modèle fondamentalement pensé pour représenter des spectres DDA. Lorsqu'il a été découvert que les données étaient acquises en DIA, il était trop tard pour changer complètement l'approche, même si des articles de recherche existent dans le domaine de la DIA comme DIA-BERT (LIU et al. 2025). Il a été décidé de poursuivre le travail avec MSBERT pour montrer l'intérêt de ce type de modèle.

La piste suivie a été de reprendre les solutions explorées par la solution SIAPARTNERS, la plateforme de NTS actuelle de l'INERIS. Comme on savait qu'ils utilisaient des scores de similarité de spectres assez simples sur les mêmes données que nous, ils devaient avoir trouvé une solution pour comparer des bases de

données spectrales avec des données acquises en DIA.

La première tâche a été de ré-implémenter le fonctionnement de la solution SIAPARTNERS. Avec La DDA, notre approche initiale aurait été de simplement projeter tous les spectres de l'échantillon ainsi que tous ceux de la base de données et de chercher les couples (x,y) avec x un spectre de l'échantillon et y un spectre de la base de données pour lesquels la similarité cosinus est la plus élevée. Comme les spectres des échantillons ne sont pas directement comparables aux bibliothèques, il faut faire une boucle dans le sens inverse, c'est-à-dire qu'il faut rechercher les substances de la base de données dans l'échantillon (voir l'étape d'extraction des raies dans le spectre ci-dessous).

Voici l'approche détaillée de la boucle :

1. **Recherche du précurseur dans les données MS1 :** L'analyse commence au niveau MS1. Pour une molécule cible donnée, on connaît la masse exacte (masse de précurseur), c'est la masse théorique de la molécule (somme des masses de tous les atomes, sans prendre d'isotope en compte) à laquelle on ajoute de l'ion utilisé pour l'ionisation, ici celle d'un proton H^+. Le système parcourt l'ensemble des scans MS1 du fichier `mzML` pour rechercher cette masse précise (avec une tolérance, par exemple ±0.003 m/z). Visuellement cela donne un *Extracted-Ion Chromatogram* (XIC) qui trace l'intensité de cette masse en fonction du temps de rétention.

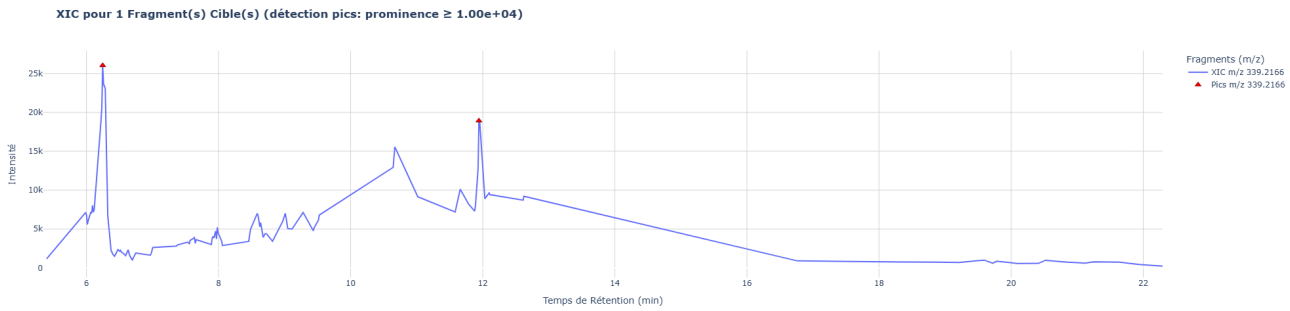


FIGURE 22 – XIC de recherche de la masse 339.216604 m/z (Piperonyl butoxide) dans le fichier Escaut

2. **Détection du temps de rétention expérimental :** L'algorithme détecte les pics dans le XIC qui est un signal à une dimension. Le sommet de chaque pic correspond à un temps de rétention expérimental auquel la masse cible (qu'on espère être la molécule cible) était à sa concentration maximale lors de son passage dans la colonne chromatographique.
   Dans la figure 22, on voit bien que la masse semble sortir tout au long de l'analyse, la réalité c'est qu'ici c'est le deuxième pic qui représente la sortie la plus probable de la molécule car un étalon de cette molécule a donné un temps de rétention d'un peu plus de 12 minutes, et ce cas se présente régulièrement.
   Dans un monde parfait, on aurait un XIC pour une masse précise (sans fenêtre) et qui afficherait un seul pic. En pratique, la première approche testée qui consiste à comparer les spectres de la base de données uniquement au spectre à l'APEX (point culminant du XIC) n'est pas viable, il faudra utiliser les spectres de chaque pic du XIC et donc pour chaque substance de la base de données, on obtiendra ainsi plusieurs identifications potentielles.
3. **Extraction des raies dans le spectre DIA :** Ce n'est qu'après avoir identifié un temps de rétention précis que la comparaison des spectres peut commencer. Les spectres MS2 à 20eV et à 40eV qui suivent le spectre MS1 dans l'échantillon vont être comparés à des spectres de références provenant pour l'instant de la base de données INERIS ou Agilent, et dans le futur à Massbank. Pour que ces spectres DIA et la bibliothèque (DDA) puissent être comparés, on va extraire dans le spectre de l'échantillon sélectionné uniquement les raies qui sont en face de celles du spectre recherché dans la bibliothèque avec une fenêtre à ±0.002 m/z. Ce processus permet de passer de la figure 23 à la figure 24.

