Développement d'outils de chimiométrie pour le suivi de contamination aux composés aromatiques polycycliques dans des matrices environnementales complexes

Merzouk Haouchine

Laboratoire Interdisciplinaire des Environnements Continentaux

LIEC UMR 7360 CNRS, Université de Lorraine

Ecole doctorale Sciences et Ingénierie des Ressources Naturelles - SIRENA

THESE DE DOCTORAT

Présentée en vue de l'obtention du titre de

Docteur de l'Université de Lorraine en Géosciences

Merzouk HAOUCHINE

Développement d'outils de chimiométrie pour le suivi de contamination aux composés aromatiques polycycliques dans des matrices environnementales complexes

Soutenance publique le 04 octobre 2023 devant le jury :

M. Ludovic Duponchel Pr, LASIRE, Université de Lille	Rapporteur
Mme. Edith Parlanti Cr CNRS, EPOC, Université de Bordeaux	Rapportrice
Mme. Véronique Sadtler Pr, LRGP, Université de lorraine	Examinatrice (Présidente du jury)
M. Xavier Luciani MdC HdR, Université de Toulon	Examinateur
M. Pierre Faure-Catteloin Dr CNRS, LIEC, Université de Lorraine	Directeur de thèse
M. Marc Offroy MdC, LIEC, Université de Lorraine	Co-directeur de thèse
Mme. Coralie Biache Cr CNRS, LIEC, Université de Lorraine	Invitée

Remerciements

Pour commencer, je tiens à remercier et à exprimer ma profonde gratitude à mes directeurs de thèse : **Pierre Faure** et **Marc Offroy** ainsi qu'à ma co-encadrante **Coralie Biache** pour leur précieuse contribution à la réalisation et à la réussite de ma thèse. Tout simplement, merci pour ces 3 belles années de thèse, merci pour votre patience et votre appui. Merci également pour votre implication et votre disponibilité. J'adresse un remerciement particulier à **Marc** pour nos débats scientifiques et philosophiques, parfois animés ! Je tiens à te remercier pour tous tes conseils concernant mes futurs choix professionnels, bien que, 'choisir, c'est renoncer' 😉. Je tiens également à te remercier de m'avoir donné l'opportunité d'intervenir dans les enseignements des méthodes numériques à l'ENSIC. Cela m'a permis de découvrir le monde de l'enseignement supérieur qui va de pair avec le monde de la recherche scientifique.

Je tiens à exprimer ma profonde gratitude envers les membres du jury, **Ludovic Duponchel, Edith Parlanti, Véronique Sadtler** et **Xavier Luciani,** pour l'honneur qu'ils m'ont fait en évaluant mon travail. Je vous remercie pour vos questions, vos commentaires, ainsi que pour vos suggestions, corrections et conseils. Je remercie également mes membres du CSI, **François Lesage** et **Laurent Jeanneau** pour leur accompagnement et leurs conseils mais aussi leurs encouragements durant ces 3 années de thèse.

Je remercie **Catherine Lorgeoux** de m'avoir accueilli avec bienveillance au sein de la plateforme de géochimie organique du laboratoire GéoRessources. J'adresse également mes remerciements à **Hermine Huot** et **Céline Caillet** pour notre collaboration fructueuse dans le cadre de l'analyse des sols par spectroscopie infrarouge. Mes remerciements vont aussi à **Mario Marchetti** pour notre travail commun sur l'identification Raman des polymorphes du carbonate de calcium dans le béton recyclé. En outre, je t'adresse mes remerciements ainsi qu'à **Marc** pour m'avoir donné l'opportunité d'intervenir dans les cours de chimiométrie destinés aux doctorants des écoles doctorales SIRENA et Paris-Est. Je tiens à remercier **Ludovic Duponchel** et **Myriam Moreau** du laboratoire de spectroscopie pour les interactions, la réactivité, et l'environnement (LASIRE) de l'université de Lille pour leur accueil et de m'avoir permis de réaliser les expérimentations Raman menées dans le cadre de ma thèse. Je tiens enfin à remercier toutes l'équipe enseignante de méthodes numériques à l'ENSIC pour leur accueil, leur bienveillance mais aussi leur bonne humeur et leur sens de l'humour.

Je remercie l'ensemble des personnes qui ont contribué, que ce soit de près ou de loin, à rendre mes 3 dernières années agréables et emplies de bonne humeur sur le plan personnel. Merci à mes chers collègues et amis **Rafael** et **Nicolas F**. Merci à mes chers amis **Lamine** et **Adel**. Merci à mes chers collègues du LIEC et de la plateforme de géochimie organique. Merci à tous !

Je remercie chaleureusement mes chers parents d'avoir contribué à forger l'homme que je suis aujourd'hui. C'est grâce à vous que j'ai pu atteindre ce point de ma vie. Vous avez toujours eu confiance en moi et m'avez toujours exprimé votre fierté à mon égard. À mon tour, je tiens à vous dire à quel point je suis extrêmement fier, heureux et reconnaissant envers vous. C'est pourquoi, je vous dédie ce travail en témoignage de mon amour et de ma reconnaissance. Je remercie chaleureusement mes chères petites sœurs d'avoir toujours été présentes. C'est un bonheur d'avoir des sœurs comme vous. Votre grand frère vous aime profondément.

Pour finir, je remercie tendrement ma chère épouse le docteur Sabrina Hamla/Haouchine pour son dévouement et sa patience. Comme je te l'ai toujours dit, tu incarnes à la perfection l'image de la femme kabyle que tu es, une véritable montagne de générosité et de patience « dadrar nesvar ». Je te remercie pour ta présence et ton soutien inébranlable. Merci infiniment, ma chérie !

Table des matières

Table des matières

Table des matières

Table des matières

Liste des figures

Introduction

CHAPITRE I

Liste des figures

CHAPITRE II

CHAPITRE III

Liste des figures

Liste des figures

CHAPITRE IV

Liste des figures

CHAPITRE V

Liste des figures

Liste des tableaux

INTRODUCTION

CHAPITRE II

CHAPITRE III

CHAPITRE IV

CHAPITRE V

Liste des tableaux

INTRODUCTION

Selon une étude réalisée par l'agence européenne de l'environnement '*European Environment Agency*' (EEA), portant sur 39 pays européens, plus de 2.8 millions de sites en Europe sont potentiellement concernés par une contamination locale du sol, dont environ 650 000 clairement identifiés[1]. Les données provenant de cette étude répartissent le niveau de pollution des sites européens suivant différentes catégories de contaminants. Elles montrent qu'en Europe et en France, les principaux polluants affectant les matrices solides (i.e. sol, boue et sédiments) sont les métaux lourds et les hydrocarbures (huiles minérales et hydrocarbures aromatiques polycycliques (HAP) - Figure 1). Les HAP, contaminants organiques de la famille des composés aromatiques polycycliques (CAP), impactent ainsi fortement les matrices solides en Europe (11%) et en France (9%)[2,3] (Figure 1).

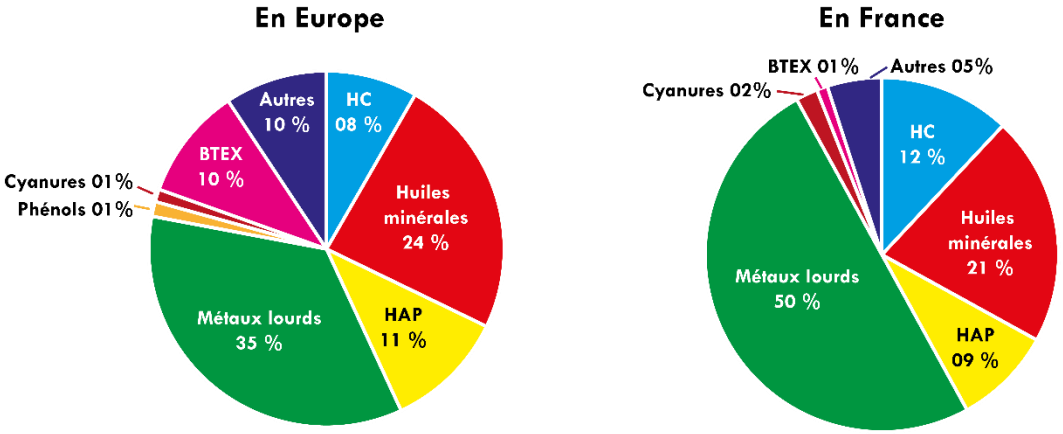


Figure 1: Contaminants affectant les matrices solides en Europe et en France. HC : Hydrocarbures chlorés ; BTEX : Benzène, toluène, Éthylbenzène, Xylènes ; HAP : Hydrocarbures aromatiques polycycliques[2].

Ces estimations constituent de bons indicateurs de la pollution mais elles doivent être considérées avec précaution, car les diagnostics actuels de la pollution des sols se basent principalement sur des substances réglementées. En effet, parmi les milliers de CAP et de HAP existants, seuls 16 HAP, sont considérés dans cette étude de l'EEA, car ils sont inclus dans la liste des polluants prioritaires à surveiller, établie par l'Agence Américaine de la Protection de l'Environnement (US-EPA)[4] et reprise, à travers le monde, par de nombreuses agences de l'environnement dont l'EEA. Certes, certains de ces HAP ont été sélectionnés d'une part, sur la base de leur toxicité et d'autre part, sur la base de leur caractère cancérogène suspecté[5]. Néanmoins, d'autres HAP sont inclus dans la liste pour des raisons historiques[6] (i) en lien avec leur utilisation fréquente en industrie comme par exemple dans la fabrication des colorants ou encore (ii) en lien avec les standards analytiques (i.e. étalons) disponibles, au moment de l'édition de la liste (Figure 2). Or il existe une grande diversité de composés organiques constituant les contaminations organiques, non répertoriée, en particulier les dérivées des HAP ou encore les CAP hétérocycliques et leurs dérivées, qui peuvent être tout autant voire même plus toxique[7] que les 16 HAP réglementaires.

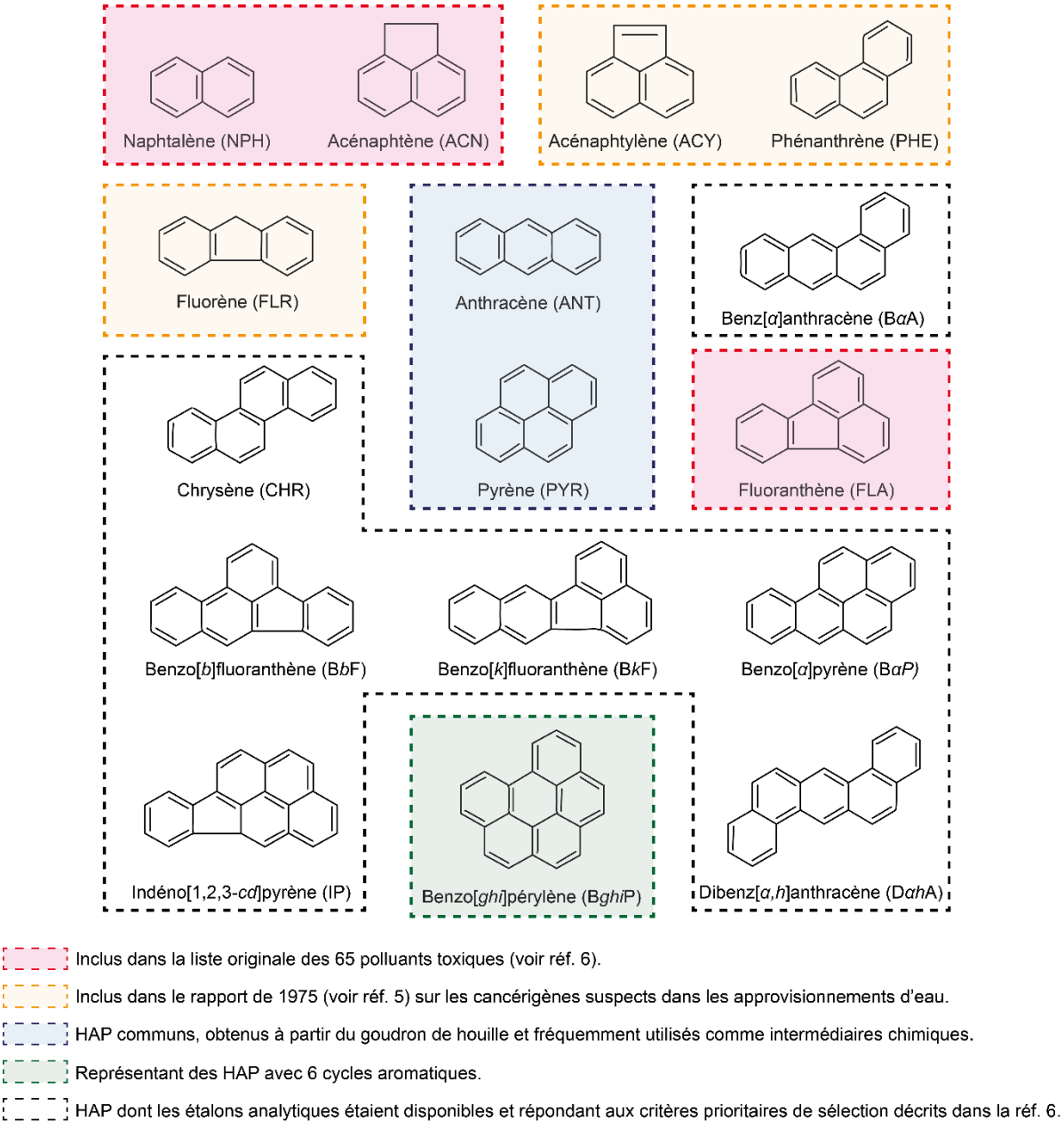
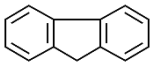
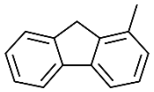
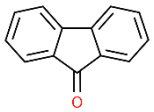
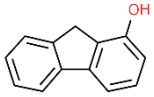
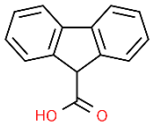
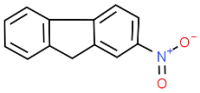
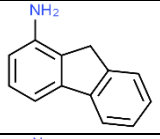
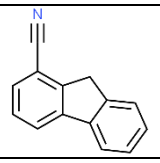
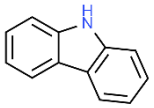


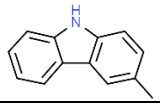
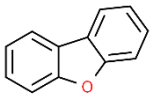
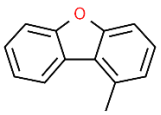
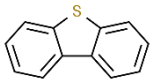
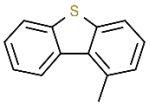
Figure 2 : Structures moléculaires des 16 HAP réglementaires inclus dans la liste de l'agence américaine de la protection de l'environnement (US-EPA) et raisons de leur inclusion dans la liste.

Il n'existe pas à l'heure actuelle de consensus établi quant à la classification des CAP. Néanmoins, dans ce travail de thèse, le choix s'est porté sur la classification proposée par C. Achten & J.T. Andersson[8] où les CAP sont classés en deux catégories principales, d'une part les hydrocarbures aromatiques polycycliques (HAP) et leurs dérivés, et d'autre part, les CAP hétérocycliques (NSO-CAP) et leurs dérivés. Chaque catégorie principale est caractérisée par des sous catégories relatives à la présence et au type de substitution (Tableau 1). Un CAP est une molécule organique constituée d'au moins deux cycles aromatiques condensés, le plus souvent benzéniques. Dans certains cas, des cycles à 5 ou à plus de 6 carbones ainsi que des hétérocycles peuvent y figurer. Par ailleurs, des alkyles ou des groupements fonctionnels peuvent se greffer à ces molécules pour former des dérivés[8]. Les HAP et leurs dérivés se

caractérisent par leurs cycles aromatiques fusionnés, ne contenant que des atomes de carbone et d'hydrogène[9]. La dérivation peut être une alkylation (i.e. substitution d'un hydrogène par un groupement alkyle), formant ainsi un HAP alkylé ou une substitution par un groupement fonctionnel formant par exemple une céto-HAP, un hydroxy-HAP, un nitro-HAP, un amino-HAP, une cyano-HAP ou un HAP acide[8]. Et enfin, contrairement aux HAP et leurs dérivés, les CAP hétérocycliques (i.e. NSO-CAP) intègrent au sein même de leurs structures cycliques, un hétéroatome d'azote, de soufre ou d'oxygène qui se substitue à un des carbones d'un des cycles aromatiques[10]. Quand il s'agit d'une substitution par un atome d'azote, on parle alors de la sous-catégorie des azaarènes (N-CAP). Dans le cas du soufre et de l'oxygène, il s'agit de thiazaarènes (S-CAP) et d'oxaarènes (O-CAP), respectivement[11]. Comme pour les HAP, il existe aussi des dérivés des NSO-CAP issus du greffage aux cycles de ces composés soit de groupements alkyles, soit de groupements fonctionnels.

Tableau 1 : Catégories principales et sous catégories des CAP. Les structures chimiques sont issues de la base de données Chemspider (https://chemspider.com).

Catégorie principale	Sous-catégorie	Nom d'un composé en guise d'exemple	Structure moléculaire
HAP et leurs dérivés	HAP	9H-Fluorène	
	HAP alkylés	1-Méthyl-9H-fluorène	
	Céto-HAP	9H-Fluorén-9-one	
	Hydroxy-HAP	9H-Fluorén-1-ol	
	HAP acides	Acide 9H-fluorène-9-carboxylique	
	Nitro-HAP	2-Nitro-9H-fluorène	
	Amino-HAP	9H-Fluorén-1-amine	
	Cyano-HAP	9H-Fluorène-1-carbonitrile	
CAP hétérocycliques	Azaarènes (N-CAP)	9H-Carbazole	

(NSO-CAP) et leurs dérivés	N-CAP alkylé	3-Méthyl-9H-carbazole	
	Oxaarènes (O-CAP)	Dibenzo[b,d]furane	
	O-CAP alkylé	1-Méthyldibenzo[b,d]furane	
	Thiazaarènes (S-CAP)	Dibenzo[b,d]thiophène	
	S-CAP alkylé	1-Méthyldibenzo[b,d]thiophène	

Plusieurs processus peuvent être à l'origine de la formation des CAP. Ils sont, souvent, générés lors de la combustion incomplète, de la pyrolyse et/ou de la diagenèse de la matière organique, ou encore, ils peuvent être, en moindre mesure, biosynthétisés par des plantes et des champignons[10,12,13]. Ils peuvent ainsi être d'origine naturelle mais restent essentiellement d'origine anthropique. Ils sont, d'une part, présents naturellement dans les sources pétrogéniques (i.e. pétrole brut et houille) et émis lors d'épisodes volcaniques et de feux de forêts. D'autres part, Ils sont largement produits et diffusés dans l'environnement via les différentes activités humaines exploitant, notamment, des ressources (e.g. combustibles) fossiles (i.e. chauffage, industrie, transports, etc.)[14–18]. De nos jours, les CAP sont omniprésents puisqu'ils peuvent être rencontrés dans l'air, les eaux de surface et souterraines, les sols et les végétaux (Figure 3). Ils sont en particulier très fréquemment rencontrés sur les sites d'anciennes cokeries ou d'usine à gaz, étant les produits majoritaires constituants les goudrons de houille, sous-produits de la cokéfaction[19]. La pollution aux CAP constitue ainsi une problématique environnementale majeure. De par leur propriété hydrophobe, ces produits présentent une forte toxicité, possèdent des capacités d'adsorption sur des particules organiques et/ou minérales qui leur permettent de migrer d'un compartiment environnemental à un autre (Figure 3). De plus, leur dégradation (e.g. photo-oxydation, biodégradation) peut conduire à la genèse de métabolites qui peuvent être encore plus toxiques, réactifs voire mobiles que les CAP originels[20].

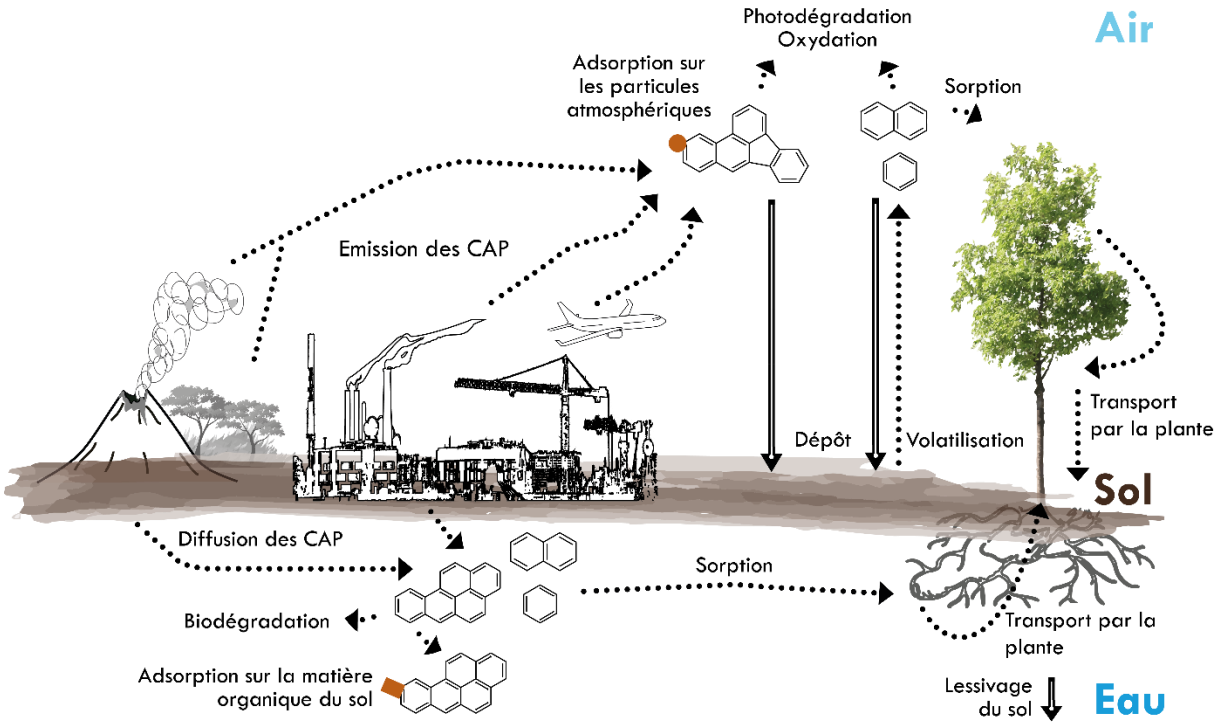


Figure 3: Origine et diffusion des CAP dans l'environnement.

Il est, par conséquent, important d'identifier et de caractériser de manière la plus exhaustive possible les CAP rencontrés dans les matrices environnementales contaminés et notamment les sols issus des sites pollués. C'est dans ce cadre général que s'inscrit ce travail de thèse qui a pour objectif de contribuer à l'amélioration des étapes de diagnostic des sols contaminés et permettre ainsi une meilleure compréhension des processus d'évolution de cette pollution. Ces connaissances permettraient in fine d'établir des stratégies de remédiation (e.g. bioremédiation, oxydation chimique in-situ[21,22]) et de gestion de sites contaminés.

Compte tenu de la forte diversité moléculaire des CAP se traduisant par des différences importantes de caractéristiques physico-chimiques (i.e. poids moléculaires, volatilité, lipophilie/hydrophilie, etc.), il est, le plus souvent, nécessaire de déployer différentes techniques analytiques complémentaires pour obtenir une caractérisation exhaustive. Préalablement à ces analyses, des étapes de préparation des sols sont indispensables : élimination de l'eau (lyophilisation, séchage à l'air libre ou à l'étuve, dessiccation), tamisage pour éliminer les éléments grossiers (généralement maille inférieure à 2mm), broyage afin d'homogénéiser l'échantillon et favoriser l'extraction au solvant (fluide supercritique, micro-onde, liquide pressurisé) pour isoler la phase organique contenant les CAP ciblés, des autres phases organiques et minérales du sol étudié[9] (Figure 4).

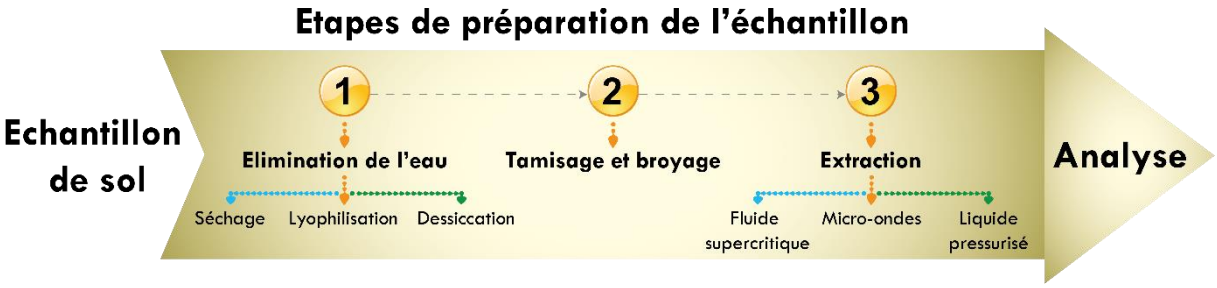


Figure 4 : Etapes principales de préparation d'un échantillon de sol avant son analyse.

La composition moléculaire de cette phase organique ainsi isolée étant très complexe, il est généralement nécessaire de séparer les CAP individuellement afin de pouvoir les détecter et les quantifier. Cette séparation s'appuie le plus souvent sur des techniques séparatives chromatographiques couplées à des systèmes de détection[14,23]. Ces dispositifs doivent être suffisamment sensibles et résolutifs pour pouvoir identifier et quantifier le maximum de CAP même s'ils sont présents en faibles quantités (i.e. en état de traces). Les systèmes chromatographiques les plus souvent utilisés sont la chromatographie liquide à haute performance (HPLC) et la chromatographie gazeuse (GC). Les systèmes de détection dans le cas d'une séparation HPLC s'appuient souvent, sur des techniques de spectroscopie électromagnétique (détecteur de fluorescence (FLD), détecteur en ultraviolet et visible (UV-Vis) ou détecteur à barrette de diodes (DAD)) ou sur la spectrométrie de masse (MS). Dans le cas d'une séparation GC, des détecteurs à ionisation de flamme (FID) ou de spectrométrie de masse (MS) sont les plus souvent utilisés[24]. Compte tenu de la très grande diversité moléculaire des CAP, la méthode séparative est sélectionnée habituellement en fonction des propriétés physico-chimiques tels que le poids moléculaire et la volatilité des espèces ciblées.

Ces techniques chromatographiques ont permis des avancées notables dans le cadre de la caractérisation de la contamination aux CAP. Néanmoins, elles présentent quelques limitations puisqu'elles sont longues, couteuses et consommatrices en solvant ou en gaz, ce qui peut limiter le nombre d'analyses réalisées. Par ailleurs, la grande diversité de poids moléculaire des CAP rend difficile leur analyse exhaustive. Pour l'analyse des 16 HAP réglementaires (de 128 à 276 g/mol), les colonnes chromatographiques sont adaptées, mais lorsque l'on souhaite aborder des CAP de poids moléculaires ou de polarité plus élevés, il devient difficile d'obtenir une résolution chromatographique satisfaisante. Les techniques chromatographiques restent néanmoins, de nos jours, les seules techniques analytiques de référence pour la caractérisation des CAP bien qu'un intérêt croissant soit porté sur des techniques de spectroscopie prometteuses. Ces dernières exploitent les différents mécanismes de l'interaction onde lumineuse-matière et offrent de multiples avantages tels que la facilité et la rapidité de mise en œuvre, la consommation limitée en solvant, la caractérisation globale parfois non destructive et parfois même directe de certains échantillons. Ces

techniques peuvent permettre de compléter les approches chromatographiques classiques voire d'être utilisé comme technique de substitution [25,26].

Parmi les nombreuses spectroscopies existantes (e.g. infrarouge et résonnance magnétique nucléaire), une technique apparait particulièrement adaptée à l'analyse des CAP. Il s'agit de la spectroscopie de fluorescence[26]. Les CAP possèdent en effet, au moins deux noyaux aromatiques[12], ce qui confèrent, pour beaucoup d'entre eux, une fluorescence intrinsèque [27]. Cette technique analytique est sensible, sélective et facile à implémenter[28]. De plus, la spectroscopie de fluorescence tridimensionnelle (3D) est une configuration instrumentale qui permet de collecter, dans une seule matrice d'excitation-émission de fluorescence (MEEF), des spectres d'émission à différentes longueurs d'excitation. Elle offre ainsi une caractérisation globale de l'échantillon analysé avec la détection simultanée de tous les composés fluorescents, que leurs longueurs d'excitation et d'émission soient connues ou inconnues. La spectroscopie Raman est une autre technique analytique prometteuse pour l'analyse des CAP. Elle est également facile à mettre en œuvre et elle présente l'avantage d'être moins spécifique aux composés fluorescents ce qui la rend complémentaire avec la spectroscopie de fluorescence. Cependant, les signaux provenant de ces spectroscopies peuvent s'avérer complexes et bruités, nécessitant ainsi des approches de traitement de données pointues couplées à des stratégies analytiques réfléchies.

C'est dans ce contexte environnemental et de santé publique que la chimiométrie a été abordé dans ce travail de thèse. La chimiométrie est définie par la société internationale de chimiométrie '*International Chemometrics Society*' comme étant la discipline scientifique qui utilise et développe des méthodes mathématiques ou statistiques pour comprendre les mesures effectuées sur un système ou un processus physico-chimique. Elle a donc pour objectif global d'extraire les informations pertinentes de données brutes, afin de les interpréter et de les exploiter au mieux. La chimiométrie est une science multidisciplinaire qui englobe, entre autres, la chimie, la physico-chimie, la métrologie, les mathématiques, les statistiques, le traitement du signal et l'informatique. Elle est ainsi utilisée dans de nombreux domaines de recherche[29].

La chimiométrie intervient à différents stades de la production et de l'analyse des données : génération et optimisation de la mesure grâce à l'élaboration de plans d'expériences, correction du bruit et/ou des interférences présents dans les données, exploration, compréhension et modélisation des tendances ou des structures particulières de celles-ci notamment leur caractère linéaire ou non linéaire. Elle s'intègre ainsi dans une approche qui met l'accent sur la mesure et la modélisation probabiliste, statistique ou par apprentissage automatique des processus physico-chimiques. Elle permet l'analyse de tous

types de données en physico-chimie tels que les données en spectroscopie[30], en chromatographie[31], en électrochimie[32], etc. Elle permet l'analyse des données univariées[33] (i.e. une seule variable), multivariées[33] (i.e. plusieurs variables) ou multivoies[34] (i.e. plusieurs variables et plusieurs dimensions).

Ainsi, l'objectif principal de ce travail de thèse est le développement et l'application de méthodologies de chimiométrie sur les données de spectroscopie de fluorescence et Raman afin de caractériser qualitativement et quantitativement, de manière aussi étendue que possible, la contamination des sols par les CAP. Les travaux de recherche menés au cours de cette thèse proposent des approches innovantes et originales pour la description, la résolution et la modélisation de ces systèmes physico-chimiques complexes.

Différents axes de la chimiométrie seront abordés. Le premier concerne le prétraitement et la correction des données spectrales du bruit et des interférences provenant de l'instrumentation et des phénomènes physiques perturbant le signal chimique. Le deuxième axe s'intéresse à l'exploration des tendances dans les données multivariées par la famille des méthodes de réduction de dimensionnalité. Le troisième axe traite de la branche de la résolution multivariée des courbes. Il s'appuie sur les méthodes de décomposition spectrale bilinéaire ou trilinéaire des signaux spectroscopiques. Cet axe a pour objectif l'extraction et la séparation des informations relatives aux espèces individuelles (i.e. CAP) regroupées et interagissant dans un mélange complexe (i.e. sol). Le quatrième axe s'inscrit dans la branche de la chimiométrie proposant des méthodes capables de modéliser les relations complexes, linéaires ou non linéaires, entre des informations spectrales et des variables quantitatives de références. Après calibration, optimisation et validation de ces modèles, l'objectif est de pouvoir prédire à partir d'une information spectrale, une valeur quantitative d'intérêt. Le dernier axe sera discuté dans ce manuscrit en tant que perspective de ce travail de recherche. Il relève du domaine de la fusion des blocs (i.e. ensembles) de données, chacun issu d'une technique analytique spécifique. Les méthodes de chimiométrie utilisées dans cet axe sont appelées méthodes d'analyse en multiblocs. Ces méthodes sont capables de prendre en compte au mieux, à travers différentes modalités spectrales ou non, toute la complexité du système physico-chimique étudié dans le but de le décrire, de le résoudre ou de le modéliser au mieux.

Ce manuscrit de thèse s'organise donc autour de cinq chapitres. Le **Chapitre I** résume les connaissances autour de la spectroscopie de fluorescence et de la spectroscopie Ultraviolet-Visible (UV-Vis) à travers un rappel des notions générales d'interactions lumière-matière, ainsi que des aspects théoriques et pratiques sur ces deux spectroscopies. Pour finir ce premier chapitre, une revue des caractéristiques de fluorescence des différentes catégories

de CAP est proposée. Le **Chapitre II**, quant à lui, expose les méthodes et les algorithmes de chimiométrie utilisés ou pouvant être utilisés pour la correction, l'analyse et la modélisation des signaux de fluorescence. Les deux chapitres suivants traitent de développements méthodologiques respectivement réalisés pour la correction et la décomposition trilinéaire des signaux de fluorescence. Plus particulièrement, le **Chapitre III** présente un nouvel algorithme, développé dans le cadre de cette thèse, nommé MT-SVD pour '*Multi-Truncated Singular Value Decomposition*' qui permet la gestion des effets de diffusion de la lumière, du bruit et des phénomènes pouvant être observé de déficience du rang matriciel dans les signaux de fluorescence. Cet algorithme sera détaillé avant d'être mis en pratique sur une série de cas de plus en plus complexes. Le **Chapitre IV** expose les résultats de l'identification de 16 HAP dans des sols industriels contaminés grâce à la décomposition trilinéaire des signaux de fluorescence par PARAFAC ('*PARAllel FACtor analysis*'). Dans ce chapitre, une approche originale d'exploitation de ce type de décomposition est proposée. Et enfin, le **Chapitre V** vient conclure la partie résultat en proposant d'une part, le résultat de la modélisation quantitative de la pollution des sols aux 16 HAP préalablement identifiés et d'autre part, une discussion autour de la caractérisation des CAP par spectroscopie Raman et de la perspective d'une analyse en multiblocs des signaux Raman et de fluorescence. En outre, une conclusion générale, étoffée de perspectives de recherche, viendra clore ce manuscrit de thèse.

CHAPITRE I

CARACTERISATION PAR SPECTROSCOPIE DE FLUORESCENCE DES COMPOSES AROMATIQUES POLYCYCLIQUES

I.1 Introduction

La spectroscopie est la science de l'acquisition des spectres. Elle peut être vue comme étant un panel de techniques instrumentales avancées exploitant divers phénomènes engendrés par l'interaction entre un rayonnement et une matière. Parmi elles, la branche des spectroscopies électromagnétiques exploite l'absorption, l'émission ou la diffusion d'un rayonnement électromagnétique par la matière.

La spectroscopie UV-Vis et la spectroscopie de fluorescence sont deux spectroscopies électromagnétiques complémentaires qui peuvent être exploitées pour caractériser des composés chimiques tels que les composés aromatiques polycycliques (CAP) qui présentent une fluorescence intrinsèque pour un certain nombre d'entre eux. Comme précisé en introduction, la caractérisation de ces polluants, omniprésents dans l'environnement et répondant aux spécifications de l'organisation mondiale de la santé, constitue un enjeu de santé publique.

Ainsi, dans ce chapitre, l'attention sera focalisée sur la spectroscopie de fluorescence qui est une technique prometteuse d'analyse des CAP. D'abord, les mécanismes moléculaires, d'absorption en UV-Vis et d'émission de fluorescence, seront définis. Ensuite, quelques aspects théoriques et pratiques liés à la mesure en spectroscopie de fluorescence seront exposés. Enfin, les phénomènes physico-chimiques affectant la mesure et la réponse en fluorescence seront détaillés.

Pour finir, un état de l'art des caractéristiques de fluorescence des CAP liées à leur structuration moléculaire sera présenté.

I.2 Notions générales d'interaction lumière-matière

Une molécule est formée d'atomes et d'électrons. Les atomes sont reliés entre eux par des liaisons chimiques covalentes formées par les électrons. Le comportement dynamique de la molécule est déterminé par les mouvements de ses électrons et de ses noyaux atomiques[35,36]. Lorsque la lumière interagit avec la matière, des transferts d'énergie peuvent se produire, ce qui peut induire un comportement dynamique des molécules qui se manifeste dès lors par des mouvements électroniques (i.e. transitions électroniques), des mouvements moléculaires d'ensemble (i.e. translation et rotation de la molécule) ou des mouvements moléculaires de vibration des noyaux (élongations, torsions ou déformations internes)[35]. Ces mouvements sont spécifiques à chaque système moléculaire qui possède des niveaux discrets d'énergie définis, reliés à la nature des atomes et des liaisons entre eux[36]. Ces mouvements fournissent ainsi des informations précieuses sur les caractéristiques de la molécule en

question, telles que sa structure et sa conformation, sa composition chimique et son environnement. En outre, les transferts d'énergie peuvent se faire soit par l'absorption d'un photon, soit par l'émission d'un photon, soit par la diffusion d'un photon. Lorsqu'un photon est absorbé par une molécule, celle-ci gagne de l'énergie et passe à un état d'excitation. À l'inverse, lorsqu'une molécule passe d'un état excité à un état moins excité, elle émet un photon et perd de l'énergie. Enfin, lorsqu'un photon est diffusé par la molécule, cela peut se produire de deux manières : soit la molécule ne gagne ni ne perd d'énergie (i.e. diffusion élastique), soit elle gagne ou perd de l'énergie (i.e. diffusion inélastique).