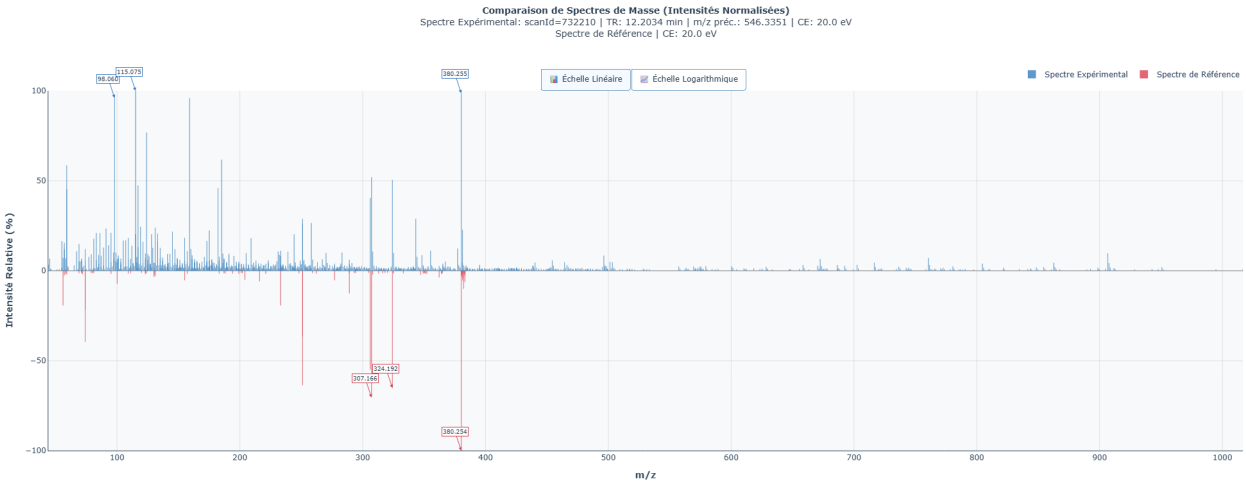


FIGURE 23 – Spectre d'échantillon brut

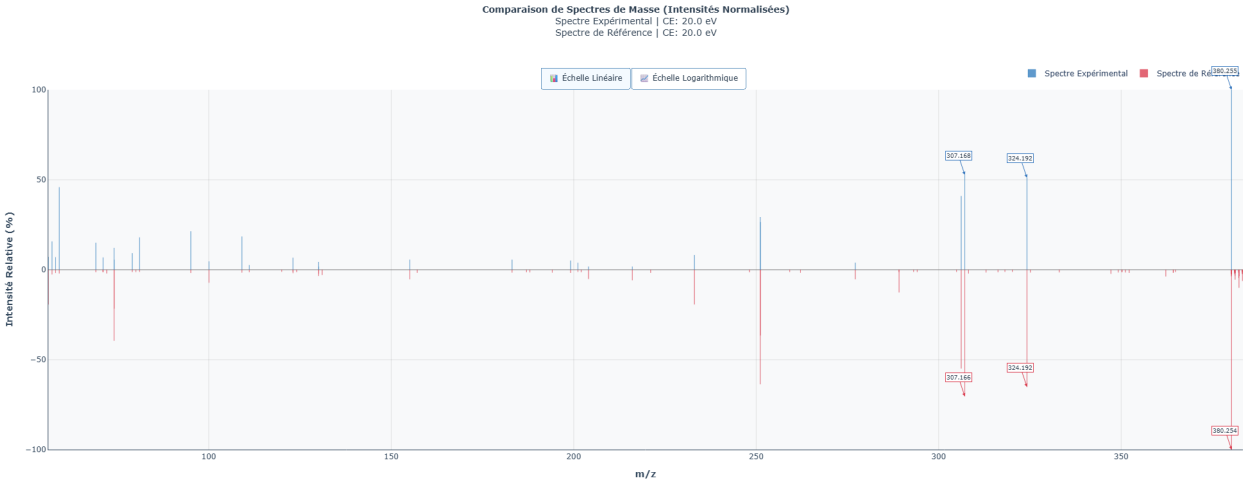


FIGURE 24 – Spectre d'échantillon filtré avec une fenêtre à ±0.002 m/z autour des pics de référence rouge)

Figures 23 et 24 : Spectre de Masse du Celiprolol de la Base INERIS (rouge) comparé au spectre de masse à l'APEX pour sa masse de précurseur dans l'échantillon (bleu).

4. **Calcul du score de similarité :** Le spectre DIA filtré est alors directement comparé au spectre de référence DDA (issu de MassBank ou de la base interne INERIS) de la molécule cible. Pour la comparaison de spectres expérimentaux avec les spectres de références, la solution SIAPARTNERS s'appuie sur plusieurs scores :

   - Score de similarité cosinus à 20eV et 40eV appliqués aux spectres de masses : une comparaison des intensités
   - Score de Scholle à 20eV et 40eV : une comparaison des intensités pondérée par les masses

5. **Comparaison MS1 et profil isotopique :** Les spectres MS1 eux vont être comparés à des profils isotopiques théoriques de la substance recherchée en utilisant les mêmes scores de similarité.
6. **Agrégation des scores pour classification binaire :** Cette étape qui n'a pas été ré-implémentée pendant le stage, consiste à prendre le maximum d'analyses antérieures pour lesquelles l'expert métier a identifié une substance dans un échantillon avec un niveau de confiance fixé à 1 c'est-à-dire le plus

fiable ou 0 pour non identifiée. Cela a fourni à SIAPARTNERS un échantillon d'un peu plus de 700 substances identifiées. À partir de cela les étapes précédentes ont permis de calculer des scores entre les spectres MS1 et profil isotopique, et entre les MS2 expérimentaux et de référence à 20 et 40 eV. Un modèle de Random Forest (RF) a ensuite été entraîné en classification binaire pour dire si oui ou non la molécule était présente à partir des scores de similarité des spectres.

L'apprentissage du modèle de la solution SIAPARTNERS est donc réalisé sur des données labellisées par un expert, il fait donc face à deux problèmes, le manque de temps pour annoter toutes les analyses et le risque de potentielles erreurs humaines.

Une fois cette approche ré-implémentée et testée pour voir si les scores étaient cohérents avec les fichiers de validation de données et les exports de la solution SIAPARTNERS. MSBERT a été intégré dans la boucle, pour chaque spectre MS1, et MS2, le score MSBERT est calculé entre les vecteurs d'embedding d'une substance expérimentale et d'une substance de référence. Pour donner plus de confiance dans les comparaisons des scores MSBERT avec les scores "traditionnels", d'autres scores ont été ajoutés parmi ceux qui étaient recommandés dans Degnan et al. 2023. Pour mieux visualiser, voici à quoi ressemblent les résultats de l'implémentation de la boucle :

nom_substance	cas	source	lision_e	apex_rt_experimental	rt_theoric	id_spectre_exp	masse_precurseur_r	score_msbert	score_cosine_greedy	score_cosine_greedy_we	score_ft_correlation	score_ams
Fluxapyroxad	907204-31-3	INERIS	0	11.671916666667	11.742	scanId=700323	382.097333	0.7507485151290894	0.9994388766267829	0.9994348146760241	0.9994388766267823	0.9355697783399012
Fluxapyroxad	907204-31-3	INERIS	20	11.671916666667	11.742	scanId=700580	382.097333	0.9877625107765198	0.9827102025700201	0.98204990124993O7	0.9827102025700198	0.7435346023669963
Fluxapyroxad	907204-31-3	INERIS	40	11.671916666667	11.742	scanId=700836	382.097333	0.9454529285430908	0.9469831484615285	0.9770901359538765	0.9469831484615285	0.8039649810755516
Chloridazon (PAC)	1698-60-8	Agilent	0	8.83025		scanId=529823	222.0428660614	0.7463668584823608	0.9995550572883918	0.9995492348282419	0.9984235761843293	0.9493946776109794
Chloridazon (PAC)	1698-60-8	Agilent	20	8.83025		scanId=530080	222.0428660614	0.9916096329689026	0.9959790311458196	0.9990312258053513	0.9337001668908487	0.8846589488714045
Chloridazon (PAC)	1698-60-8	Agilent	40	8.83025		scanId=530336	222.0428660614	0.9159589409828186	0.9869889533547461	0.977172631209497	0.9800430387103917	0.9127190837164229

Figure 25 – Première lignes d'un export CSV du dataframe de résultat de la fonction implémentée

Méthodologie

Certains points de méthode doivent être gardés en tête. MSBERT n'a été entraîné que sur des spectres MS2 en mode d'ionisation positif et en DDA, cependant, nous l'avons utilisé pour projeter des spectres MS2 DIA qui n'ont pas de masse de précurseur car tout est fragmenté, la masse de précurseur du spectre de référence est utilisée pour la projection du spectre de référence et du spectre expérimental.
Nous avons également projeté des spectres MS1 avec MSBERT, la masse de précurseur de la base de données de référence a aussi été fournie au modèle pour faire ces projections.