La lumière se caractérise par sa double nature ondulatoire et corpusculaire. D'une part, elle est constituée, selon la théorie classique, d'un ensemble d'ondes électromagnétiques monochromatiques planes progressives. Chaque onde se distingue par sa fréquence ν définit par la relation c/λ où c représente la vitesse de la lumière ($\approx 3 \times 10^8$ m.s^{-1}) et λ la longueur de l'onde électromagnétique, exprimée généralement en nanomètre. D'autre part, la lumière se comporte, selon la théorie quantique, comme un ensemble de corpuscules de masse identique nulle et ayant la propriété de transporter une quantité précise d'énergie (i.e. un flux de photons) définit par la relation [35,37] :

$$E = h\nu = \frac{hc}{\lambda} \quad \text{(I.1)}$$

Avec, l'énergie des photons notée E en électron-volt (eV), la constante de Planck notée h ($\approx$4.13×10^{-15} eV.s) et la fréquence ν en hertz (Hz).

Une molécule quant à elle possède une énergie potentielle (E_{mol}) qui est définit par l'approximation de Born-Oppenheimer (i.e. approximation adiabatique) comme la somme de trois énergies[38] : $E_{mol} = E_{éle} + E_{vib} + E_{rot}$, avec :

- $E_{éle}$, énergie des électrons
- E_{vib}, énergie des vibrations des noyaux
- E_{rot}, énergie de la rotation de la molécule.

La dissociation de ces énergies est permise par l'approximation de Born-Oppenheimer car la masse des électrons est nettement plus faible que celle des noyaux[36]. Ainsi, les mouvements rapides des électrons (de l'ordre de la femtoseconde) s'ajustent aux mouvements plus lents de vibration des noyaux et de rotation de la molécule (de l'ordre de la picoseconde)[36]. Par ailleurs, l'énergie de translation n'est pas quantifiée car elle est considérée comme étant très faible, à l'exception du cas des gaz[35]. Ce dernier cas ne sera pas abordé dans le cadre de cette thèse.

Un mouvement électronique ou moléculaire est induit lorsque le rayonnement lumineux incident caractérisé par l'énergie E interagit avec le système moléculaire via un transfert d'énergie. Ce transfert se produit, par absorption des photons, uniquement dans le cas où l'énergie véhiculée correspond à la différence ΔE entre deux niveaux d'énergie électroniques, vibrationnels ou rotationnels de la molécule[38] :

$$E = h\nu = \Delta E \qquad \text{(I.2)}$$

Ces mouvements électroniques ou moléculaires sont donc à la fois dépendants des niveaux d'énergie moléculaire, inhérents à chaque molécule, et de l'énergie du rayonnement incident et donc de sa position dans le spectre électromagnétique qui regroupe et catégorise l'ensemble des rayonnements électromagnétiques. Ce dernier est composé de sept régions principales allant de celle des ondes radio (la région la moins énergétique avec de grandes longueurs d'ondes) jusqu'à celle des rayons gamma (la région la plus énergétique avec de courtes longueurs d'ondes) en passant par la région de l'UV-Vis qui se situe à des longueurs d'ondes comprises entre 190 et 800 nm[36,39] (Figure I.1).

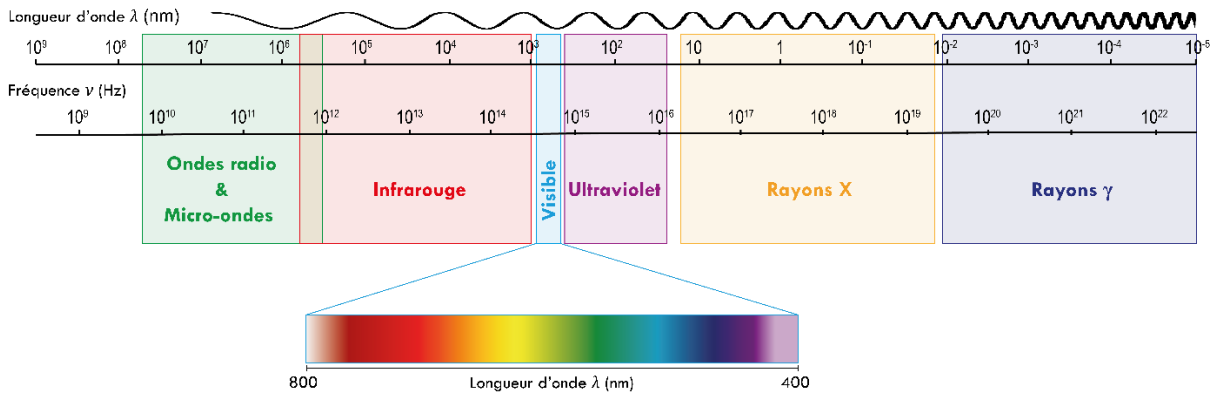


Figure I.1 : Le spectre électromagnétique[40].

Chacun des mouvements électroniques ou moléculaires peut être détecté et analysé généralement par une technique spectroscopique spécifique adaptée et exploitant une région typique du spectre électromagnétique[35]. Par exemple, la spectroscopie infrarouge est une technique qui permet de caractériser les mouvements de vibration moléculaire en mesurant l'absorption d'énergie du rayonnement infrarouge par les molécules. La spectroscopie Raman est également une technique qui permet de caractériser les mouvements de vibration moléculaire, mais elle exploite le phénomène de diffusion plutôt que l'absorption. Dans la spectroscopie Raman, le rayonnement incident est diffusé inélastiquement par les molécules, et les changements d'énergie associés aux mouvements de vibration moléculaire sont détectés par analyse de la fréquence et de l'intensité du rayonnement diffusé. La

spectroscopie d'absorption micro-onde est quant à elle un exemple de spectroscopie électromagnétique d'absorption utilisée pour étudier les mouvements de rotation moléculaire.

Dans le cadre de cette thèse, l'intérêt principal est porté sur la spectroscopie d'émission de fluorescence, qui fait partie des techniques analytiques exploitant les interactions ayant lieu dans la région UV-Vis. Elle permet d'obtenir des informations sur la configuration moléculaire de composés spécifiques, que nous allons définir ultérieurement, par l'analyse de leurs spectres. Elle peut également, sous certaines conditions, permettre la quantification d'analytes d'intérêts. Par ailleurs, cette technique est intrinsèquement liée avec une autre spectroscopie du même domaine spectral qui est la spectroscopie d'absorption en UV-Vis. La spectroscopie de fluorescence mesure, en effet, une émission radiative de quanta (i.e. photons) de lumière consécutive à l'absorption de l'énergie UV-Vis par la molécule[41] tandis que la spectroscopie UV-Vis réalise une mesure indirecte de cette quantité d'énergie UV-Vis absorbée par la molécule. Ces dernières peuvent ainsi fournir des informations complémentaires intéressantes.

I.3 Spectroscopies d'absorption UV-Vis et d'émission de fluorescence

I.3.1 Mécanisme d'absorption de la lumière UV-Vis

Lorsque le rayonnement lumineux UV-Vis, caractérisé par des énergies E allant, selon l'équation (I.1), de ≈ 6.52 à 1.55 eV, interagit avec un système moléculaire capable d'absorber les photons de cette gamme, un transfert d'énergie se produit et des mouvements électroniques sont observés. Les énergies véhiculées par ces ondes correspondent en effet, aux énergies ΔE (équation (I.2)) nécessaires aux transitions électroniques des molécules[39].

Plus précisément, dans les conditions de température ambiante et de pression atmosphérique standards, les électrons remplissant les orbitales moléculaires liantes (i.e. les orbitales σ et π) et non-liantes (i.e. les orbitales n) de ces systèmes moléculaires sont dans un état fondamental stable. Lorsque l'énergie apportée par le rayonnement UV-Vis est suffisante (i.e. $E \geq \Delta E$), ces électrons passent alors vers des orbitales moléculaires anti-liantes (i.e. les orbitales σ^* et π^*) non remplies pour se retrouver dans un état électronique excité. On parle alors de transitions électroniques (Figure I.2).

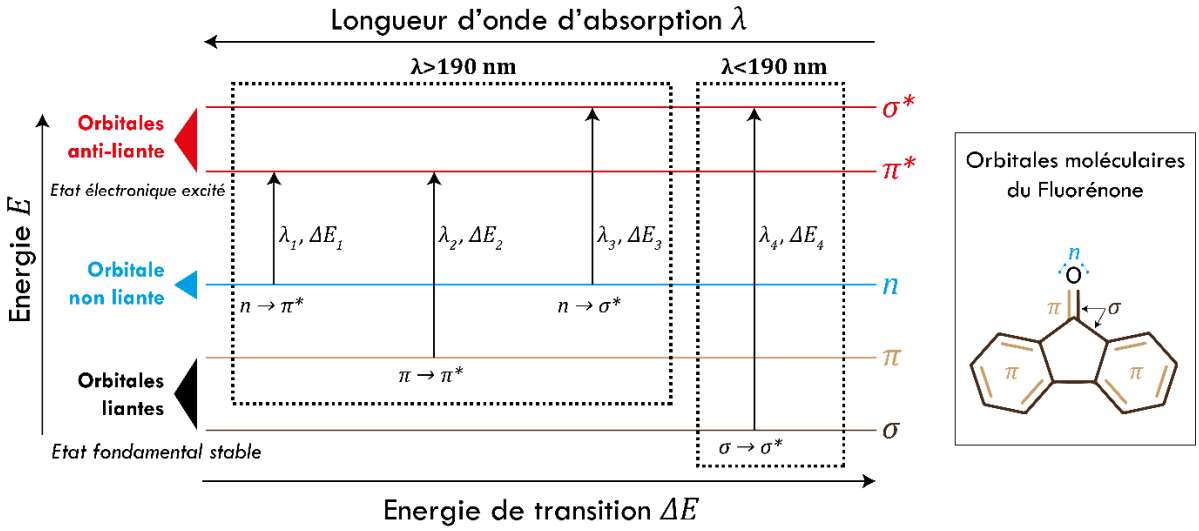


Figure I.2 : Niveaux d'énergie des orbitales moléculaires et différentes transitions électroniques d'un système moléculaire : Exemple du Fluorénone. $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$ et $\Delta E_1 < \Delta E_2 < \Delta E_3 < \Delta E_4$.

La capacité d'un système moléculaire à absorber un rayonnement dépend donc de la nature de ses orbitales moléculaires. Tous les composés saturés sont théoriquement capables d'absorber un rayonnement à condition qu'il soit assez énergétique (i.e. λ<190 nm) pour permettre la transition $\sigma \rightarrow \sigma^*$. Les systèmes moléculaires conjugués se caractérisent par des transitions $n \rightarrow \pi^*$ et/ou $\pi \rightarrow \pi^*$ permises, sous certaines conditions dans la gamme de l'UV-Vis (i.e. λ>190 nm). Par conséquent, ils sont, pour la plupart, d'excellents absorbants (i.e. chromophores) dans cette gamme. Cette absorption permise d'énergie (i.e. de photons) est à la base de la spectroscopie d'absorption moléculaire dans l'UV-Vis mais elle est également la première étape du phénomène de fluorescence.

I.3.2 Mécanisme d'émission de fluorescence

Après absorption du rayonnement et donc de l'énergie UV-Vis, certaines molécules typiques, appelées fluorophores, s'excitent, passant d'un état fondamental stable à un état électronique excité avant d'émettre de l'énergie sous forme de photons pour retourner à un état fondamental stable (i.e. photoluminescence). Ce processus peut être divisé en deux types d'émissions radiatives, à l'origine de deux phénomènes différents que sont la fluorescence (i.e. émission rapide de photons) et la phosphorescence (i.e. émission lente de photons exploitée dans le cadre de la spectroscopie de phosphorescence). Un fluorophore ou fluorochrome est une molécule ou une structure supramoléculaire capable d'émettre ce type de rayonnement. Il est soit intrinsèque (i.e. émission de fluorescence native), soit extrinsèque (i.e. émission de fluorescence induite après dérivation chimique)[42]. Son pouvoir fluorescent provient alors de sa capacité à absorber le rayonnement UV-Vis dû à sa structure chimique, et donc, à la nature même de ses orbitales moléculaires. En outre, des composés moléculaires proches structurellement des fluorophores et appelés chromophores sont également capables

d'absorber de la lumières UV-Vis, mais ne réémettent pas de la lumière sous forme de fluorescence.

Après excitation du fluorophore par un rayonnement photonique monochromatique caractérisée par une longueur d'onde λ, ce dernier absorbe en 10^{-15} secondes de l'énergie E (définit par l'équation (I.1)) et passe ainsi d'un état électronique fondamental stable S_0 vers un état électronique singulet excité S_1 ou S_2. Le phénomène radiatif de fluorescence se produit rapidement, de l'ordre de grandeur de 10^{-10} à 10^{-7} secondes[43] par la relaxation de l'état excité singulet S_1, ou très rarement S_2, vers l'état électronique fondamental S_0[28]. Cet effet radiatif est quantifié par une grandeur nommée le rendement quantique de fluorescence ou tout simplement rendement de fluorescence et elle est notée Φ_F. Il s'agit du rapport entre le nombre de photons émis et le nombre de photons absorbés ($\Phi_F = \frac{Nombre\ de\ quanta\ émis}{Nombre\ de\ quanta\ absorbés}$). Il est possible de retrouver cette grandeur exprimée théoriquement par les différentes constantes de vitesses relatives aux différents phénomènes radiatifs et non radiatifs mis en jeu, ou encore, en fonction de l'intensité de fluorescence mesurée et d'un certain nombre de paramètres liés à l'appareillage, en particulier à sa configuration optique. Les différentes relations et équations sont disponibles dans les ouvrages de B. Valeur et G. G. Guilbault[44–46]. Le mécanisme de fluorescence ainsi que les différents processus mis en jeu sont décrits par le diagramme de Perrin-Jablonski (Figure I.3).

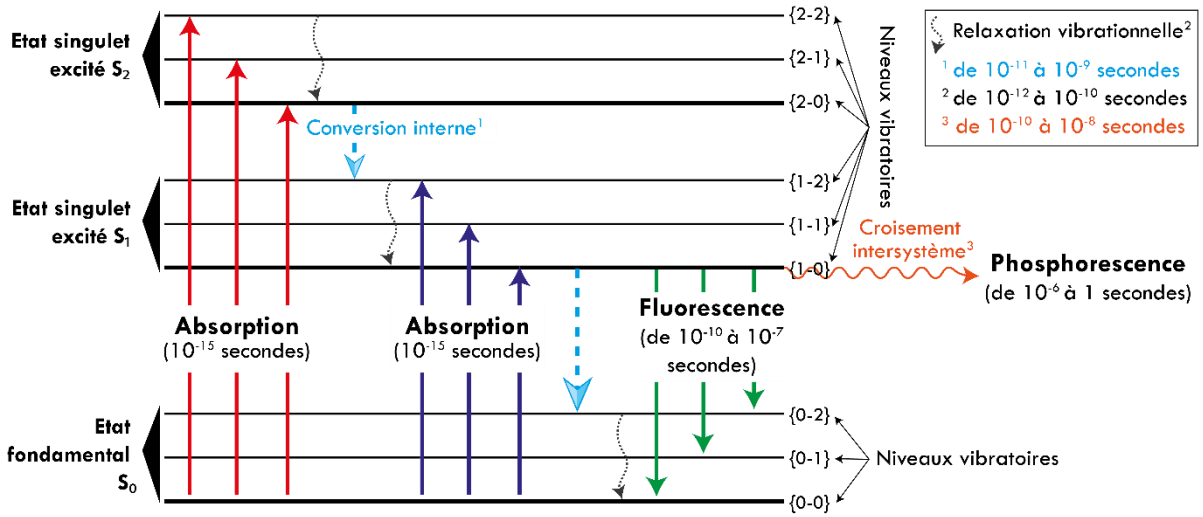


Figure I.3 : Diagramme de Perrin-Jablonski décrivant le phénomène de fluorescence. Adapté des ouvrages de B. Valeur [44,45].

Le phénomène de fluorescence se caractérise par trois propriétés principales communes à la plupart des fluorophores sauf, pour certains cas particuliers de composés au comportement atypique ou se trouvant dans un environnement particulier. En premier lieu, les spectres d'émission d'un fluorophore ont généralement une allure invariante, peu importe la longueur d'onde d'excitation. En effet, bien que la longueur d'onde d'excitation puisse varier,

cela ne provoque que des variations d'intensité dans les spectres d'émission (Figure I.4). Ce principe est décrit par la loi de Kasha. Il s'agit d'une des conséquences de la relaxation vibrationnelle (i.e. perte d'énergie non radiative) qui se matérialise, avant l'émission de fluorescence, par une décroissance rapide (de 10^{-12} à 10^{-10} secondes) vers le niveau vibratoire le plus bas de l'état excité singulet (Figure I.3). Ainsi, l'émission se produit à partir de cet état vibrationnel quelle que soit la longueur d'onde d'excitation.

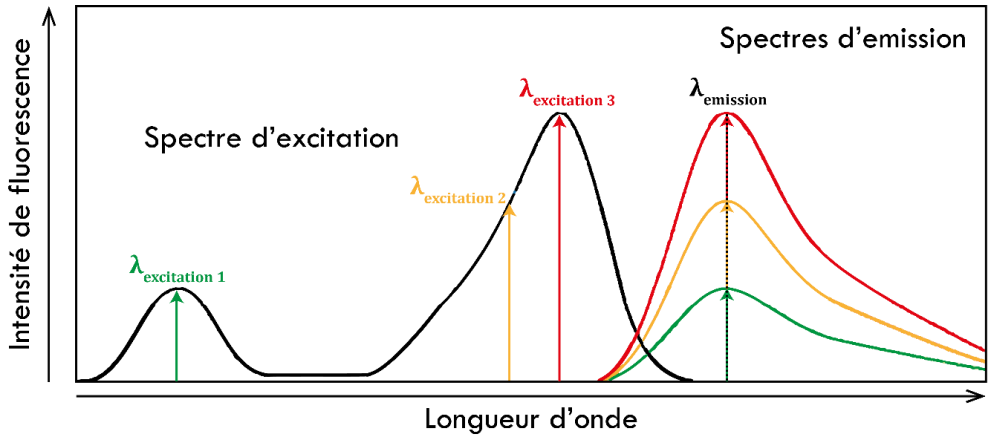


Figure I.4 : Spectre d'excitation et d'émission d'un fluorophore, illustrant la règle de Kasha[27].

Ensuite, le décalage Stokes est une autre propriété du phénomène de fluorescence qui résulte d'une perte d'énergie entre l'excitation et l'émission. Cet écart est encore principalement dû à la perte d'énergie lors du processus de la relaxation vibrationnelle. En pratique, le décalage Stokes s'observe par des longueurs d'ondes d'émission supérieures aux longueurs d'ondes d'excitation correspondantes. Enfin, la dernière propriété est que le spectre d'émission (i.e. $S_1 \rightarrow S_0$) est l'image inversée du spectre d'absorption $S_0 \rightarrow S_1$ et non du spectre d'absorption global (i.e. $S_0 \rightarrow S_1$ et $S_0 \rightarrow S_2$). Cette caractéristique est la conséquence de l'effet de la conversion interne qui se produit rapidement (de 10^{-11} à 10^{-9} secondes) du niveau S_2 vers le niveau S_1. En effet, l'émission de fluorescence se produit principalement à partir de l'état excité singulet S_1 et non de l'état S_2 (Figure I.3)[27,28,47]. Les propriétés mentionnées ci-dessus sont illustrées à travers l'exemple des spectres d'absorption et d'émission de deux fluorophores différents, à savoir la quinine et l'anthracène (Figure I.5).

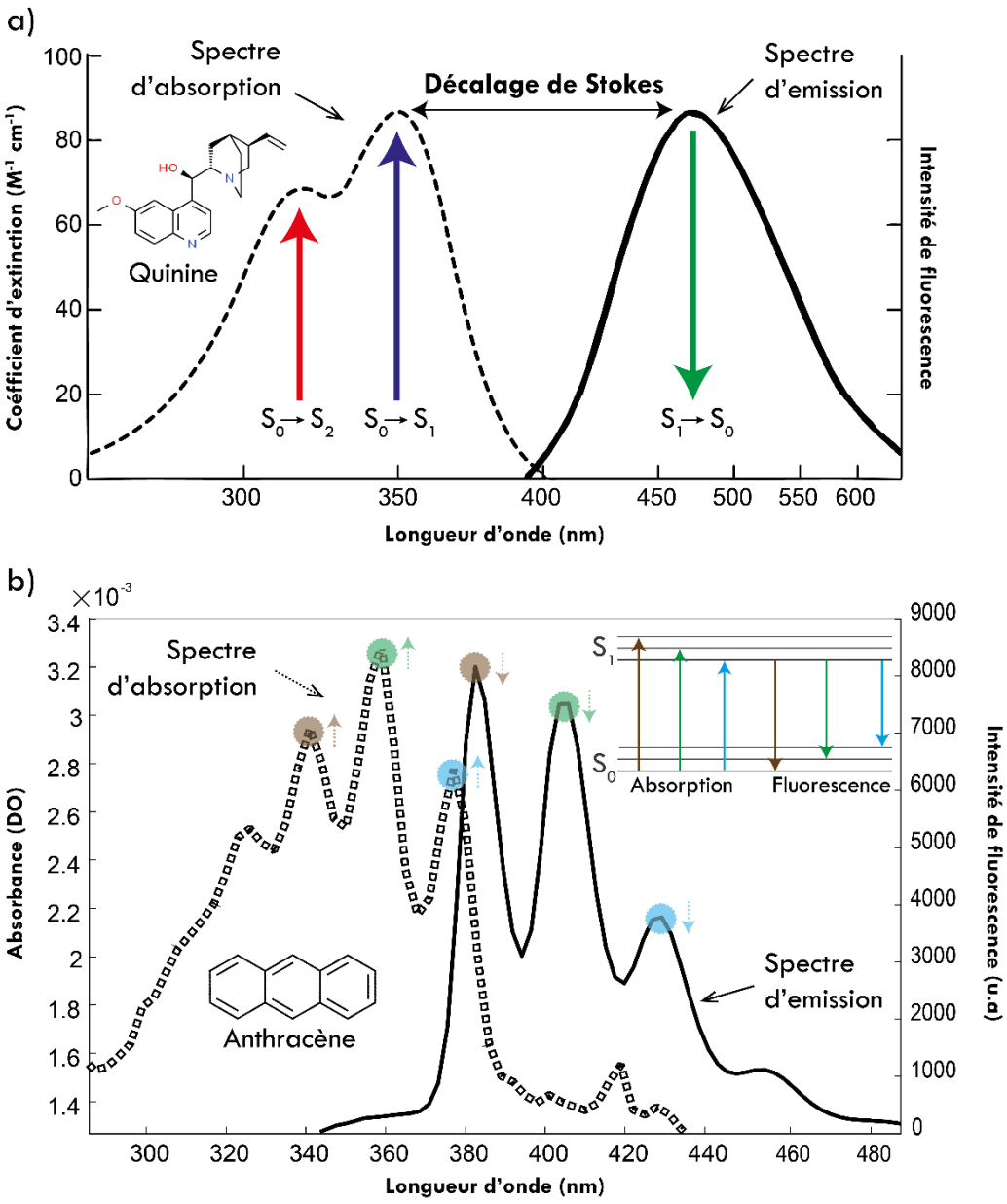


Figure I.5 : Spectre d'absorption et d'émission de la quinine (a)[27]. Spectre d'absorption et d'émission de l'anthracène (b).

I.4 Quantification par spectroscopie d'absorption UV-Vis et d'émission fluorescence

Comme mentionné précédemment, la spectroscopie d'absorption UV-Vis et la spectroscopie d'émission de fluorescence, également appelée spectrofluorimétrie, sont deux techniques spectroscopiques imbriquées et complémentaires qui permettent l'analyse qualitative (i.e. caractérisation structurale) et l'analyse quantitative des échantillons liquides dans les conditions linéaires de la loi de Beer-Lambert valable dans le cas de milieux très dilués.

I.4.1 Loi de Beer-Lambert

La spectroscopie d'absorption UV-Vis permet la mesure indirecte de la quantité d'énergie UV-Vis absorbée par une molécule (i.e. absorbance notée $A(\lambda)$). Il s'agit, en effet, d'une mise en relation dans le cas d'un milieu très dilué, via la loi de Beer-Lambert (Equation I.3), de l'intensité du rayonnement UV-Vis incident noté $I_0(\lambda)$ et de celle du rayonnement transmis via le milieu noté $I_T(\lambda)$ (mesuré expérimentalement)[39] :

$$A(\lambda) = \log\frac{I_0(\lambda)}{I_T(\lambda)} \quad \text{(I.3)}$$

Sachant que $I_T(\lambda)$ est reliée par l'équation (I.4) à la concentration de la molécule, au coefficient d'absorption molaire (i.e. une mesure de l'efficacité avec laquelle une molécule absorbe la lumière dans un solvant particulier) et au trajet optique de la cellule de mesure (i.e. épaisseur de l'échantillon traversé), notés c, $\epsilon(\lambda)$ et l, respectivement, l'équation (I.3) devient donc l'équation (I.5) qui relie linéairement la concentration d'un chromophore à son absorbance[39] :

$$I_T(\lambda) = I_0(\lambda)e^{-2{,}3\epsilon(\lambda)lc} \quad \text{(I.4)}$$

$$A(\lambda) = \log\frac{I_0(\lambda)}{I_T(\lambda)} = \epsilon(\lambda)lc \quad \text{(I.5)}$$

Avec c exprimée en mole par litre (mol.L^{-1}), $\epsilon(\lambda)$ en litre par mole et par centimètre (L.mol^{-1}.cm^{-1}) et l en centimètre (cm). Concernant la spectroscopie de fluorescence, les développements mathématiques[44–46] qui suivent, permettent d'expliquer la nature linéaire ou non linéaire de la relation entre l'intensité du signal de fluorescence émise et la concentration d'un fluorophore.

I.4.2 Linéarité et non linéarité du signal de fluorescence

L'intensité de l'énergie absorbée par un fluorophore qui est excitée à une longueur d'onde d'excitation λ_{ex} est notée $I_A(\lambda_{ex})$. Elle est définie, par l'équation (I.6) comme l'écart entre l'intensité d'énergie incidente, notée $I_0(\lambda_{ex})$ et l'intensité de la lumière transmise par le milieu, notée $I_T(\lambda_{ex})$:

$$I_A(\lambda_{ex}) = I_0(\lambda_{ex}) - I_T(\lambda_{ex}) \quad \text{(I.6)}$$

$I_0(\lambda_{ex})$ peut-être notée simplement I_0 puisque les dispositifs expérimentaux actuels permettent de s'affranchir des fluctuations de l'intensité incidente, dépendantes de λ_{ex} et liées à la source excitatrice. Ainsi, l'équation (I.6) devient :

$$I_A(\lambda_{ex}) = I_0 - I_T(\lambda_{ex}) \quad \text{(I.7)}$$

La loi de Beer-Lambert stipule que l'intensité de la lumière transmise par le milieu peut s'exprimer, en fonction de $c, \varepsilon(\lambda_{ex})\ et\ l$, par la relation (I.8) similaire à la relation (I.4) :

$$I_T(\lambda_{ex}) = I_0\ e^{-2.3\varepsilon(\lambda_{ex})lc} \quad \text{(I.8)}$$

Ainsi, l'équation (I.7) devient :

$$I_A(\lambda_{ex}) = I_0(1 - e^{-2.3\varepsilon(\lambda_{ex})lc}) \quad \text{(I.9)}$$

Par ailleurs, l'intensité de fluorescence émise à la longueur d'onde d'émission λ_{em} par le fluorophore excité à λ_{ex} est notée $I_F(\lambda_{ex}, \lambda_{em})$. Elle est reliée au rendement quantique de fluorescence du fluorophore Φ_F et à l'intensité du rayonnement qu'il absorbe par :

$$I_F(\lambda_{ex}, \lambda_{em}) = \Phi_F I_A(\lambda_{ex}) \quad \text{(I.10)}$$

Ainsi, d'après les relations (I.9) et (I.10) l'équation basique reliant la fluorescence à la concentration peut s'écrire :

$$I_F(\lambda_{ex}, \lambda_{em}) = \Phi_F I_0(1 - e^{-2.3\varepsilon(\lambda_{ex})lc}) \quad \text{(I.11)}$$

La relation (I.11) démontre que l'intensité de fluorescence est liée non linéairement à la concentration du fluorophore c. Néanmoins, un développement limité, appelé développement de McLaren, du terme $1 - e^{-2.3\varepsilon(\lambda_{ex})lc}$ peut être utilisé dans le cas des solutions très diluées i.e. fluorophore(s) en faible(s) concentration(s) :

$$1 - e^{-2.3\varepsilon(\lambda_{ex})lc} = 2.3\varepsilon(\lambda_{ex})lc - \frac{1}{2}(2.3\varepsilon(\lambda_{ex})lc)^2 + \cdots$$

En effet, il est à observer qu'au fur et à mesure que la concentration décroit, au fur et à mesure que les termes d'ordre supérieur deviennent négligeables. De ce fait, en ne gardant que le premier terme, l'équation (I.11) devient :

$$I_F(\lambda_{ex}, \lambda_{em}) \cong 2.3\Phi_F I_0 \varepsilon(\lambda_{ex})lc \quad \text{(I.12)}$$

La relation I.12 prédit donc que la fluorescence est liée linéairement, dans le cas des solutions très diluées, à la concentration du fluorophore. Afin de s'assurer que les mesures en fluorescence sont effectuées dans des conditions linéaires autorisant l'utilisation du développement limité de McLaren, il est possible de les précéder par des mesures

d'absorbance en spectroscopie d'absorption UV-Vis sachant que, d'après l'équation (I.5), $A_{ex} = \varepsilon(\lambda_{ex})lc$ et donc l'équation (I.12) devient :

$$I_F(\lambda_{ex}, \lambda_{em}) \cong 2.3\Phi_F I_0 A_{ex} \quad \text{(I.13)}$$

Une préconisation est fournie dans la littérature pour garantir la condition de linéarité de la loi de Beer-Lambert entre la concentration et l'intensité de fluorescence. Il s'agit de ne pas dépasser la valeur de 0.05 pour l'absorbance mesurée du fluorophore (i.e. $A_{ex} < 0.05$) pour garder un écart de la linéarité acceptable, entre l'intensité de fluorescence et la concentration associée (Figure I.6).

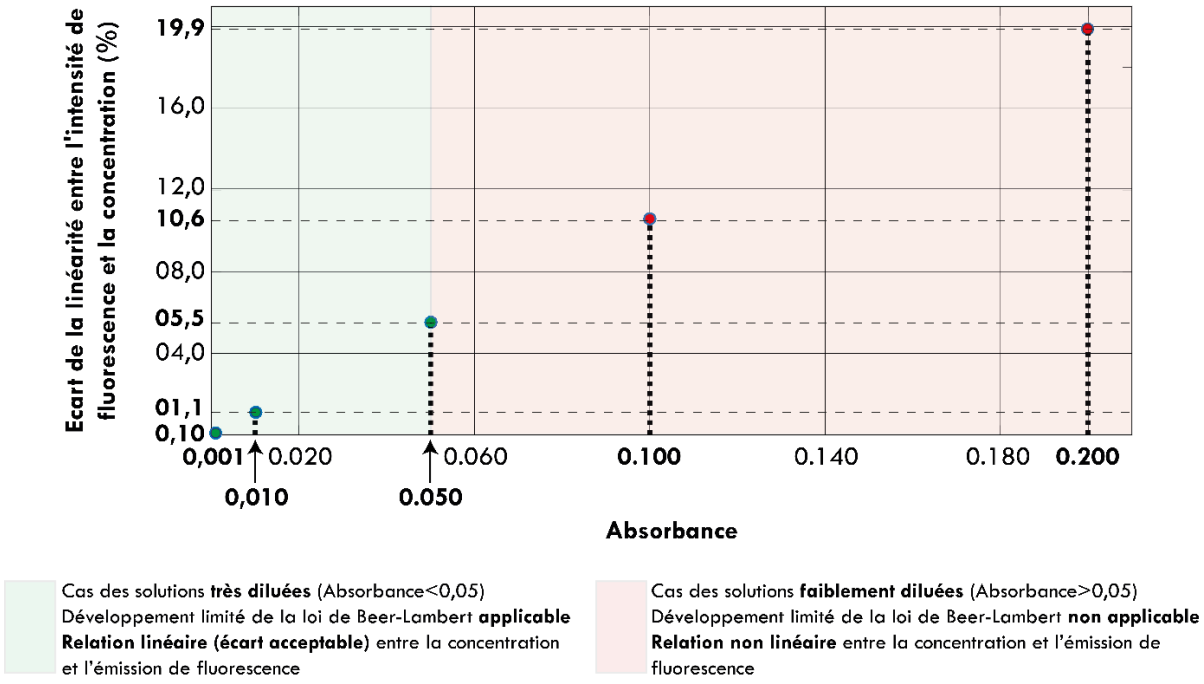


Figure I.6 : Absorbance versus écart de la linéarité entre l'intensité de fluorescence et la concentration. Graphique réalisé à partir des données fournies dans[44].

I.5 Aspects pratiques de la spectroscopie de fluorescence

En pratique, les trois grandeurs λ_{ex}, λ_{em} et I_F sont prises en compte lors des mesures en spectrofluorimétrie. Les deux longueurs d'ondes sont généralement exprimées en nanomètres (nm) tandis que l'intensité de fluorescence est exprimée en unité arbitraire (u.a). Au niveau instrumental, trois configurations principales peuvent être utilisées en spectrofluorimétrie : la spectrofluorimétrie classique, synchrone et la fluorescence 3D. Le résultat d'une mesure en fluorescence est dépendant de la configuration utilisée, soit un spectre d'émission, soit un spectre d'excitation, soit un spectre synchrone ou encore une carte tridimensionnelle (i.e. 3D) appelée matrice d'excitation-émission de fluorescence (MEEF).

I.5.1 La spectrofluorimétrie classique

La spectrofluorimétrie classique permet la mesure de l'I_F dans un spectre d'émission (resp. d'excitation) sur l'ensemble des λ_{em} (resp. λ_{ex}) mais uniquement en une seule λ_{ex} (resp. λ_{em}) préalablement fixée. Cette technologie est adaptée à l'analyse d'échantillons très simples tels que ceux ne contenant qu'un seul composé chimique pur, pour lequel l'λ_{ex} (resp. λ_{em}) optimale est identifiée et fixée. Le spectre d'émission réel (i.e. non mesuré) d'un fluorophore est noté $F(\lambda_{em})$. Il représente la distribution de la probabilité des différentes transitions qui peuvent avoir lieu depuis le plus bas niveau vibrationnel de S_1 vers les différents niveaux vibrationnels de S_0 (cf. §I.3.2). Cette distribution de probabilité est reliée au nombre de photons émis et à ceux absorbés par le fluorophore (i.e. rendement quantique de fluorescence) via la relation :

$$\int_0^{\infty} F(\lambda_{em}) = \Phi_F \tag{I.14}$$

Ce spectre d'émission réel peut être relié à l'$I_F(\lambda_{ex}, \lambda_{em})$ mesurée en l'introduisant dans les équations (I.11) et (I.13), pour le cas non linéaire et le cas linéaire, respectivement. Ce développement aboutit aux relations (I.15) et (I.16), respectivement :

$$I_F(\lambda_{ex}, \lambda_{em}) = kF(\lambda_{em})I_0(1 - e^{-2.3\varepsilon(\lambda_{ex})lc}) \tag{I.15}$$

$$I_F(\lambda_{ex}, \lambda_{em}) \cong 2.3kF(\lambda_{em})I_0A_{ex} \tag{I.16}$$

La grandeur k est un facteur de proportionnalité qui dépend de plusieurs paramètres liés majoritairement à la configuration du dispositif expérimental. Dans le cas de la mesure du spectre d'émission, seule λ_{em} évolue. Par conséquent, A_{ex} reste constante. Ainsi, dans le cas des solutions très diluées, la relation généraliste (I.16) peut être simplifiée en considérant $K = 2.3kI_0A_{ex}$ comme une constante :

$$I_F(\lambda_{em}) \cong KF(\lambda_{em}) \tag{I.17}$$

Dans le cas du spectre d'excitation, λ_{em} est fixe et donc $F(\lambda_{em})$ est constant. Ceci signifie que la variation du spectre est reliée uniquement aux variations de A_{ex}. Ainsi, dans le cas des solutions très diluées, la relation généraliste (I.16) peut être simplifiée en considérant $\acute{K} = 2.3kI_0F(\lambda_{em})$ comme une constante :

$$I_F(\lambda_{ex}) \cong \acute{K}A_{ex} \tag{I.18}$$

Jusqu'à maintenant, les développements mathématiques exposés tiennent compte uniquement d'un seul fluorophore. Dans le cas d'un mélange de N fluorophores présents à de faibles concentrations, les équations (I.16), (I.17) et (I.18) deviennent (I.19), (I.20) et (I.21), respectivement.

$$I_F(\lambda_{ex}, \lambda_{em}) \cong 2.3kI_0 \sum_{n=1}^{N} F_n(\lambda_{em}) A_{n_{ex}} \quad \text{(I.19)}$$

$$I_F(\lambda_{em}) \cong K \sum_{n=1}^{N} F_n(\lambda_{em}) \quad \text{(I.20)}$$

$$I_F(\lambda_{ex}) \cong \acute{K} \sum_{n=1}^{N} A_{n_{ex}} \quad \text{(I.21)}$$

I.5.2 La spectrofluorimétrie synchrone

Dans le cas des échantillons complexes, il est difficile, voire impossible, d'avoir en un seul spectre d'excitation ou d'émission, une signature spectrale de l'ensemble des composés fluorescents qui les composent. Chaque molécule, de par sa structure chimique, possède une absorption maximale et donc un rendement de fluorescence optimal, à une ou plusieurs λ_{ex} spécifiques (cf. §I.3.1). Par conséquent, la spectrofluorimétrie synchrone fût développée comme alternative. Il s'agit dans ce cas de fixer un décalage constant noté $\Delta\lambda$ entre plusieurs λ_{ex} et λ_{em} qui elles, varient simultanément. λ_{ex} varie avec un pas choisi par l'utilisateur (i.e. résolution spectrale) tandis que $\lambda_{em} = \lambda_{ex} + \Delta\lambda$[27]. Le décalage $\Delta\lambda$ est généralement choisi en lien avec le décalage de Stokes de la molécule présentant le plus faible rendement de fluorescence ou ayant la concentration la plus faible[43]. Le résultat d'une mesure de l'I_F avec cette technique est un spectre synchrone dépendant à la fois des λ_{ex}, λ_{em} et du $\Delta\lambda$ (Figure I.7).