Des différences notables entre les scores MSBERT et les scores cosinus (cosinus weighted correspond à scholle) sont visibles. Les scores cosinus sont très tranchés, ces scores sont soit très hauts soit très bas ce qui fournit peu de nuance dans la comparaison des spectres. Le fait d'apprendre un modèle de RF sur ces scores comme dans la solution SIAPARTNERS peut donc être moins pertinent qu'avec un score nuancé car les seuils de décisions appris par les arbres de décisions de la forêt auront peu de marge d'erreur étant donné qu'un score à 0.991 représente une bien moins bonne similarité qu'un score à 0.998 par exemple. Les scores MSBERT montrent plus de nuances d'après l'expert métier, cela peut aussi rendre la prise de décision plus difficile. Bien que ce dernier score fournisse une information plus riche, son utilisation dépendra du côté pratique de son utilisation.

Calcul des scores de similarité. Voici les deux formules principales permettant de calculer les scores cosinus. Les scores AMS et par transformée de Fourrier (score de corrélation FT) ne seront pas traités car ils ont été implémentés (voir section 9.2) mais pas analysés. La similarité cosinus appliquée directement sur les vecteurs d'embedding produits par MSBERT, de dimension fixe 512, c'est celle qu'on appelle score MSBERT :

$$\text{cosinus}_{\text{MSBERT}}(\mathbf{u},\mathbf{v}) = \frac{\sum_{i=1}^{512} u_i\, v_i}{\sqrt{\sum_{i=1}^{512} u_i^2}\,\sqrt{\sum_{i=1}^{512} v_i^2}}.$$

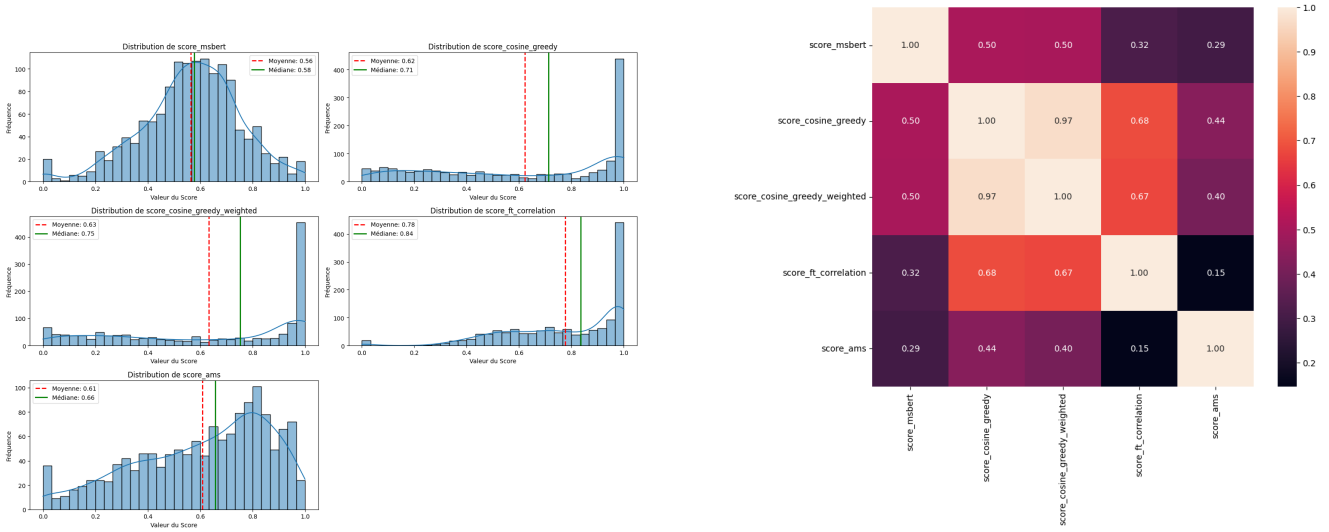
La seconde est le `cosine_greedy` pondéré (*weighted cosine*) de la librairie `matchms`, appliqué à des spectres de masse $(m/z, I)$. Après appariement *greedy* des pics des deux spectres avec une tolérance $\Delta m/z$, la

similarité pondérée par les masses est calculée par :

$$\text{cosinus}_{\text{greedy weighted}}(A,B) = \frac{\sum_{k\in\text{matches}} m_k \cdot I_{A,k}\, I_{B,k}}{\sqrt{\sum_{p\in A}(m_p \cdot I_{A,p})^2}\,\sqrt{\sum_{q\in B}(m_q \cdot I_{B,q})^2}},$$

où m_k désigne la valeur m/z du pic apparié et $I_{N,k}$ l'intensité associée à cette masse dans le spectre N. La version *non pondérée* correspond exactement à la même expression en supprimant les coefficients m_k. L'option *greedy* est privilégiée ici car la masse de précurseur n'existe pas dans les spectres DIA, ce qui rend inutile l'utilisation de variantes intégrant un *mass shift* calculé à partir de cette masse.

> Il est essentiel de comprendre que les scores cosinus présentés se calculent d'une manière similaire mais ne s'appliquent pas du tout sur les mêmes objets, utiliser les informations brutes du spectre ou utiliser les représentations vectorielles créées par MSBERT n'est pas du tout équivalent.



(a) distributions des scores (Escaut)

(b) matrice de corrélation des scores (Escaut)

FIGURE 26

Les scores MSBERT n'ont pas été comparés directement au score RF car ces derniers sont attribués pour chaque substance et prennent en compte les scores cosinus des spectres MS1 et MS2 à 20eV et 40eV, ce qui entraîne une dimension supplémentaire qui n'est pas présente dans les scores MSBERT.

Les figure 27 et figure 28 montrent des exemples de cas où les scores MSBERT et cosinus sont en désaccord. Il a été remarqué que MSBERT affiche régulièrement des scores bien plus bas que le cosinus dans le cas où beaucoup de raies de références (rouge) ne sont pas matchées car il n'y a aucune raie dans le spectre théorique qui correspond (avec la fenêtre recherche).
Cela s'explique par le fait que dans les implémentations des scores cosinus de la librairie Matchms, le score est maximisé en supprimant toute raie qui n'est pas matchée dans la référence comme dans l'échantillon. Les spectres finaux sur lesquels s'appliquent les formules de cosinus contiennent donc plus de raies de références non matchées lorsque MSBERT les compare.

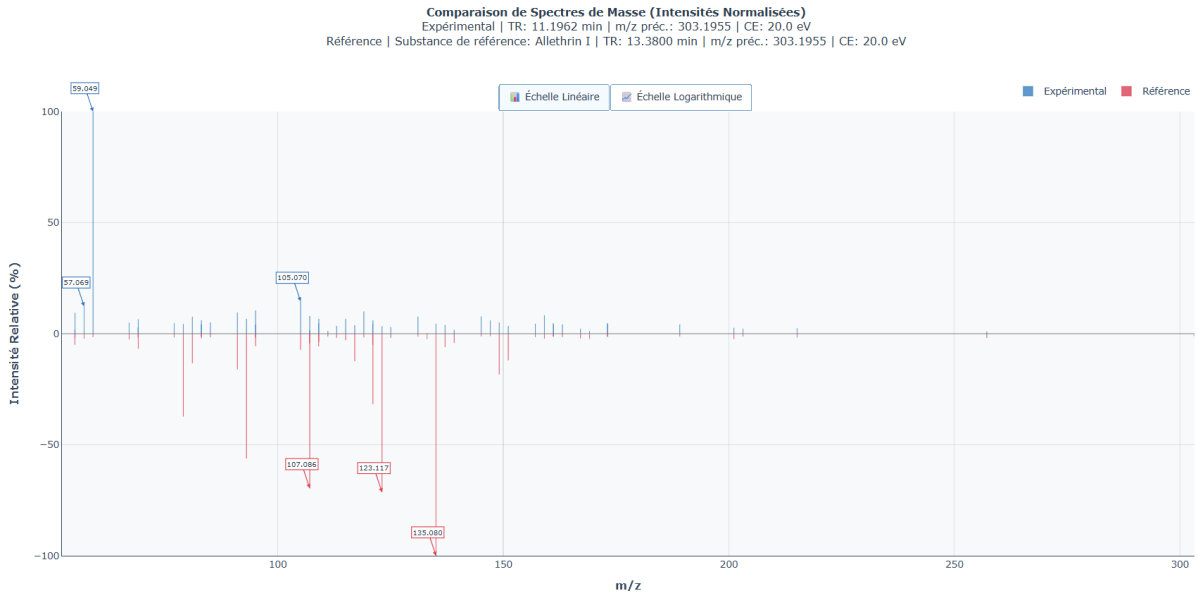


FIGURE 27 – Comparaison de spectres à 20eV où le score MSBERT (0.665) est largement supérieur au score cosinus (0.173)