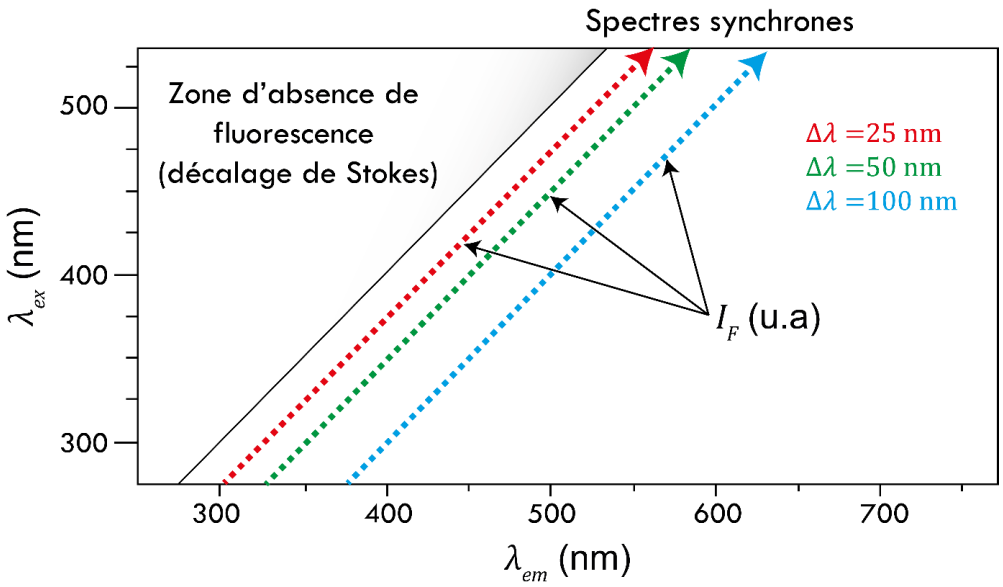


Figure I.7 : Schématisation des spectres synchrones qui auraient pu être obtenus expérimentalement à $\Delta\lambda = 25\ nm\ (rouge), 50nm\ (vert) ou\ 100\ nm (en\ bleu)$ *avec un* $\lambda_{ex} = 275\ nm$ *au départ.*

La spectrofluorimétrie synchrone permet une meilleure caractérisation des mélanges complexes que la spectrofluorimétrie classique, néanmoins, elle nécessite une connaissance préalable du système étudiée. Elle est en effet, dépendante du choix du $\Delta\lambda$ et elle n'offre qu'une caractérisation partielle de l'échantillon complexe.

I.5.3 La fluorescence 3D

La fluorescence 3D est la seule technique de fluorescence qui offre une caractérisation globale de l'échantillon analysé. Dans ce système, l'I_F est mesurée sur l'ensemble des λ_{em} et λ_{ex}. Le résultat des mesures est une collection de spectres d'émission et d'excitation qui sont arrangés en une MEEF. Cette dernière peut être visualisée comme une image contenant sur chacun de ces pixels une I_F (Figure I.8).

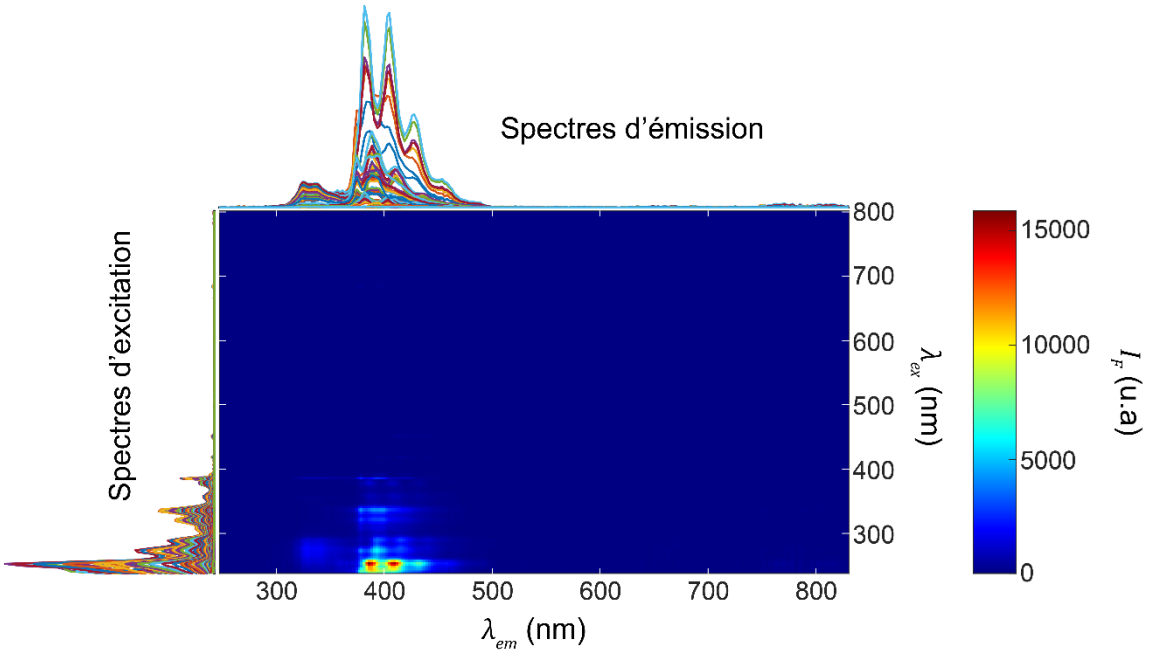


Figure I.8 : MEEF d'un mélange de quatre molécules : Anthracène, naphtalène, benz[α]anthracène et pyrène. Si nous regardons dans la direction X ou Y de la MEEF, nous avons accès respectivement aux spectres d'excitation ou aux spectres d'émission.

Une MEEF est donc un ensemble d'$I_F(\lambda_{ex}, \lambda_{em})$ pour toutes les λ_{ex} et les λ_{em} :

$$MEEF = \{I_F^{i\times j}(\lambda_{ex}^i, \lambda_{em}^j) | i \in [1,n] \; et \; j \in [1,m]\} \quad \text{(I.22)}$$

En écriture matricielle, une MEEF est une matrice $\mathbf{X}$ (n, m) comportant sur ces vecteurs lignes les spectres d'émission (i.e. $I_F(\lambda_{em})$) et sur ces vecteurs colonnes les spectres d'excitation (i.e. $I_F(\lambda_{ex})$). La première (resp. deuxième) dimension de la matrice notée n (resp. m) représente le nombre de vecteurs lignes (resp. colonnes) correspondant au nombre de λ_{ex} (resp. λ_{em}). Ces deux dimensions peuvent être identiques, il s'agira dans ce cas d'une matrice $\mathbf{X}$ carré. Dans le cas de plusieurs échantillons, les MEEF de mêmes dimensions peuvent donc être empilées les unes aux autres pour former un cube de données noté $\underline{\mathbf{X}}(n, m, l)$, où l est la dimension du cube relative au nombre d'échantillon (Figure I.9).

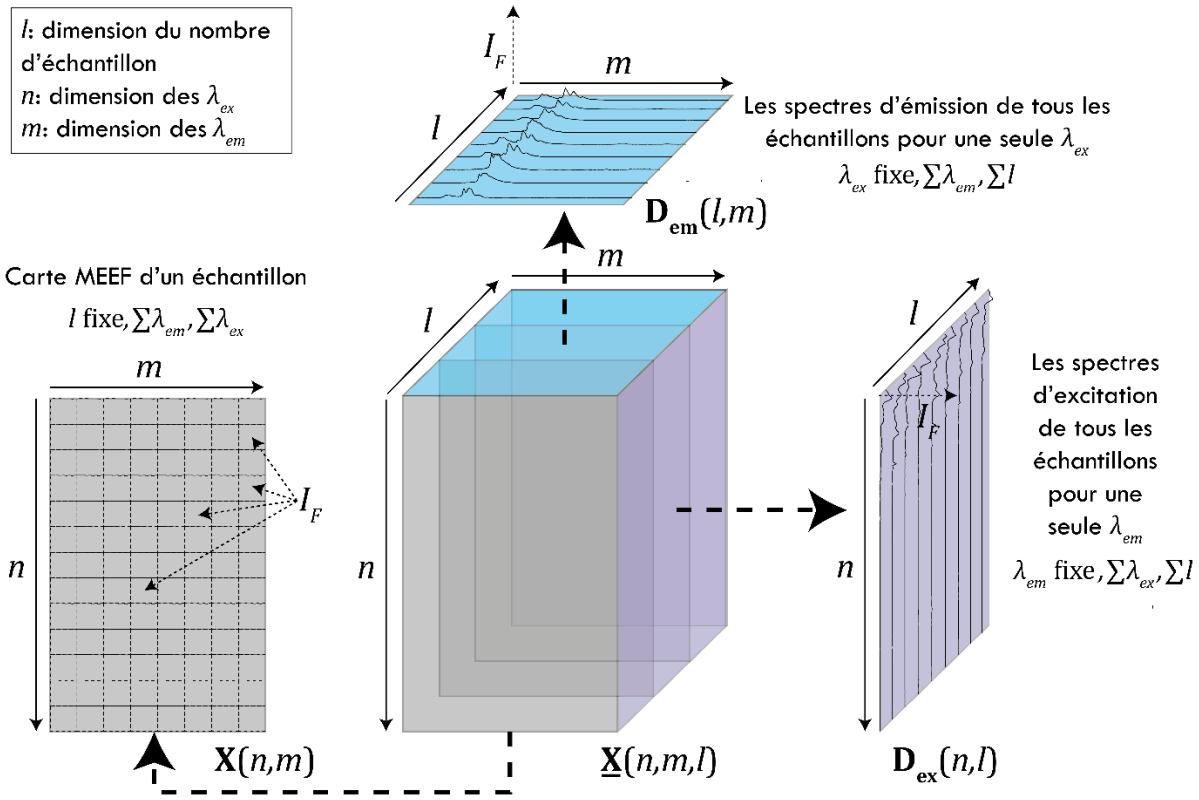


Figure I.9 : Schématisation d'une MEEF et d'un cube de plusieurs MEEF.

I.5.3.1 Instrumentation en fluorescence 3D

Avec les anciens dispositifs expérimentaux de fluorescence 3D dotés d'un monochromateur d'excitation et d'un autre d'émission, cette technique souffrait d'un temps d'analyse long, qui pouvait être optimisé mais au détriment du rapport signal sur bruit. L'arrivée des détecteurs CCD *'charge coupled device'* a permis un gain en efficacité de cette technique puisque le monochromateur d'émission n'est plus nécessaire et le spectre d'émission (i.e. toutes les λ_{em} sur une seule λ_{ex}) est obtenu en une seule mesure. Dans ce cas, les paramètres influant sur la vitesse de la mesure sont la résolution en émission et le pas fixé pour l'excitation. Il est à noter que la plupart de ces instruments permettent la mesure simultanée de l'absorbance en UV-Vis et l'émission de fluorescence (Figure I.10). Cette mesure en absorbance permet notamment la surveillance des conditions de linéarité du signal (cf. §I.4) mais aussi de corriger l'effet du filtre interne. Cet effet et les autres phénomènes physico-chimiques affectant le signal de fluorescence seront décrit dans la prochaine section.

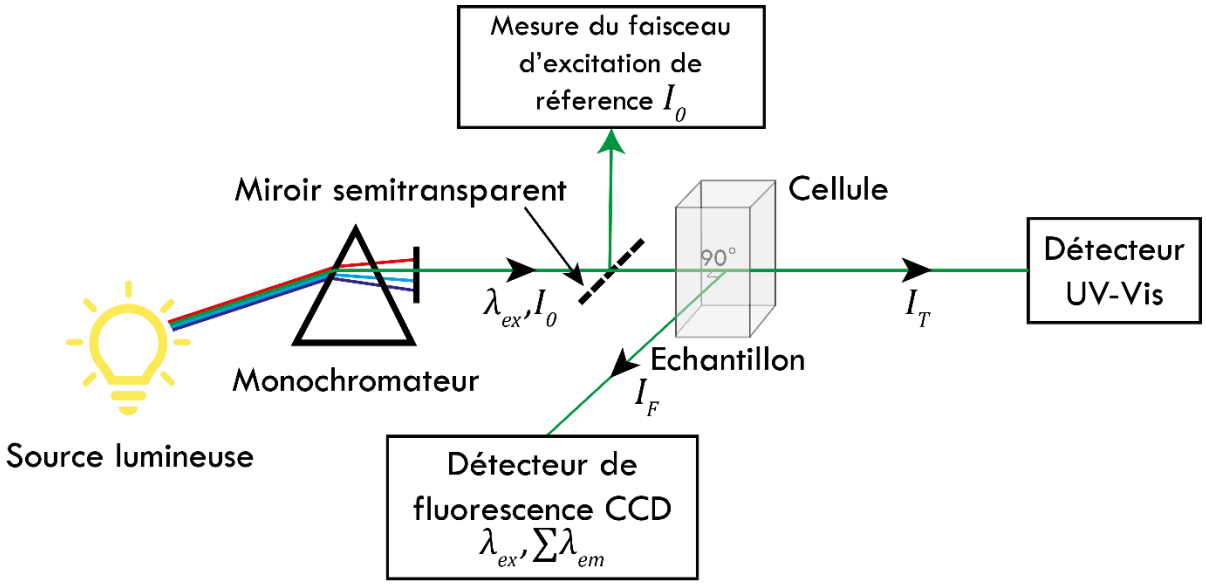


Figure I.10 : Schéma simplifié d'un spectrofluorimètre 3D doté d'un détecteur CCD.

I.6 Les phénomènes physico-chimiques affectant le signal de fluorescence

Différents phénomènes physico-chimiques sont à l'origine de perturbations rencontrées en spectroscopie de fluorescence. Ces derniers peuvent avoir un effet sur la structure, la position ou l'intensité du signal de fluorescence. En premier lieu, la perturbation de la structure du signal est principalement dû aux phénomènes de diffusion de la lumière qui engendrent de fortes interférences au sein des spectres ou des MEEF. En second lieu, le décalage de la position du signal émis est la conséquence de la polarité du solvant ou celle de la formation d'excimères ou d'exciplexes. Enfin, la diminution de l'intensité de fluorescence peut être la conséquence de plusieurs processus responsables de l'extinction de fluorescence. Ces différents phénomènes physico-chimiques sont résumés dans la Figure I.11 et décrits par la suite.

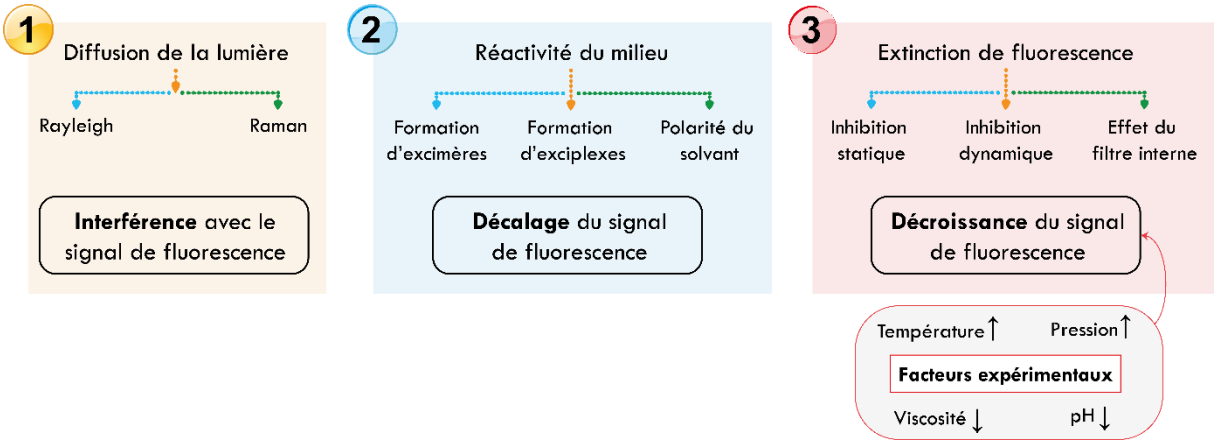


Figure I.11 : Origines des perturbations du signal de fluorescence.

I.6.1 Effet de la diffusion de la lumière

Comme évoqué précédemment (cf. §I.2), l'absorption et l'émission de la lumière ne sont pas les seuls mécanismes mis en jeu lors de l'interaction d'un rayonnement lumineux avec un fluorophore. En effet, la lumière incidente peut également être diffusée par les molécules de manière élastique ou, avec une plus faible probabilité (i.e. 1 photon sur 10^8), de manière inélastique. Dans le cas de la diffusion élastique de la lumière, appelée diffusion Rayleigh, l'énergie des photons diffusés est la même que celle des photons incidents, seule la direction de la lumière est modifiée. Cette diffusion est la conséquence d'une collision entre les photons incidents et les électrons liés à un atome ou à une molécule[27]. Dans le cas de la diffusion inélastique de la lumière, appelée diffusion Raman, l'énergie des photons diffusés est soit plus faible (i.e. Diffusion Raman Stokes), soit plus importante que celle des photons incidents (i.e. Diffusion Raman anti-Stokes). Cette diffusion est la conséquence d'une modification de la polarisation de la molécule par le champ électrique du rayonnement électromagnétique[27] (Figure I.12).