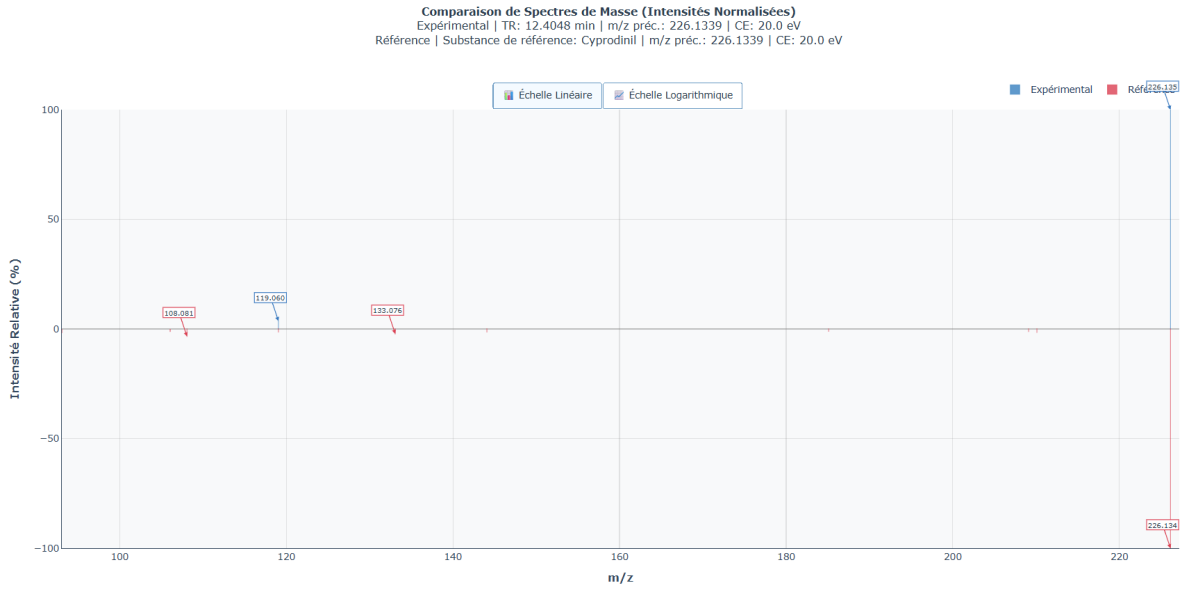


FIGURE 28 – Comparaison de spectres à 20eV où le score cosinus (0.998) est largement supérieur au score MSBERT (0.254)

7 Développement durable et responsabilité sociétale

Au-delà de ses aspects techniques, ce stage s'inscrit pleinement dans une démarche de développement durable et de responsabilité sociétale. Ce projet de recherche en constitue une application concrète et innovante, tournée vers l'avenir de la surveillance environnementale.

La contribution première de ce stage réside dans le développement d'une capacité d'analyse améliorée et accélérée de l'empreinte chimique dans les milieux aquatiques. En cherchant à automatiser l'identification de substances, les travaux menés visent directement à renforcer la protection de la ressource en eau et des écosystèmes. La capacité à détecter plus rapidement des polluants, notamment des contaminants émergents, est un enjeu majeur pour la santé publique et la préservation de la biodiversité.

Sur le plan de la responsabilité scientifique et technologique, la démarche adoptée durant ce stage a privilégié des outils et des approches relevant de la science ouverte (*open science*). Le choix de s'appuyer sur des modèles de fondation open-source comme MSBERT, et de les évaluer sur des bases de données publiques telles que MassBank DSFP (qui contient les échantillons INERIS), améliore la transparence, la reproductibilité et la pérennité des recherches. Cette approche favorise une dynamique collaborative au sein de la communauté scientifique.

Enfin, ce projet participe à une gestion plus durable des ressources au sein du laboratoire. L'automatisation d'une tâche aussi chronophage que l'interprétation des spectres de masse permet d'optimiser le capital humain. En libérant du temps aux experts du laboratoire ANAE, pour leur permettre de se concentrer sur des analyses à plus haute valeur ajoutée, comme l'interprétation de cas complexes ou la recherche sur des phénomènes de pollution inédits. Cette optimisation permet à l'INERIS de renforcer son efficacité et de faire évoluer son expertise face à des défis environnementaux toujours plus complexes.

En somme, ce projet illustre comment l'intelligence artificielle, appliquée de manière réfléchie, devient un levier stratégique pour renforcer les missions de service public au cœur de la transition écologique.

8 Conclusion et perspectives

L'adaptation de MSBERT aux données de l'INERIS a démontré la capacité des modèles de fondation à créer des représentations vectorielles cohérentes des spectres de masse, comme en témoignent les visualisations de clustering dans l'espace latent. Les fondations établies par cette preuve de concept ouvrent des perspectives prometteuses pour l'expert métier, et l'approche développée pour traiter les données DIA, inspirée de la méthodologie de la solution SIAPARTNERS, constitue une avancée pour améliorer les performances de l'analyse de suspects (NTS) au laboratoire ANAE.

Les livrables concrets de ce stage comprennent une implémentation fonctionnelle de MSBERT adaptée aux données spectrales de l'INERIS, une méthodologie de traitement des spectres DIA pour la caractérisation de substances, des visualisations interactives permettant l'analyse approfondie des représentations latentes, une évaluation comparative démontrant le potentiel face à la solution SIAPARTNERS, et des pistes pour les futurs stages qui permettront de poursuivre ces recherches. Le code est rendu accessible via un module Python installable et un notebook de présentation du travail utilisant ce module a été produit, il synthétise le travail effectué durant le stage. Ainsi, la reprise du projet et de la base de code sera plus efficace.

8.1 Défis surmontés et apprentissages

Ce stage orienté recherche et développement a présenté plusieurs défis techniques stimulants qui ont conduit à des résultats prometteurs. La transition réussie des spectres DDA vers les spectres DIA plus riches en informations et le développement d'approches de débruitage, en utilisant des fichiers de blancs sont deux pistes qui méritent d'être approfondies. Les résultats encourageants obtenus avec MSBERT, entraîné uniquement sur des spectres MS2 issus de DDA, ont révélé l'intérêt d'explorer des modèles plus adaptés comme DIA-BERT ou des variantes de l'architecture MSBERT plus adaptées à une comparaison de spectres bruts en remplaçant par exemple la masse de précurseur en entrée par le temps de rétention.

Ce stage m'a permis d'acquérir une compréhension approfondie des modèles de fondation, et particulièrement des architectures basées sur BERT. Cette maîtrise des architectures Transformer, qui donnent des performances à l'état de l'art dans d'innombrables domaines, constituera un atout précieux pour mes projets futurs.

Les apprentissages techniques incluent la maîtrise du développement sur environnement distant avec Onyxia, l'utilisation efficace de formats optimisés comme Parquet pour le stockage de données, une connaissance approfondie de PyTorch et la découverte de Keras 3, ainsi que la création de visualisations

interactives avec Plotly. Comparativement à mes projets précédents centrés sur l'image et la vidéo, ce stage m'a permis de développer de nouvelles compétences sur les signaux spectraux dont la compréhension et le traitement présentent des défis analytiques originaux, enrichissant ainsi ma palette de compétences en traitement du signal et en Machine Learning.

8.2 Perspectives de recherche

Plusieurs axes d'amélioration prometteurs se dégagent de ce travail. L'optimisation du traitement des données offre des opportunités concrètes à travers l'amélioration de la sélection des raies spectrales et le nettoyage rigoureux des échantillons à l'aide des blancs terrains POCIS mais aussi des blancs de laboratoire qui n'ont pas encore été exploités. L'agrégation intelligente des scores aux différentes énergies de collision (0, 20 et 40 eV) constitue une piste d'optimisation nécessaire pour faciliter la prise de décision des experts. L'intégration d'informations physico-chimiques pertinentes (pKa, polarité, solubilité dans l'eau) dans les bases de données permettra peut-être d'améliorer la pertinence des caractérisations et la validation des temps de rétention par l'expert métier.

L'augmentation de l'utilisation des données DIA, qui génèrent des informations plus riches que la DDA traditionnelle, ouvre la voie à l'exploration de modèles spécialisés pour la DIA comme DIA-BERT, l'objectif serait que l'intégralité du processus de NTS puisse être appris avec des techniques de Machine Learning ce qui permettrait d'améliorer l'efficacité de l'ensemble du processus avec l'augmentation de la qualité et la quantité des données. Pour le moment les étapes permettant de débruiter ou de sélectionner des raies pour comparer les spectres DIA aux spectres DDA sont faites "algorithmiquement" mais c'est peut-être là qu'une marge d'amélioration se cache.

L'exploration des Graph Neural Networks (GNN) pour représenter les données spectrales sous forme de graphes permettrait d'intégrer naturellement des informations supplémentaires comme les temps de rétention et les énergies de collision, offrant une représentation plus complète et robuste. De plus, des liens entre les spectres pourraient être envisagés pour que le modèle gère lui-même plusieurs spectres MS1 et MS2 à différentes énergies de collision.

L'utilisation des représentations MSBERT pour la détection d'anomalies dans les prélèvements environnementaux constitue une perspective pouvant apporter une réelle plus-value dans le quotidien de l'expert métier. Cette approche permettrait d'identifier automatiquement les spectres anormaux et de prioriser efficacement l'analyse des substances potentiellement dangereuses. La stratégie d'analyse non-ciblée représente un horizon de développement majeur particulièrement prometteur : en utilisant les embeddings MSBERT pour projeter les spectres dans un espace latent et en appliquant des méthodes de clustering, il devient possible d'isoler des groupes de spectres similaires aux substances toxiques connues, même non répertoriées dans les bases de données, toute autre caractéristique des molécules pourrait faire l'objet d'un clustering ou d'une classification dans l'espace latent.

Une perspective qui peut être testée rapidement consiste à enrichir l'architecture MSBERT pour créer un modèle de fondation spécialement conçu pour reconnaître tous types de spectres (MS1 et MS2) en intégrant harmonieusement des informations contextuelles pertinentes : énergie de collision, temps de rétention, et autres paramètres expérimentaux jugés utiles par l'expert métier. Cette approche permettrait potentiellement de créer des représentations plus riches et plus robustes aux variations expérimentales.

Les fondations posées durant ces trois mois constituent une base pour les développements futurs de l'INERIS dans le domaine de l'intelligence artificielle appliquée à la chimie analytique. L'approche développée ouvre des perspectives pour transformer les capacités de NTS et de NTA du laboratoire ANAE, avec un impact potentiel sur la surveillance environnementale et la détection de contaminants émergents.

Références

Beck, Armen G. et al. (juin 2024). “Recent Developments in Machine Learning for Mass Spectrometry”. en. In : *ACS Measurement Science Au* 4.3, p. 233-246. ISSN : 2694-250X, 2694-250X. DOI : 10.1021/acsmeasuresciau.3c00060. URL : https://pubs.acs.org/doi/10.1021/acsmeasuresciau.3c00060 (visité le 03/06/2025).

Bui-Thi, Danh et al. (mai 2024). “TransExION : a transformer based explainable similarity metric for comparing IONS in tandem mass spectrometry”. en. In : *Journal of Cheminformatics* 16.1, p. 61. ISSN : 1758-2946. DOI : 10.1186/s13321-024-00858-5. URL : https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00858-5 (visité le 03/06/2025).

Bushuiev, Roman et al. (mai 2025). “Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS”. en. In : *Nature Biotechnology*. ISSN : 1087-0156, 1546-1696. DOI : 10.1038/s41587-025-02663-3. URL : https://www.nature.com/articles/s41587-025-02663-3 (visité le 03/06/2025).

Degnan, David J. et al. (oct. 2023). “Characterizing Families of Spectral Similarity Scores and Their Use Cases for Gas Chromatography–Mass Spectrometry Small Molecule Identification”. en. In : *Metabolites* 13.10, p. 1101. ISSN : 2218-1989. DOI : 10.3390/metabo13101101. URL : https://www.mdpi.com/2218-1989/13/10/1101 (visité le 04/08/2025).

Hansen, Michael Edberg et Jørn Smedsgaard (août 2004). “A new matching algorithm for high resolution mass spectra”. en. In : *Journal of the American Society for Mass Spectrometry* 15.8, p. 1173-1180. ISSN : 1044-0305. DOI : 10.1016/j.jasms.2004.03.008. URL : https://pubs.acs.org/doi/10.1016/j.jasms.2004.03.008 (visité le 12/08/2025).

Horai, Hisayuki et al. (juill. 2010). “MassBank : a public repository for sharing mass spectral data for life sciences”. en. In : *Journal of Mass Spectrometry* 45.7, p. 703-714. ISSN : 1076-5174, 1096-9888. DOI : 10.1002/jms.1777. URL : https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jms.1777 (visité le 08/08/2025).

Hupatz, Henrik et al. (jan. 2025). “Critical review on in silico methods for structural annotation of chemicals detected with LC/HRMS non-targeted screening”. en. In : *Analytical and Bioanalytical Chemistry* 417.3, p. 473-493. ISSN : 1618-2642, 1618-2650. DOI : 10.1007/s00216-024-05471-x. URL : https://link.springer.com/10.1007/s00216-024-05471-x (visité le 03/06/2025).

Jin, Zhuo-Lin et al. (août 2025). “Application of machine learning in LC-MS-based non-targeted analysis”. en. In : *TrAC Trends in Analytical Chemistry* 189, p. 118243. ISSN : 01659936. DOI : 10.1016/j.trac.2025.118243. URL : https://linkinghub.elsevier.com/retrieve/pii/S0165993625001116 (visité le 03/06/2025).