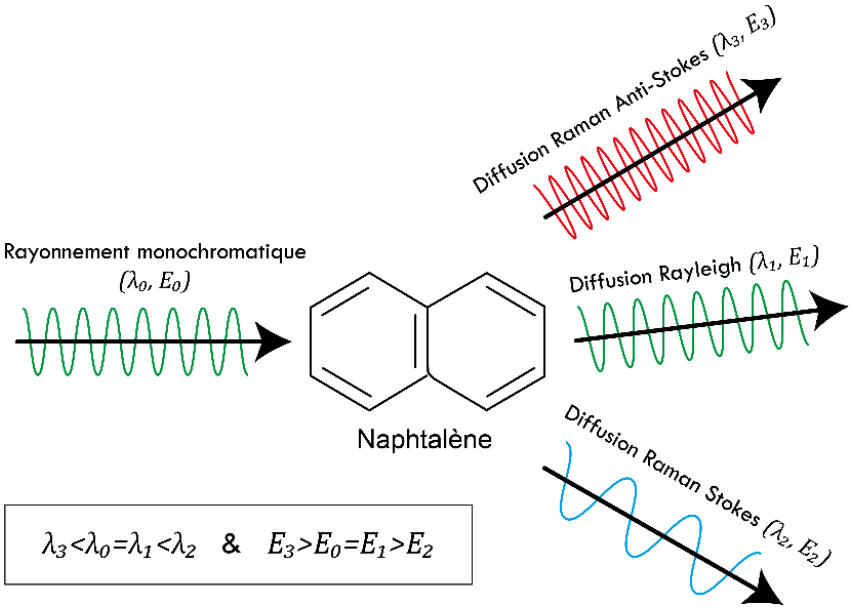


Figure I.12 : Les trois processus de diffusion de la lumière par une molécule.

Le signal Rayleigh sur une MEEF est très intense à cause de sa forte probabilité d'apparition ; 10000 fois supérieur à celle de la diffusion Raman. Il peut être du premier ou du second ordre. Dans le premier cas, il est caractérisé par des longueurs d'onde d'émission très proches des longueurs d'onde d'excitation. Pour le second cas, les longueurs d'onde d'émission sont le double des longueurs d'onde d'excitation. Le signal Raman est quant à lui plus faible mais néanmoins très proche de la zone de fluorescence (Figure I.13).

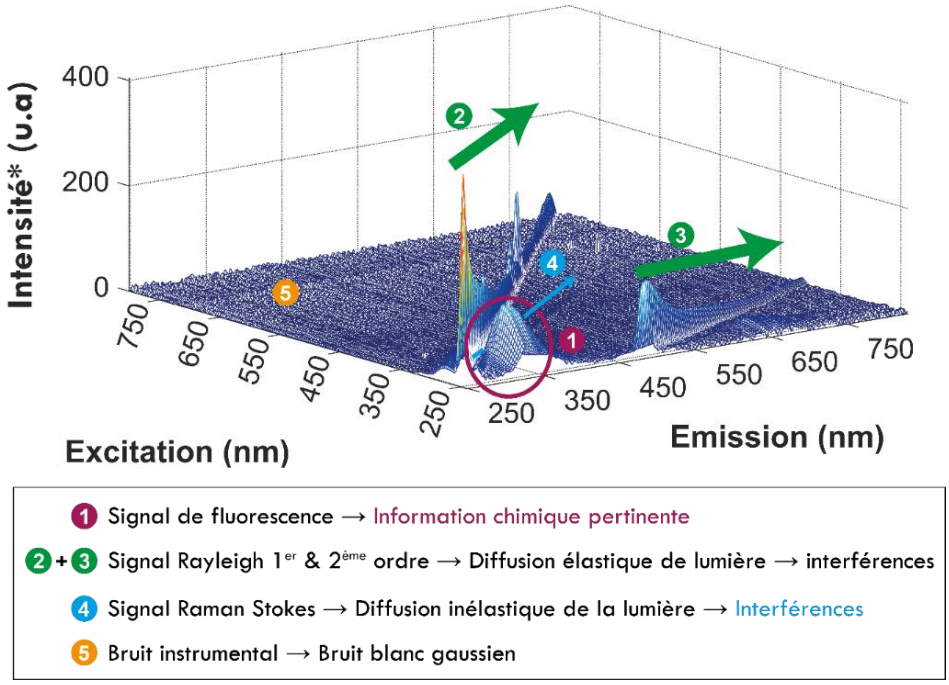


Figure I.13 : Localisation des différents signaux interférents et du bruit dans une MEEF du naphtalène.

Le signal Raman Stokes est celui qui est observé. En effet, pour qu'une molécule puisse diffuser en Raman anti-Stokes, il faut qu'elle soit initialement (i.e. avant irradiation) dans un état vibrationnel excité ce qui est peu fréquent. Le signal Raman Stokes se caractérise par des longueurs d'onde d'excitation toujours plus courtes que les longueurs d'onde d'émission (i.e. perte d'énergie) (Figure I.14).

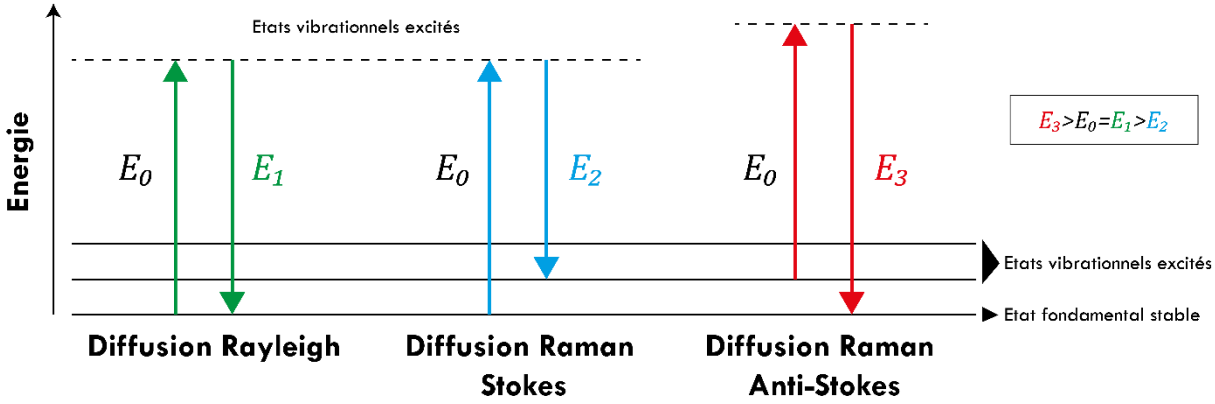


Figure I.14 : Diagramme de Perrin-Jablonski illustrant les trois proccessus de diffusion de la lumière par une molécule[48].

I.6.2 Effet de la polarité du solvant

La polarité du solvant engendre des décalages dans la position d'émission d'un fluorophore (i.e. accentuation du décalage de Stokes). Cet effet est dû à une stabilisation de l'état excité par les molécules du solvant polaire. Cette stabilisation de l'état excité est d'autant plus importante que la polarité du solvant est élevée. Elle induit une émission à une énergie plus faible et donc un décalage de l'émission vers des λ_{em} plus grandes. Généralement, ce

sont les fluorophores polaires qui sont le plus impactés par cet effet de la polarité du solvant. Les molécules non polaires comme les HAP sont beaucoup moins sensibles à la polarité du solvant[28].

I.6.3 Effet de la formation d'excimères ou d'exciplexes

Le terme excimère (resp. exciplexe) est issu de la contraction du *mot 'exc**ited** di**mer**'* (resp. *'excited complex'*). Il s'agit donc de dimères (resp. complexes) à l'état excité. Un excimère (resp. exciplexe) se forme par collision entre une molécule excitée et une molécule identique (resp. différente), non excitée. En conséquence, l'énergie d'excitation portée initialement par la molécule excitée, se délocalise sur les deux molécules qui forment soit un excimère, soit un exciplexe. Ainsi, l'émission de fluorescence se voit décaler vers de plus grandes longueurs d'ondes. Les effets de ce phénomène s'observent aux fortes concentrations puisqu'il faut suffisamment de molécules dans le milieu pour que des rencontres puissent se faire pendant la courte durée de vie de l'état excité S_1 (i.e. de 10^{-10} à 10^{-7} s). Par exemple, le naphtalène et le pyrène sont deux molécules capables de former des excimères. Par ailleurs, l'anthracène est une molécule connue pour former des exciplexes avec des molécules amines[44].

I.6.4 Extinction de la fluorescence

Un premier type d'extinction de fluorescence, appelé inhibition dynamique, se produit par la collision entre le fluorophore à l'état excité et une molécule en solution qualifiée d'extincteur. Cette collision engendre une désactivation de l'état excité du fluorophore par un retour diffusif vers l'état fondamental[27,28]. Beaucoup de composés agissent comme extincteurs de fluorescence. L'oxygène moléculaire est l'un des extincteurs collisionnels les plus connus[28]. Les atomes lourds tels que l'ion iodure (I^-) et l'ion bromure (Br^-) sont aussi connus pour engendrer une inhibition dynamique[44]. Un deuxième type d'extinction de fluorescence, appelé inhibition statique, se produit par la formation d'un complexe non fluorescent entre le fluorophore et un extincteur ce qui le neutralise et engendre une diminution de l'I_F[27,28]. L'effet du filtre interne est également responsable de l'extinction de fluorescence. Il s'agit d'une réabsorption (resp. absorption) interne par certains fluorophores (resp. chromophores), de la fluorescence émise par d'autres (resp. de la lumière incidente). Dans les deux cas, le détecteur ne reçoit pas tous les photons émis, ce qui se traduit par la baisse de l'I_F. Par ailleurs, plusieurs facteurs expérimentaux peuvent accentuer le phénomène d'extinction de fluorescence (i.e. décroissance de l'I_F) tels que la baisse du pH ou de la viscosité du milieu. La température et la pression sont quant à eux inversement proportionnelles à l'I_F[27].

Bien que les phénomènes physico-chimiques qui influencent le signal de fluorescence soient nombreux, la fluorescence demeure une technique analytique très sensible, sélective et facile à mettre en œuvre[28]. Cette technique est prometteuse pour l'analyse des composés chimiques fluorescents tels que certains CAP dont beaucoup de HAP, particulièrement en anticipant les effets parasites et les conditions de linéarité.

I.7 Propriétés des composés aromatiques polycycliques en fluorescence

Comme dans le cadre de ce travail de thèse, les CAP (et en particulier les HAP) sont ciblés (cf. §Introduction), cette partie du chapitre présente les grandes caractéristiques des signaux de fluorescence de ces composés. Les CAP sont des systèmes moléculaires conjugués présentant des transitions $n \rightarrow \pi^*$ et/ou $\pi \rightarrow \pi^*$ permises dans la gamme de l'UV-Vis. C'est pourquoi, la plupart d'entre eux sont d'excellents chromophores (cf. §I.3.1) et sont donc potentiellement de bons fluorophores. Leur grande diversité chimique rend néanmoins, difficile toute généralisation quant à leur rendement et leurs caractéristiques de fluorescence. Toutefois, dans le cadre de la classification adoptée dans ce travail (i.e. HAP, NSO-CAP et leurs dérivés, cf. §Introduction), certaines caractéristiques propres à chaque sous-catégorie peuvent être citées.

I.7.1 Propriétés des HAP en fluorescence

Les HAP sont des molécules planes et rigides avec un système d'électrons π et donc des transitions $\pi \rightarrow \pi^*$ permises dans la gamme de l'UV-Vis (i.e. la transition la plus basse en énergie de ces molécules, Figure I.2). Ces molécules se caractérisent, grâce à ce type de transition, par des coefficients d'absorption molaires et des rendements de fluorescence élevés. Par ailleurs, ces composés présentent un certain nombre de caractéristiques[44–46] :

- Beaucoup de HAP émettent de la fluorescence dans la région de l'UV-Vis.
- Au fur et à mesure que le nombre de cycles aromatiques s'accroit (i.e. extension du système π et donc du degré de conjugaison), l'émission de fluorescence des HAP se décale vers de plus grandes longueurs d'onde (i.e. plus basse d'énergie)[49]. Le rendement de fluorescence est également proportionnel au degré de conjugaison[50].
- De manière générale, les HAP non linéaires (e.g. phénanthrène) absorbent et fluorescent à de courtes longueurs d'onde (i.e. haute énergie) comparés aux composés linéaires comme l'anthracène, qui est un isomère linéaire du phénanthrène (Figure I.15).

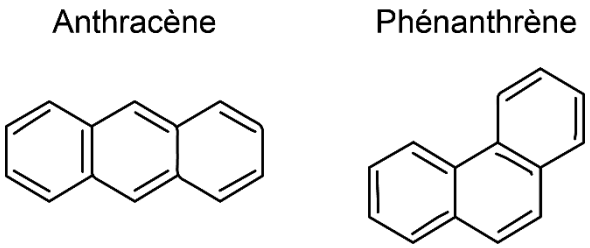


Figure I.15 : Structures chimiques de l'anthracène et du phénanthrène.

- Le spectre d'émission d'un HAP en solution est résolu en se caractérisant par la présence de plusieurs bandes d'émission, relatives aux niveaux vibrationnels (Figure I.5). Dans le cas d'un mélange complexe de HAP, cette résolution peut être perdue à cause du recouvrement spectral.

I.7.2 Propriétés des NSO-CAP en fluorescence

Contrairement aux HAP et leurs dérivés, les caractéristiques de fluorescence des NSO-CAP, à l'exception des azaarènes, ont été très peu étudiées. Pour rappel, les azaarènes, les oxaarènes et les thiazaarènes, qui composent cette catégorie, sont des composés aromatiques contenant au sein d'un ou de plusieurs de leurs cycles aromatiques un ou plusieurs hétéroatomes d'azote, d'oxygène ou de soufre, respectivement. La présence de l'hétéroatome implique souvent une transition de plus basse énergie (la transition $n \rightarrow \pi^*$) avec pour conséquence, un faible rendement de fluorescence. Néanmoins, ces composés restent solvant dépendant. Leur rendement de fluorescence augmente significativement quand ils sont dans des solvants polaires capables de former des liaisons hydrogènes tels que les alcools[44].

Lorsqu'un azote, un oxygène ou un soufre est lié avec une liaison σ à des atomes de carbone dans un hétérocycle (e.g. cycles pyrrole, furane et thiophène, respectivement) les transitions, impliquant les électrons non liés, ont des propriétés similaires à celles des transitions $\pi \rightarrow \pi^*$ et non à celles des transitions $n \rightarrow \pi^*$. Cette propriété est dû à la géométrie moléculaire particulière des orbitales non liantes. Cela explique aussi le rendement de fluorescence relativement élevé de ces composés tels que le carbazole ou le benzo[α]carbazole pour les azaarènes, le dibenzofuran pour les oxaarènes et le dibenzothiophène pour les thiazaarènes[45] (Figure I.16).

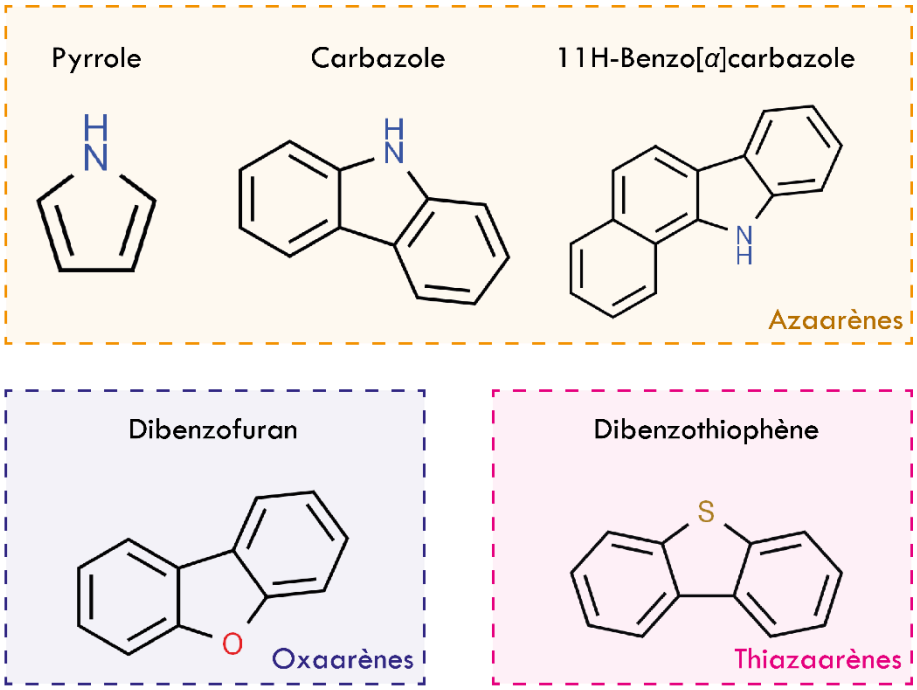


Figure I.16 : Structures chimiques de quelques NSO-CAP. Les structures chimiques sont issues de la base de données Chemspider (https://chemspider.com).

I.7.3 Propriétés de fluorescence des dérivés des HAP et des dérivés des NSO-CAP

Il est prudent, dans le cas des dérivées des HAP et des NSO-CAP, de ne pas faire de généralisation quant à leurs caractéristiques de fluorescence. En effet, plusieurs facteurs tels que la nature, le nombre et la position du (des) groupement(s) substituant(s) ou encore la nature du solvant influencent la réponse en fluorescence. De plus, l'effet d'une mono substitution ne peut être extrapolé, sans précaution, à un composé poly-substitué. De manière générale, la présence d'un hétéroatome qui serait impliqué dans le système π peut conduire à la permission de la transition $n \rightarrow \pi^*$ de plus basse énergie, au détriment de la transition $\pi \rightarrow \pi^*$. Ces transitions se caractérisent par des coefficients d'absorption molaire 100 fois plus petits et une cinétique de désexcitation plus faible que les transition $\pi \rightarrow \pi^*$. Par conséquent, le rendement de fluorescence des molécules est plus faible puisque les mécanismes de désexcitation non radiatifs deviennent prédominants[44]. Ci-dessous sont citées, de manière non exhaustive, quelques caractéristiques en fonction du groupement fonctionnel substituant :

I.7.3.1 Groupement carbonyle

L'effet de la substitution par un groupement carbonyle ne peut être généralisée. En effet, cette substitution induit pour certains composés (e.g. anthrone) la permission de la transition $n \rightarrow \pi^*$ et donc un déclin de fluorescence. Pour d'autres composés (e.g. fluorénone), la transition de plus basse énergie reste $\pi \rightarrow \pi^*$ malgré la substitution carbonyle. Par conséquent, le rendement de tels composés reste élevé. Pour d'autres composés (e.g.

xanthone) caractérisés par une transition $n \rightarrow \pi^*$ dont l'énergie est légèrement supérieure à la transition $\pi \rightarrow \pi^*$, la polarité du solvant sera déterminante. En effet, au fur et à mesure que la polarité du solvant augmente, des liaisons hydrogène se forment induisant une baisse d'énergie des transitions $\pi \rightarrow \pi^*$. Ces dernières deviennent de plus basse énergie que les transitions $n \rightarrow \pi^*$ en engendrant une augmentation du rendement de fluorescence[44,46] (Figure I.17). Un tableau répertoriant différents solvants organiques couramment utilisés pour l'extraction et/ou la solvatation des HAP, triés selon leur polarité, est disponible en annexe A.