Jonge, Niek de, Justin J. J. van der Hooft et Daniel Probst (fév. 2025). *To Bin or not to Bin : Alternative Representations of Mass Spectra*. en. arXiv :2502.10851 [cs]. DOI : 10.48550/arXiv.2502.10851. URL : http://arxiv.org/abs/2502.10851 (visité le 15/07/2025).

Liu, Zhiwei et al. (avr. 2025). “DIA-BERT : pre-trained end-to-end transformer models for enhanced DIA proteomics data analysis”. en. In : *Nature Communications* 16.1. Publisher : Springer Science and Business Media LLC. ISSN : 2041-1723. DOI : 10.1038/s41467-025-58866-4. URL : https://www.nature.com/articles/s41467-025-58866-4 (visité le 17/07/2025).

Oberacher, Herbert et al. (déc. 2020). “A European proposal for quality control and quality assurance of tandem mass spectral libraries”. en. In : *Environmental Sciences Europe* 32.1, p. 43. ISSN : 2190-4707, 2190-4715. DOI : 10.1186/s12302-020-00314-9. URL : https://enveurope.springeropen.com/articles/10.1186/s12302-020-00314-9 (visité le 08/08/2025).

Wang, Mingxun et al. (août 2016). “Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking”. en. In : *Nature Biotechnology* 34.8, p. 828-837. ISSN : 1087-0156, 1546-1696. DOI : 10.1038/nbt.3597. URL : https://www.nature.com/articles/nbt.3597 (visité le 08/08/2025).

Zhang, Hailiang et al. (oct. 2024). “MSBERT : Embedding Tandem Mass Spectra into Chemically Rational Space by Mask Learning and Contrastive Learning”. en. In : *Analytical Chemistry* 96.42, p. 16599-16608. ISSN : 0003-2700, 1520-6882. DOI : 10.1021/acs.analchem.4c02426. URL : https://pubs.acs.org/doi/10.1021/acs.analchem.4c02426 (visité le 03/06/2025).

Résumé

Ce rapport présente les travaux menés lors d'un stage à l'Institut National de l'Environnement Industriel et des Risques (INERIS), axés sur l'application du Deep Learning à l'analyse de données de spectrométrie de masse. Face au défi que représente l'interprétation des volumes massifs de données générées par la chromatographie liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS) pour la surveillance environnementale, ce projet a exploré le potentiel des modèles de fondation, inspirés des avancées en traitement du langage naturel.

La démarche a débuté par une étude approfondie de l'état de l'art, identifiant des architectures de type Transformer telles que MSBERT comme des solutions de pointe pour apprendre des représentations sémantiques de spectres de masse. Un travail conséquent de collecte, de nettoyage et d'uniformisation de données hétérogènes a été mené, agrégeant des sources publiques (MassBank EU) et internes à l'INERIS (analyses de rivières, bases d'étalons).

Le cœur du projet a consisté à adapter et évaluer le modèle MSBERT. L'analyse de son espace latent a confirmé sa capacité à regrouper les molécules par similarité structurelle, ouvrant des perspectives pour l'analyse non-ciblée (NTA). Un défi majeur a été l'adaptation du modèle, initialement conçu pour des données DDA (Data-Dependent Acquisition), aux données DIA (Data-Independent Acquisition) spécifiques à l'INERIS. Une nouvelle méthodologie a été développée, combinant l'extraction de signaux pertinents des spectres DIA et le calcul d'un score de similarité basé sur les représentations vectorielles de MSBERT.

Les résultats constituent une preuve de concept démontrant que ce score offre une mesure de similarité plus nuancée et potentiellement plus robuste que les méthodes traditionnelles. Ce travail jette les bases d'un outil d'analyse de suspects (NTS) amélioré et ouvre des pistes de recherche prometteuses, notamment l'utilisation de modèles dédiés à la DIA et l'application à la détection d'anomalies chimiques dans les échantillons environnementaux.

Abstract

This report details the work conducted during an internship at the French National Institute for Industrial Environment and Risks (INERIS), focusing on the application of Deep Learning to mass spectrometry data analysis. Addressing the challenge of interpreting the massive volumes of data generated by Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) for environmental monitoring, this project explored the potential of foundation models, inspired by recent breakthroughs in Natural Language Processing.

The approach began with a thorough state-of-the-art review, identifying Transformer-based architectures like MSBERT as leading solutions for learning semantic representations of mass spectra. Significant effort was dedicated to collecting, cleaning, and standardizing heterogeneous data, merging public sources (MassBank EU) with internal INERIS datasets (river sample analyses, standard compound libraries).

The core of the project involved adapting and evaluating the MSBERT model. Analysis of its latent space confirmed its ability to cluster molecules by structural similarity, opening perspectives for non-target analysis (NTA). A major challenge was adapting the model, originally designed for Data-Dependent Acquisition (DDA), to the Data-Independent Acquisition (DIA) data specific to INERIS. A novel methodology was developed, combining the extraction of relevant signals from complex DIA spectra with the calculation of a similarity score based on MSBERT's vector representations.

The results provide a proof of concept demonstrating that this new score offers a more nuanced and potentially more robust similarity measure than traditional methods. This work lays the groundwork for an improved non-target screening (NTS) tool and opens promising research avenues, including the use of DIA-specific models and the application of these techniques for chemical anomaly detection in environmental samples.

9 Annexe

9.1 Transformer

Entraîner un Transformer repose sur le même principe que pour tout réseau de neurones : une propagation avant (*forward pass*) suivie d'une rétropropagation (*backpropagation*) afin d'ajuster les paramètres pour minimiser une fonction de perte. La différence majeure réside dans la structure interne : en plus des couches entièrement connectées classiques, chaque bloc Transformer contient un **mécanisme d'attention multi-tête** qui apprend trois matrices de projection — W_Q, W_K et W_V — servant à transformer les embeddings d'entrée en vecteurs *Query*, *Key* et *Value*.

Ces matrices sont **spécifiques à chaque bloc** : si le modèle comporte N blocs, on a N ensembles distincts de (W_Q, W_K, W_V), optimisés indépendamment. Elles sont entraînées en même temps que les autres paramètres via la rétropropagation.

9.1.1 Attention

Définition des Vecteurs d'Entrée

Soient x_1, x_2, x_3 les vecteurs d'embedding correspondant aux trois tokens de notre séquence d'entrée. Chaque vecteur x_i est un vecteur de dimension d_{model} :

$$x_1 = (x_{1,1}, x_{1,2}, \ldots, x_{1,d_{\text{model}}})$$
$$x_2 = (x_{2,1}, x_{2,2}, \ldots, x_{2,d_{\text{model}}})$$
$$x_3 = (x_{3,1}, x_{3,2}, \ldots, x_{3,d_{\text{model}}})$$

Chaque $x_i \in \mathbb{R}^{1 \times d_{\text{model}}}$. Ces vecteurs sont typiquement issus d'une couche d'embedding et augmentés d'un encodage positionnel (détaillé plus loin).

Dérivation des Vecteurs Query

Pour chaque vecteur d'entrée x_i (c'est-à-dire chaque ligne de la matrice X, correspondant souvent à l'embedding d'un token), on dérive un triplet de vecteurs spécifiques :

- Une **Query** $q_i = x_i W_Q$ ex : Ce que ce livre cherche/demande
- Une **Key** $k_i = x_i W_K$ ex : Ce que ce livre offre/propose
- Une **Value** $v_i = x_i W_V$ ex : L'information à extraire de ce livre

Les matrices $W_Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, et $W_V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ sont des matrices de poids qui sont apprises pendant l'entraînement du modèle. Q, K et V représentent la même information initiale mais différemment, l'analogie proposée dans les exemples ci-dessus n'est pas parfaite et il n'est pas nécessaire de chercher un sens aussi précis pour ces matrices.