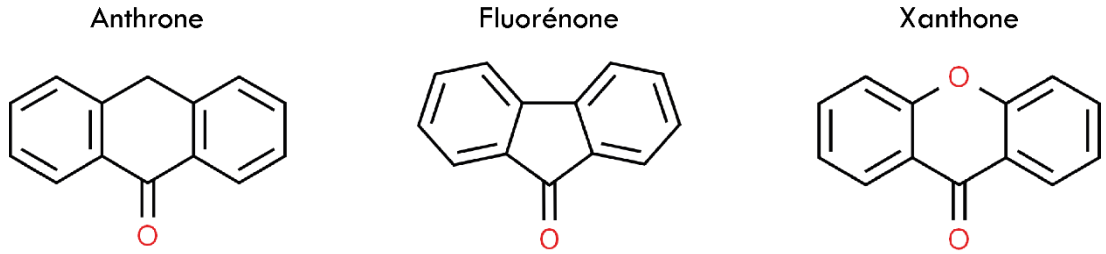


Figure I.17 : Structure chimique de l'anthrone, du fluorénone et du xanthone. Les structures chimiques sont issues de la base de données Chemspider (https://chemspider.com).

I.7.3.2 Groupement nitro

La substitution par un groupement NO_2 provoque généralement un déclin sévère de fluorescence et a contrario, un gain en efficacité de phosphorescence. Néanmoins, certains composés nitrés ne sont ni fluorescents ni phosphorescents. Ces derniers se caractérisent par un retour non radiatif vers l'état fondamental (i.e. conversion interne efficace de S_1 vers S_0). Ainsi, pour ce type de composés, il s'avère nécessaire de compléter une caractérisation approfondie de la contamination par les CAP en utilisant, d'autres techniques complémentaires à la spectroscopie de fluorescence telles que la spectroscopie Raman. Cependant, il est important de noter que de nombreux composés nitrés sont sensibles au rayonnement lumineux et se photodégradent facilement[44,45] (e.g. le 9-nitroanthracène, Figure I.18).

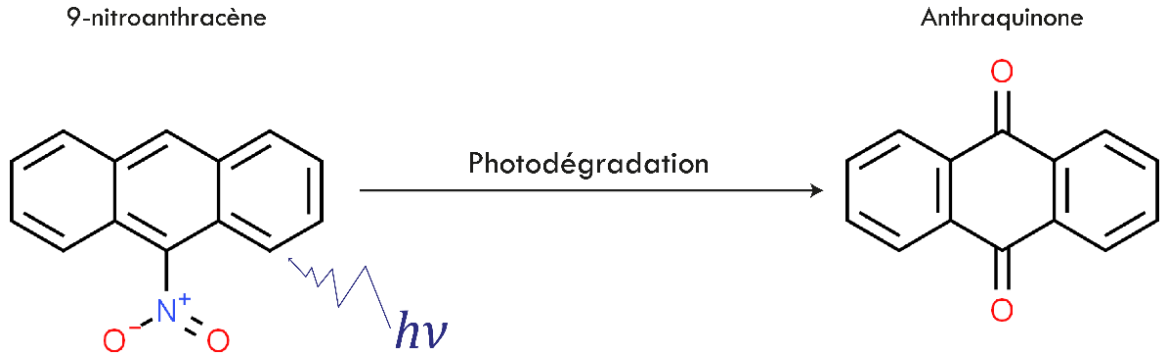


Figure I.18 : Photodégradation du 9-nitroanthracène en anthraquinone. Les structures chimiques sont issues de la base de données Chemspider (https://chemspider.com).

I.8 Contextualisation de la thèse et conclusion

I.8.1 Contextualisation de la thèse

Pour réaliser une caractérisation aussi exhaustive que possible, à la fois qualitative et quantitative, de la pollution environnementale par les CAP, diverses techniques analytiques peuvent être envisagées. Parmi elles, on compte la chromatographie pour la séparation à l'échelle moléculaire des CAP, la spectrométrie de masse pour leur identification et quantification, ainsi que la spectroscopie pour obtenir des empreintes moléculaires globales de cette pollution.

Dans la littérature, les méthodes chromatographiques liquides (LC) et gazeuses (GC), souvent couplées à la spectrométrie de masse (MS), sont les plus fréquemment utilisées pour l'analyse des CAP[26,51] en raison de leur efficacité dans l'analyse des 16 HAP réglementaires et de leur standardisation, et ce malgré le fait qu'elles soient souvent plus coûteuses et chronophages que les méthodes spectroscopiques. Beaucoup de travaux s'appuient donc sur ces techniques pour l'étude des HAP.

Malgré leur grand potentiel, les méthodes spectroscopiques sont moins fréquemment utilisées pour l'analyse des CAP en raison de l'absence d'une séparation préalable de ces composés, ce qui conduit à l'obtention d'un signal global complexe. De plus, des obstacles techniques tels que le manque de puissance et l'instabilité des lasers ou encore l'absence de détecteurs *'Charge Coupled Device'* (CCD), ont conduit, par le passé, à un intérêt limité pour ces techniques s'appuyant sur l'interaction entre la lumière et la matière pour l'étude des CAP. De nos jours, grâce aux avancées technologiques, l'intérêt pour la spectroscopie devient nettement plus important et le nombre de travaux ciblant les CAP augmente ces dernières années. Par exemple, le potentiel de la spectroscopie Raman pour l'analyse des HAP a été discuté dans deux études théoriques simulant des spectres Raman de plusieurs HAP[52,53]. Ces études ont ouvert la voie à des études Raman expérimentales, telles que celle menée sur 48 HAP différents, dans le but de déterminer dans quelle mesure cette spectroscopie pouvait être utilisée pour identifier de manière individuelle les différentes espèces. Cette étude discute également les facteurs qui contrôlent les positions des pics Raman majeurs, de l'impact de la fluorescence induite sur l'intensité des pics Raman, ainsi que des effets de la variation de la longueur d'onde d'excitation sur les spectres Raman de certains HAP[54]. La spectroscopie en réflectance dans le domaine visible et proche infrarouge est une autre technique spectroscopique qui a été utilisée pour étudier 47 poudres fines de CAP à l'état solide. Cette étude a montré la faisabilité de l'étude des CAP par cette technique, ainsi que les propriétés spectrales des CAP, qui varient en fonction du nombre et de la connectivité des cycles aromatiques fusionnés, ainsi que de la présence et du type d'hétérocycles et de groupements

alkyles ou fonctionnels[55]. En outre, une autre étude propose une méthodologie analytique utilisant la spectroscopie infrarouge à réflexion totale atténuée (ATR-FTIR), exploitant le domaine électromagnétique du moyen infrarouge, pour quantifier rapidement divers HAP présents dans l'eau[56].

La spectroscopie de fluorescence est une autre méthode adaptée à l'analyse des CAP et en particulier des HAP, en raison de leur nature aromatique qui induit leur fluorescence intrinsèque. Il existe trois configurations expérimentales en spectroscopie de fluorescence : classique, synchrone et fluorescence 3D. Plusieurs études ont été menées pour analyser les HAP en utilisant la fluorescence classique ou la spectroscopie synchrone. La première est généralement employée pour l'analyse de composés purs, ainsi que comme système de détection après séparation chromatographique[57]. Cependant, elle peut également être utilisée en conjonction avec la microscopie confocale, comme cela a été fait dans une étude portant sur l'analyse de sols contaminés en laboratoire par 4 HAP[58]. La spectroscopie de fluorescence synchrone fournit une caractérisation plus complète que la fluorescence classique. Par exemple, elle a été utilisée pour l'analyse structurale de la partie organique de deux types de charbon de rang différent. L'étude a été menée sur différents extraits obtenus à l'aide de divers solvants organiques. Elle a mis en évidence les différences entre ces charbons et a fait ressortir les caractéristiques spectrales de chacun d'entre eux. De plus, elle a permis de détecter les changements structuraux survenant dans la partie aromatique des charbons après déminéralisation ou oxydation[49]. La fluorescence 3D est la seule configuration expérimentale en spectroscopie de fluorescence à pouvoir fournir une caractérisation globale de l'échantillon étudiée. Toutefois, jusqu'à récemment, les appareils permettant les analyses, n'était pas équipés de détecteur CCD, induisant des temps de mesure longs (une demi-journée pour un échantillon). C'est pourquoi les analyses en spectroscopie de fluorescence classique ou synchrone étaient les plus couramment réalisées. Aujourd'hui, avec les détecteurs CCD, les analyses par fluorescence 3D deviennent nettement plus fréquentes, mais leur exploitation reste néanmoins freinée en raison de la grande quantité de données que génère cette technique, de la structure tridimensionnelle de celles-ci, ainsi que de la complexité des signaux globaux obtenus[27]. C'est ainsi que la chimiométrie s'avère essentielle pour ce type et cette quantité de données[59]. Dans le chapitre II, nous exposerons donc en détail, ces verrous inhérents à la mesure en fluorescence 3D et nous proposerons des méthodes chimiométriques permettant de les lever.

I.8.2 Conclusion

Dans ce chapitre d'état de l'art, les notions générales d'interaction lumière-matière ont été abordées, et une revue des aspects théoriques et pratiques de deux spectroscopies

électromagnétiques complémentaires, à savoir la spectroscopie d'absorption UV-Vis et la spectroscopie d'émission de fluorescence, a été présentée. Parmi les différentes configurations expérimentales en spectrofluorimétrie possible, la fluorescence 3D, semble particulièrement adaptée pour une caractérisation plus globale des échantillons grâce à l'acquisition de spectres d'émission et d'excitation donnant accès à une « carte » ou une « image » pseudo-chimique véhiculant des informations sur les fluorophores contenus dans l'échantillon complexe analysé.

De plus, les propriétés chimiques des CAP en fluorescence permettant de caractériser leurs différentes catégories, à savoir, les HAP et les NSO-CAP, mais également leurs dérivés associés ont été exposées. Néanmoins, le signal mesuré issus de l'instrument peut s'avérer complexe à interpréter avec les différents mécanismes et phénomènes physico-chimiques qui peuvent l'influencer ou le perturber. Les phénomènes de diffusion Rayleigh, de diffusion Raman ou de faible rapport signal sur bruit liés à la nature des échantillons environnementaux, sont autant de verrous à lever pour permettre une caractérisation chimique de la pollution la plus exhaustive possible. En outre, la méthode de fluorescence 3D est une technique analytique qui produit une quantité importante de données, parfois non linéaires, pouvant également être un frein ou dans le pire des cas mener à des interprétations biaisées de la réalité chimique.

Dès lors, il devient évident et même indispensable de se tourner vers des stratégies efficaces de gestion et de traitement des données avec la chimiométrie.

CHAPITRE II

ANALYSES DES SIGNAUX DE FLUORESCENCE 3D PAR CHIMIOMETRIE

II.1 Introduction

Les laboratoires de chimie analytique sont de nos jours équipés d'une multitude d'instruments aux technologies avancées, générant des quantités importantes de données de natures multiples et multi-échelle. Ces dernières sont enregistrées, traitées et analysées via des systèmes informatiques performants. Dans les années soixante-dix, le scientifique suédois Svante Wold et le scientifique américain Bruce Kowalski ont compris qu'il serait de plus en plus nécessaire de savoir manipuler une masse importante de données, particulièrement avec l'avènement des ordinateurs dans les laboratoires. C'est ainsi qu'est née la discipline scientifique appelée chimiométrie. Ce nom vient du mot anglais '*chemometrics*', inventé par Svante Wold, il est apparu dans la littérature scientifique, pour la première fois, en 1975[60]. Svante Wold et Bruce Kowalski ont rapidement compris l'importance de cette nouvelle discipline et ont ainsi fondé en 1974 la Société Internationale de Chimiométrie, connue sous le nom de '*International Chemometrics Society*'. Ces deux chercheurs sont considérés comme les pères fondateurs de la discipline[61].

Les méthodes de chimiométrie d'analyse des MEEF peuvent être divisées, dans le cas de ce travail de thèse, en trois catégories principales : les méthodes de prétraitements, les méthodes d'analyse et les méthodes d'apprentissage quantitatif. Dans un premier temps, les prétraitements en chimiométrie ont pour objectif de corriger les données spectrales d'interférences inhérentes à la mesure de l'échantillon analysé, du bruit provenant de l'instrumentation, mais également, des phénomènes physiques gênant ou masquant l'information chimique pertinente. Cette étape est cruciale pour préparer les données en vue de leurs analyses ultérieures. Dans un deuxième temps, les méthodes d'analyse s'intéressent à la visualisation et à l'exploration des données ainsi qu'à l'identification de structures particulières (e.g. l'état d'avancement d'un processus chimiques). Parmi ces méthodes, nous retrouvons la famille des méthodes exploratoires de réduction de dimensionalité telles que l'analyse en composantes principales (ACP) et la famille des méthodes de résolution multivariée de courbes (i.e. décomposition spectrale bilinéaire ou trilinéaire). Ces dernières visent à séparer les signaux sources relatifs, idéalement, aux espèces chimiques composant un mélange complexe. Les plus populaires d'entre elles sont la méthode de résolution multivariée des courbes par moindres carrés alternés *'Multivariate Curve resolution-Alternating Least Squares'* (MCR-ALS) pour les données multivariées[62] et la méthode d'analyse des facteurs parallèles *'PARAllel FACtor'* (PARAFAC) pour les données multivoies[63] (e.g. MEEF). Enfin, les méthodes d'apprentissage permettent de construire des modèles mathématiques de régression linéaire ou non linéaire qui capturent les relations complexes entre des informations spectrales (e.g. MEEF) et des informations de référence qualitatives ou quantitatives (e.g. origine géographique de chaque échantillon ou encore concentration d'une espèce chimique

d'intérêt)[64]. Dans le cas des analyses quantitatives, l'objectif est d'utiliser un modèle de régression construit, optimisé et validé, pour effectuer des prédictions sur les concentrations des espèces chimiques d'intérêts à partir uniquement de données spectrales et/ou de chromatogrammes. Les approches les plus fréquemment utilisées pour modéliser les données multivariées ou multivoies incluent l'utilisation de la régression par les moindres carrés partiels '*Partial Least Squares Regression'* (PLSR ou N-PLSR pour les données multivoies) pour les modèles linéaires[65], ainsi que la régression par vecteurs de support '*Support Vector Regression*' (SVR) pour les modèles linéaires ou non linéaires[66].

Selon l'objectif visé et la nature des données à analyser, il est tout à fait possible de combiner des méthodes d'analyse et d'apprentissage pour maximiser et modéliser au mieux l'information chimique la plus pertinente extraite de données telles que les MEEF. Ainsi, ce chapitre d'état de l'art débutera par une description des caractéristiques numériques des MEEF, notamment sur l'aspect trilinéarité quand l MEEF sont compilées en un cube de données (i.e. $\underline{\mathbf{X}}(l,m,n)$, cf. §I.5.3) et leur éventuelle restructuration bilinéaire notamment lors de l'augmentation matricielle (i.e. restructuration d'un cube de MEEF en une seule matrice augmentée). Ce point est important car il conditionne la compréhension et l'application de différentes méthodologies en chimiométrie lors de l'analyse et de l'apprentissage de MEEF. En effet, la majorité des méthodes de réduction de dimensionalité, d'analyse de mélange et de régression (e.g. ACP, MCR-ALS, PARAFAC et PLSR), sont applicables en spectroscopie, de par, l'approximation linéaire de la loi de Beer-Lambert.

II.2 Caractéristiques des MEEF

Nous avons vu précédemment que l'intensité de fluorescence mesurée ($I_F(\lambda_{ex},\lambda_{em})$) en spectrofluorimétrie 3D est liée linéairement à la concentration c d'un fluorophore lorsque la condition de la loi de Beer-Lambert est vérifiée (i.e. $A_{ex} < 0.05$, cf. §I.4.2). Les MEEF compilées en un cube de données (i.e. espace tridimensionnel) et dont les signaux respectent cette condition, sont dites trilinéaires.

II.2.1 Trilinéarité des MEEF

Prenons d'abord le cas simple d'un échantillon pour lequel nous réalisons une mesure d'un spectre d'émission (i.e. mesure de m $I_F(\lambda_{em})$, avec λ_{ex} fixe et m correspondant au nombre de λ_{em}). La donnée obtenue est un vecteur $\mathbf{d}(1,m)$ portant les éléments d_j qui correspondent aux $I_F(\lambda_{em})_j$ avec $j = 1,\dots,m$ (Figure II.1a). Le vecteur $\mathbf{d}$ est qualifié de système linéaire car ces éléments (d_j) satisfont, par analogie à l'équation (I.12) du chapitre I, à la relation[67] :

$$d_j = I_F(\lambda_{em})_j = \alpha_1 c_j + \alpha_0 + e_j \quad \text{(II.1)}$$

Où α_1 est un coefficient multiplicatif, α_0 est un coefficient additif et e_j le terme résiduel de l'ajustement. Si nous étendons ensuite le système à plusieurs l échantillons, la donnée obtenue est alors une matrice $\mathbf{D}(l,m)$ regroupant les l différents spectres d'émission et comportant les éléments d_{ij} avec $i = 1,\dots,l$ et $j = 1,\dots,m$ (Figure II.1b). Cette matrice est dite bilinéaire si la relation linéaire est satisfaite dans les deux directions de la matrice (i.e. direction l pour les échantillons et direction m pour les λ_{em}). En d'autres termes, pour un système avec F fluorophores, chaque éléments d_{ij} de la matrice $\mathbf{D}$ doit satisfaire la relation :

$$d_{ij} = I_F(\lambda_{em})_{ij} = \left\{\sum_{f=1}^{F} c_{if} s_{fj}\right\} + e_{ij} \quad \text{(II.2)}$$

Où c_{if} représente la concentration du fluorophore f pour l'échantillon i et s_{fj} son absorptivité à la $j^{ème}$ λ_{em}. e_{ij} est le terme résiduel de l'ajustement. Enfin, la MEEF est un cas particulier de matrice ayant sur ces deux dimensions des variables spectrales (i.e. dimension n pour les λ_{ex} et m pour les λ_{em}) (Figure II.1c). Une MEEF ne peut être qualifiée de système bilinéaire. Cependant, l MEEF compilées en un cube de données $\underline{\mathbf{X}}(l,m,n)$ est un système trilinéaire de F fluorophores, quand les éléments x_{ijk} de ce dernier, avec $i = 1,\dots,l, j = 1,\dots,m$ et $k = 1,\dots,n$ (Figure II.1d), satisfont à la relation[59] :

$$\forall_{i,j,k} \mid x_{ijk} = I_F(\lambda_{ex},\lambda_{em})_{ijk} = \sum_{f=1}^{F} a_{if}\, b_{jf} c_{kf} + e_{ijk} \quad \text{(II.3)}$$

Où a_{if} représente la concentration du fluorophore f pour l'échantillon i. b_{jf} et c_{kf} représentent les valeurs des profils d'émission et d'excitation à la $j^{ème}$ λ_{em} et à la $k^{ème}$ λ_{ex}, respectivement. e_{ijk} est le terme résiduel de l'ajustement.