Construction des Matrices

En pratique, les vecteurs d'embedding sont empilés pour former la matrice d'entrée X pour faire des calculs matriciels/tensoriels. Pour une séquence de n tokens, la matrice X est de dimension $n \times d_{\text{model}}$:

$$X = \begin{pmatrix} — & x_1 & — \\ — & x_2 & — \\ — & x_3 & — \end{pmatrix}$$

C'est cette matrice X qui sera utilisée comme base pour calculer les matrices Query (Q), Key (K), et Value (V) dans le mécanisme d'attention :

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

où $W_Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, et $W_V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ sont des matrices de poids apprises.
Les matrices Q, K, et V sont donc constituées des vecteurs q_i, k_i, et v_i, où $q_i = x_i W_Q$, $k_i = x_i W_K$, et $v_i = x_i W_V$. Pour notre exemple avec $n = 3$ tokens, ces matrices se présentent comme suit :

$$Q = \begin{pmatrix} — & q_1 & — \\ — & q_2 & — \\ — & q_3 & — \end{pmatrix} \quad K = \begin{pmatrix} — & k_1 & — \\ — & k_2 & — \\ — & k_3 & — \end{pmatrix} \quad V = \begin{pmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{pmatrix}$$

Formule de l'Attention (une Tête d'Attention)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$\text{où : } Q = XW_q, \quad K = XW_k, \quad V = XW_v \tag{2}$$

Pourquoi diviser par $\sqrt{d_k}$?

Le **scaling factor** $\sqrt{d_k}$ évite que les scores d'attention deviennent trop grands, ce qui pousserait le softmax vers des distributions très "pointues" (gradients faibles). C'est essentiel pour la stabilité d'entraînement.

9.1.2 Attention Masquée

La différence entre une couche d'attention multi-tête normale et la version masquée est simple, pour obtenir la version masquée on veut donner une attention nulle aux tokens futurs, donc en entrée de la fonction softmax on fournit -Inf. Cela est principalement nécessaire pour les modèles génératifs car les tokens suivants n'ont pas encore été générés. En revanche, dans un modèle de type BERT, l'objectif est d'avoir accès à l'entièreté de la séquence, permettant ainsi à chaque token d'être contextualisé de manière bidirectionnelle.

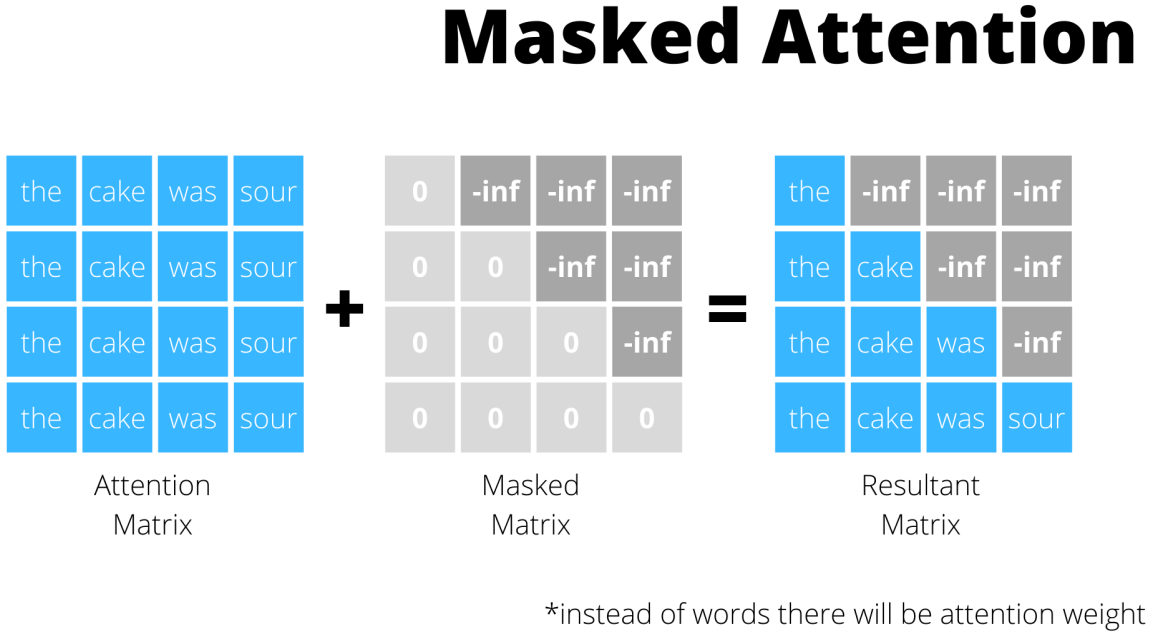


FIGURE 29 – (source krypticmouse)

9.2 Informations supplémentaires sur les scores

Le score AMS (*Accurate Mass Similarity*) repose sur la distance de Jeffreys–Matusita entre les distributions normalisées d'intensité, pondérée par la différence de m/z. On définit :

$$D_{\mathrm{AMS}}(A,B) = \frac{1}{n_A + n_B} \sqrt{\sum_{(i,j)\in \mathrm{assign}} \left(\sqrt{\hat{I}_{A,i}} - \sqrt{\hat{I}_{B,j}}\right)^2 + \frac{\left(\sqrt{m_{A,i}} - \sqrt{m_{B,j}}\right)^2}{w_{mz}}}$$

où (i,j) sont les paires de pics optimisant l'appariement, $\hat{I}$ les intensités normalisées, m les valeurs m/z et w_{mz} un facteur de pondération. La similarité AMS est ensuite donnée par :

$$\mathrm{AMS}(A,B) = e^{-\gamma \cdot D_{\mathrm{AMS}}(A,B)},$$

avec γ un facteur d'échelle.

L'algorithme permettant d'implémenter ce score provient de HANSEN et SMEDSGAARD 2004 son intérêt c'est qu'il ne nécessite aucun binning contrairement à tous les autres scores.

Le score de corrélation FT compare deux spectres de masse après discrétisation sur une grille m/z commune On transforme ainsi chaque spectre en un vecteur d'intensités $\mathbf{x}$ et $\mathbf{y}$ dont chaque case correspond à un intervalle m/z. Ces vecteurs sont ensuite comparés dans le domaine fréquentiel, en appliquant la transformée de Fourier discrète $\mathcal{F}(\cdot)$, ce qui permet de calculer efficacement la corrélation croisée entre tous les décalages possibles de pics.

En posant $\mathbf{X} = \mathcal{F}(\mathbf{x})$ et $\mathbf{Y} = \mathcal{F}(\mathbf{y})$, la corrélation croisée dans le domaine fréquentiel s'écrit :

$$\mathbf{C} = \mathcal{F}^{-1}(\mathbf{X} \cdot \overline{\mathbf{Y}}),$$

où $\bar{\cdot}$ désigne le conjugué complexe et $\mathbf{C} = (C_k)_{k=0,\ldots,N-1}$ le vecteur des corrélations pour chaque décalage de grille. Le score de corrélation FT correspond alors au rapport du maximum de corrélation croisée sur la moyenne géométrique des maxima d'auto-corrélations :

$$\mathrm{FT\_corr}(\mathbf{x},\mathbf{y}) = \frac{\max_k |C_k|}{\sqrt{\max_k \left|\mathcal{F}^{-1}(\mathbf{X}\cdot\overline{\mathbf{X}})_k\right| \cdot \max_k \left|\mathcal{F}^{-1}(\mathbf{Y}\cdot\overline{\mathbf{Y}})_k\right|}}.$$

Cette approche permet de mesurer la similarité des profils d'intensité entre spectres, tout en tolérant un léger décalage global des pics, ce qui peut se produire en raison de variations instrumentales.

Pour tous les calculs de similarité (autres que MSBERT), nous avons utilisé les réglages suivants :

- **Cosine Greedy (non pondéré)** : `tolerance` $= 0.01$, `mz_power` $= 0.0$
- **Cosine Greedy (pondéré)** : `tolerance` $= 0.01$, `mz_power` $= 1.0$
- **Corrélation FT** : taille de bin $\Delta = 0.005$ m/z
- **AMS** : facteur de pondération $w_{mz} = 0.1$, facteur d'échelle $\gamma = 25$

9.3 Courbes De Loss entraînement VAE

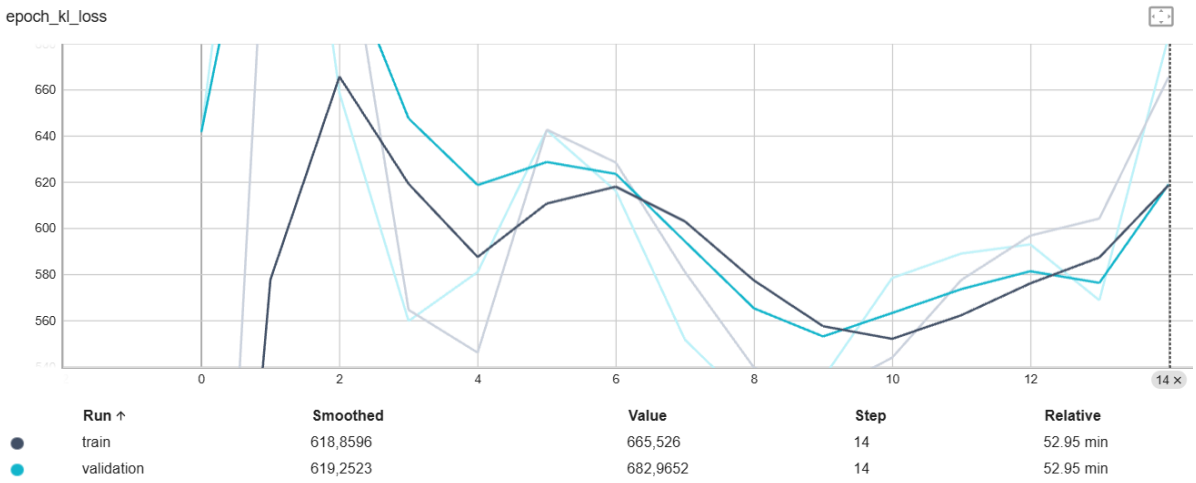


FIGURE 30 – Courbes d'apprentissage de la loss de kullback leibler

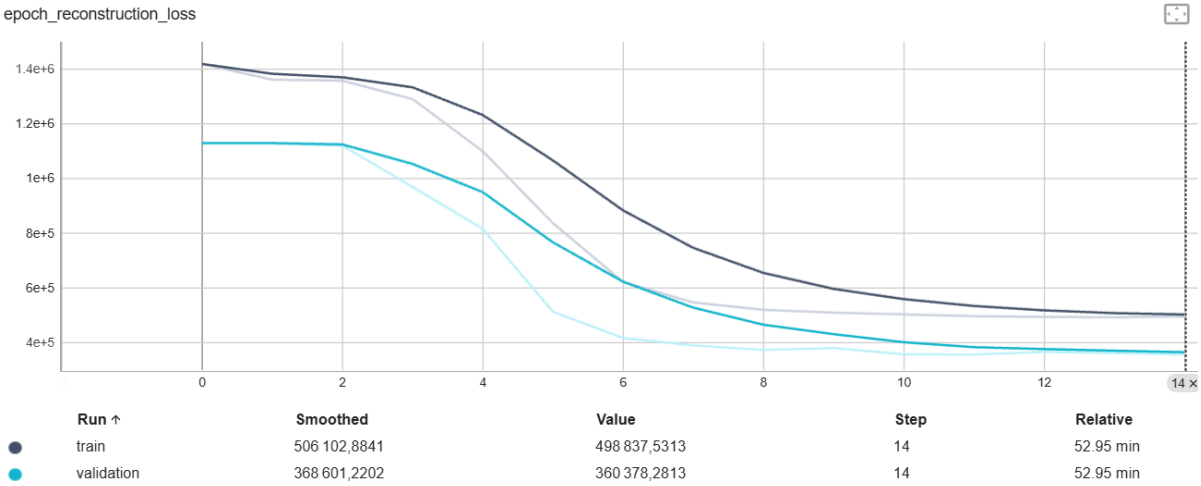
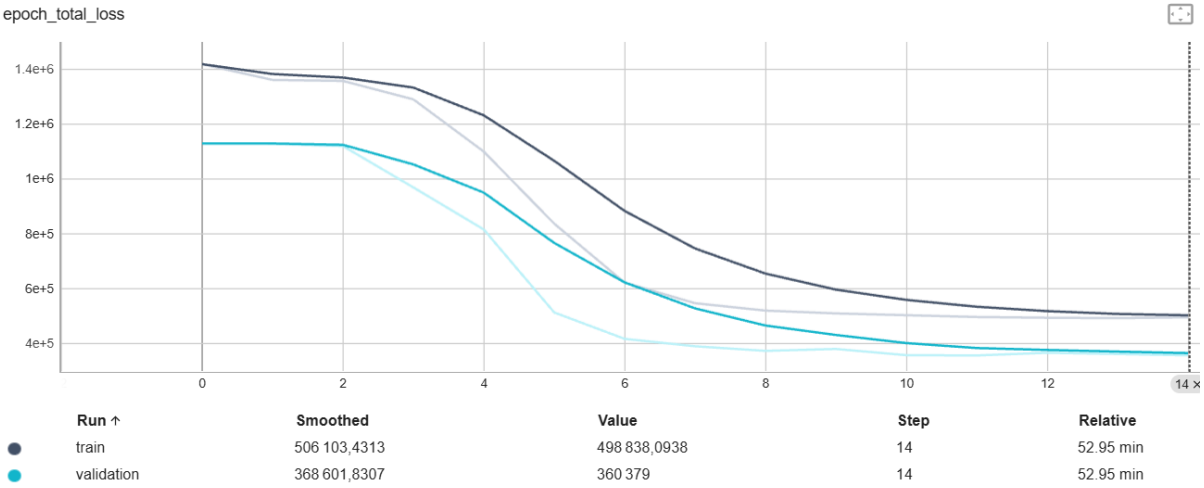


FIGURE 31 – Courbes d'apprentissage de la loss de reconstruction (Mean Square Error)



Run ↑	Smoothed	Value	Step	Relative
train	506 103,4313	498 838,0938	14	52.95 min
validation	368 601,8307	360 379	14	52.95 min

FIGURE 32 – Courbes d'apprentissage de la loss composite