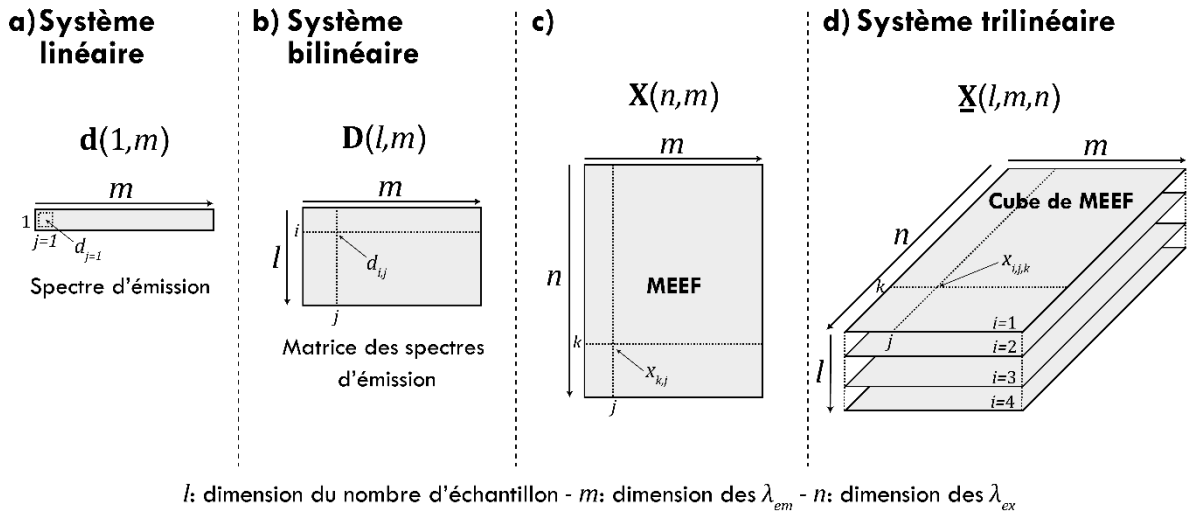


Figure II.1 : Représentation schématique d'un vecteur (a), d'une matrice (b), d'une MEEF (c) et d'un cube de MEEF (d).

Plusieurs méthodes de chimiométrie ne sont pas adaptables directement au traitement de données trilinéaires, mais seulement au traitement de données bilinéaires (e.g. ACP, PLS et MCR-ALS). Ainsi, comme évoqué précédemment, en effectuant certaines étapes de restructuration du cube des MEEF $\underline{\mathbf{X}}(l,m,n)$, nous pouvons obtenir une matrice augmentée qui peut être considérée comme un système bilinéaire et qui peut être traitée par ces méthodes. Il convient ainsi de choisir la méthode de restructuration du cube de MEEF la plus adaptée pour un traitement avec une méthode d'analyse de données multivariées bilinéaires. Ces dernières ont, en effet, été développées en chimiométrie pour être appliquées sur des matrices contenant sur leurs lignes des données chimiques d'échantillons ou de plusieurs observations (i.e. spectres, chromatogrammes, etc.) et sur leurs colonnes les variables associées aux signaux (i.e. longueurs d'onde, temps de rétention, etc.), ce qui est, par exemple, le cas de la matrice $\mathbf{D}(l,m)$ qui pour rappel, comporte l spectres d'émission chacun contenant m λ_{em}.

II.2.2 Restructuration d'un cube de MEEF en une matrice augmentée

Dans le cas du cube $\underline{\mathbf{X}}(l,m,n)$, trois types de restructuration peuvent être envisagés. Pour les deux premiers cas, le cube des MEEF est d'abord déplié en un ensemble de matrices $\mathbf{X}_l(n,m)$. Ensuite, ces matrices (i.e. MEEF) sont concaténées en ligne dans le premier cas pour former une matrice augmentée $\mathbf{D_1}(n, l \times m)$ ou en colonne dans le deuxième cas pour former une matrice augmentée $\mathbf{D_2}(l \times n, m)$ (Figure II.2). Ces deux configurations ne sont pas idéales car la dimension l relative aux échantillons ne représente ni les lignes ni les colonnes de la matrice augmentée.

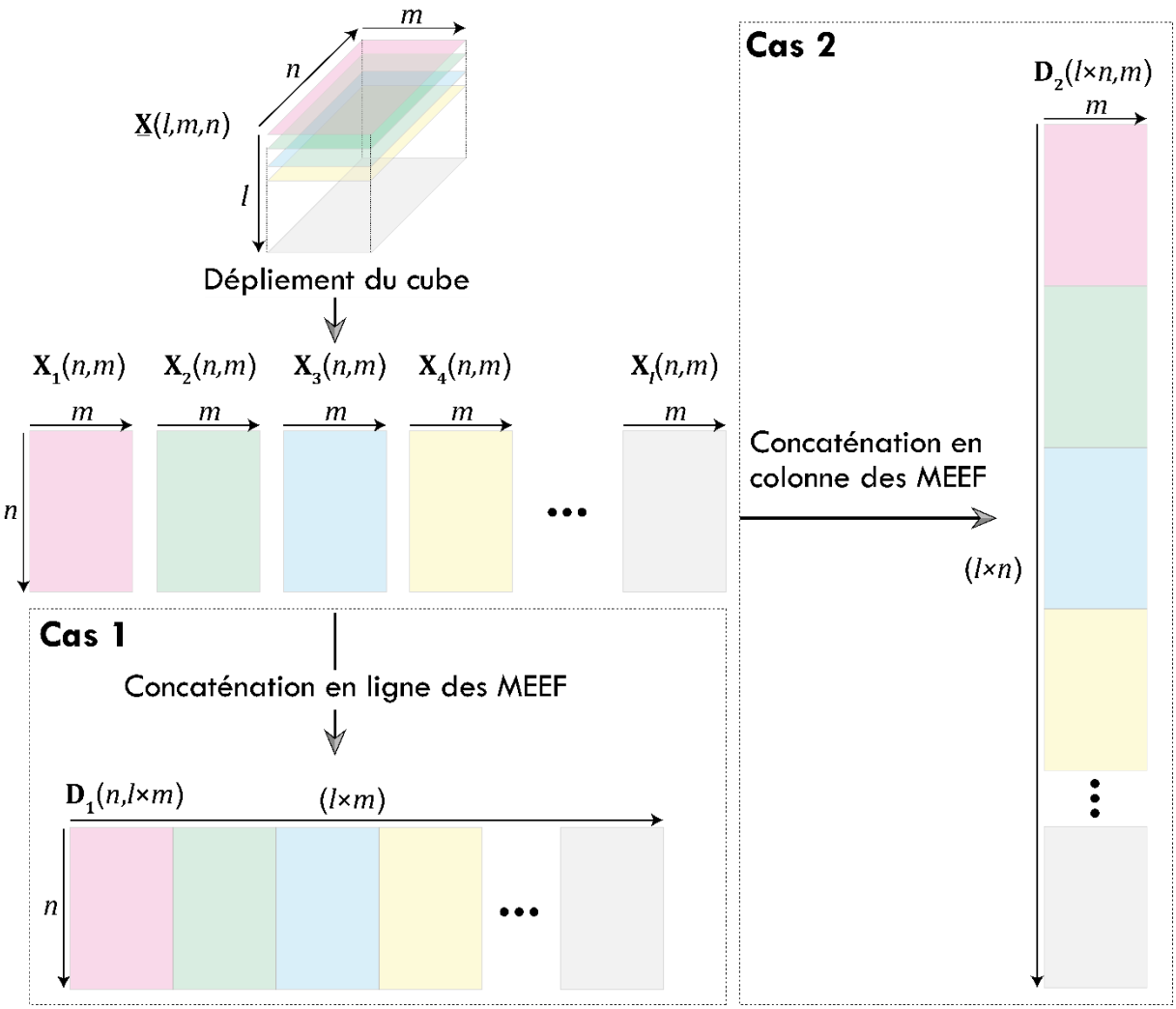


Figure II.2 : Dépliement du cube $\underline{\boldsymbol{X}}(l, m, n)$ et concaténation, en ligne pour le cas 1 et en colonne pour le cas 2, des MEEF.

Pour le troisième cas, le cube des MEEF est aussi déplié en un ensemble de matrices $\mathbf{X}_l(n, m)$ ou $\mathbf{X}_l(m, n)$. Cependant, contrairement aux deux premiers cas, chaque MEEF est ensuite déplié à son tour en un ensemble de vecteurs $\mathbf{x}_n(1, m)$ ou $\mathbf{x}_m(1, n)$. Ces vecteurs sont ensuite concaténés pour former des vecteurs augmentés $\mathbf{x}_l(1, n \times m)$ ou $\mathbf{x}_l(1, m \times n)$. Enfin, ces vecteurs augmentés sont concaténés à leur tour pour former une matrice augmentée $\mathbf{D_3}(l, n \times m)$ ou $\mathbf{D_4}(l, m \times n)$ (Figure II.3). Cette configuration est plus appropriée aux méthodes d'analyse de données multivariées bilinéaires car la dimension l relative aux échantillons représente les lignes de la matrice augmentée.

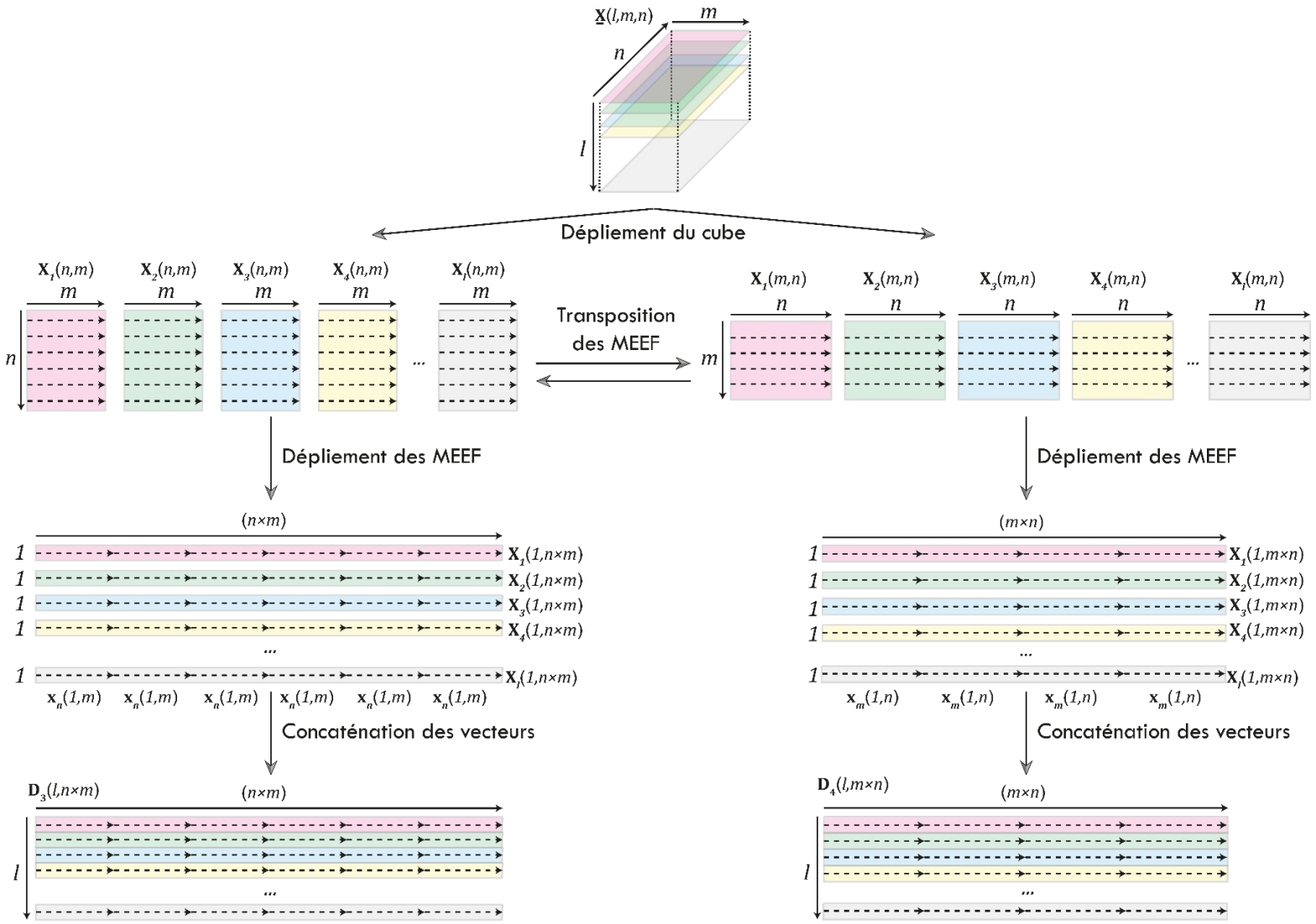


Figure II.3 : Dépliement du cube $\underline{\boldsymbol{X}}(l,m,n)$, puis dépliement des MEEF $\boldsymbol{X}_l(n,m)$ ou $\boldsymbol{X}_l(m,n)$ et concaténation des vecteurs $\boldsymbol{x}_n(1,m)$ ou $\boldsymbol{x}_m(1,n)$ en vecteurs augmentés $\boldsymbol{x}_l(1,n\times m)$ ou $\boldsymbol{x}_l(1,m\times n)$, et enfin, concaténation en colonnes de ces derniers pour obtenir une matrice augmentée bilinéaire.

Grâce à cette étape de préparation des données, il est désormais possible d'appliquer des méthodes d'analyse de données multivariées, à la fois trilinéaires et bilinéaires. Cependant, comme abordé dans le premier chapitre, divers phénomènes physico-chimiques peuvent altérer la structure, la position et l'intensité du signal de fluorescence lorsqu'une mesure de MEEF est effectuée (cf. §I.6). Par conséquent, il est important de prendre en compte ces facteurs, mais également de comprendre leur(s) impact(s) ou leur origine dans le but de les minimiser, voire même de les supprimer expérimentalement ou de les corriger en développant des méthodes de prétraitements en chimiométrie.

II.3 Correction des MEEF

II.3.1 Correction de l'effet du filtre interne

L'effet du filtre interne induit une extinction de fluorescence à cause d'une absorption de la lumière incidente sans émission de fluorescence ou d'une réabsorption de la fluorescence émise. Cette diminution de l'I_F engendre le non-respect de la loi de Beer-Lambert ce qui est souvent le cas d'échantillons complexes tels que les extraits de sols pollués aux CAP (i.e. présence simultanée de plusieurs chromophores et fluorophores). Par conséquent, une perturbation du caractère trilinéaire du cube $\underline{\mathbf{X}}(l,m,n)$ est systématiquement observée lorsqu'on est face à plusieurs échantillons de différentes origines. Pour corriger ou, du moins, atténuer cet effet lors de la mesure d'une MEEF, il convient de mesurer avec elle le spectre d'absorbance de l'échantillon et d'y appliquer ensuite un facteur correctif.

Dans le cadre de cette thèse, le spectrofluorimètre 3D (Aqualog ®) utilisé pour mener les expérimentations en fluorescence permet d'une part la mesure du spectre d'absorbance, d'autre part, il est fourni avec une méthode numérique permettant la correction de l'effet du filtre interne. Cet algorithme tient compte des différents paramètres de la géométrie optique fixe de l'appareil (i.e. non modifiable par l'utilisateur). Il s'agit de multiplier chaque intensité de fluorescence, mesurée à une paire de longueurs d'onde $(\lambda_{ex},\lambda_{em})$ notée $I_F^{mesurée}(\lambda_{ex},\lambda_{em})$, par un facteur correctif tenant compte des absorbances mesurées $A(\lambda_{ex})$ et $A(\lambda_{em})$ de l'échantillon, suivant la relation :

$$I_F^{Corrigée}(\lambda_{ex},\lambda_{em}) = I_F^{mesurée}(\lambda_{ex},\lambda_{em})10^{\frac{A(\lambda_{ex})+A(\lambda_{em})}{2}} \quad \text{(II.4)}$$

II.3.2 Correction des effets de la diffusion de la lumière

Les signaux de fluorescence 3D véhiculent de l'information chimique d'intérêt. Malheureusement, la structure des MEEF est fortement impactée par les effets de la diffusion de la lumière (i.e. Diffusion Rayleigh et diffusion Raman). Ces derniers engendrent une

perturbation sévère de la nature trilinéaire du cube $\underline{\mathbf{X}}(l,m,n)$ aboutissant nécessairement à des difficultés dans la décomposition linéaire du signal de fluorescence par MCR-ALS ou PARAFAC afin d'obtenir la caractérisation chimique la moins biaisée. Le risque associé est donc ne pas réussir à séparer les signaux sources correctement, soit en surestimant ou en sous-estimant le nombre d'espèces chimiques conduisant ainsi à extraire des composantes mathématiques qui ne reflètent pas une réalité chimique. Ce point est d'une importance capitale dans l'analyse chimiométrique des MEEF car, comme mentionné précédemment, la majorité des méthodes de décomposition en chimiométrie sont basées sur l'approximation linéaire de la loi de Beer-Lambert. Elles ne tiennent, en effet, pas compte des non linéarités dans les signaux associés à une MEEF, comme les signaux provenant de phénomènes de diffusion. Par conséquent, ces signaux doivent être systématiquement corrigés ou, du moins, atténués avant toute application d'une méthode de décomposition bilinéaire ou trilinéaire. C'est pourquoi, dans la littérature, différentes approches sont proposées pour éliminer ou gérer les effets de la diffusion de la lumière dans les MEEF : soustraction du blanc, recadrage du signal, insertion de valeurs manquantes ou de valeurs nulles, pondération des signaux de diffusion et enfin, modélisation du signal de fluorescence par interpolation dans les zones de diffusion.

En premier lieu, la soustraction du signal du blanc (i.e. signal du solvant seul, Figure II.4b) s'avère efficace pour éliminer le signal Raman, mais elle ne permet pas de supprimer entièrement le signal Rayleigh. En outre, elle peut générer des signaux négatifs[68]. En second lieu, le recadrage du signal vers une zone d'intérêt contenant uniquement le signal de fluorescence s'avère délicat puisqu'il peut générer une perte d'informations chimiques d'intérêt, en particulier dans les zones proches des effets de diffusion de la lumière[69]. Troisièmement, l'insertion de valeurs manquantes ou de valeurs nulles à la place des valeurs relatives aux signaux de diffusion (Figure II.4c et II.4d)[70,71] peut entraîner une perte d'informations chimiques et/ou perturber la nature des données[72], et même empêcher l'exécution de certains algorithmes qui sont sensibles aux valeurs manquantes[73]. Enfin, d'autres approches ont été proposées pour gérer les effets de diffusion, comme la modélisation des points de données de fluorescence où les effets de diffusion sont observés sur la MEEF et leur remplacement par des valeurs interpolées corrigées (Figure II.4e)[73,74]. Ces méthodes sont relativement efficaces pour gérer les effets de diffusion. Cependant, elles restent sensibles aux paramètres d'interpolation, tels que son type unidimensionnel ou bidimensionnel et sa taille de la fenêtre, afin d'obtenir une approximation optimale qui ne génère pas d'artefacts ou de biais lors de l'ajustement spectral, avec par exemple une méthode de décomposition type PARAFAC. De plus, l'ensemble de ces méthodes ne traitent pas le bruit blanc (i.e. bruit provenant de l'instrumentation), ce qui peut être un problème dans les cas où le rapport signal sur bruit est faible.

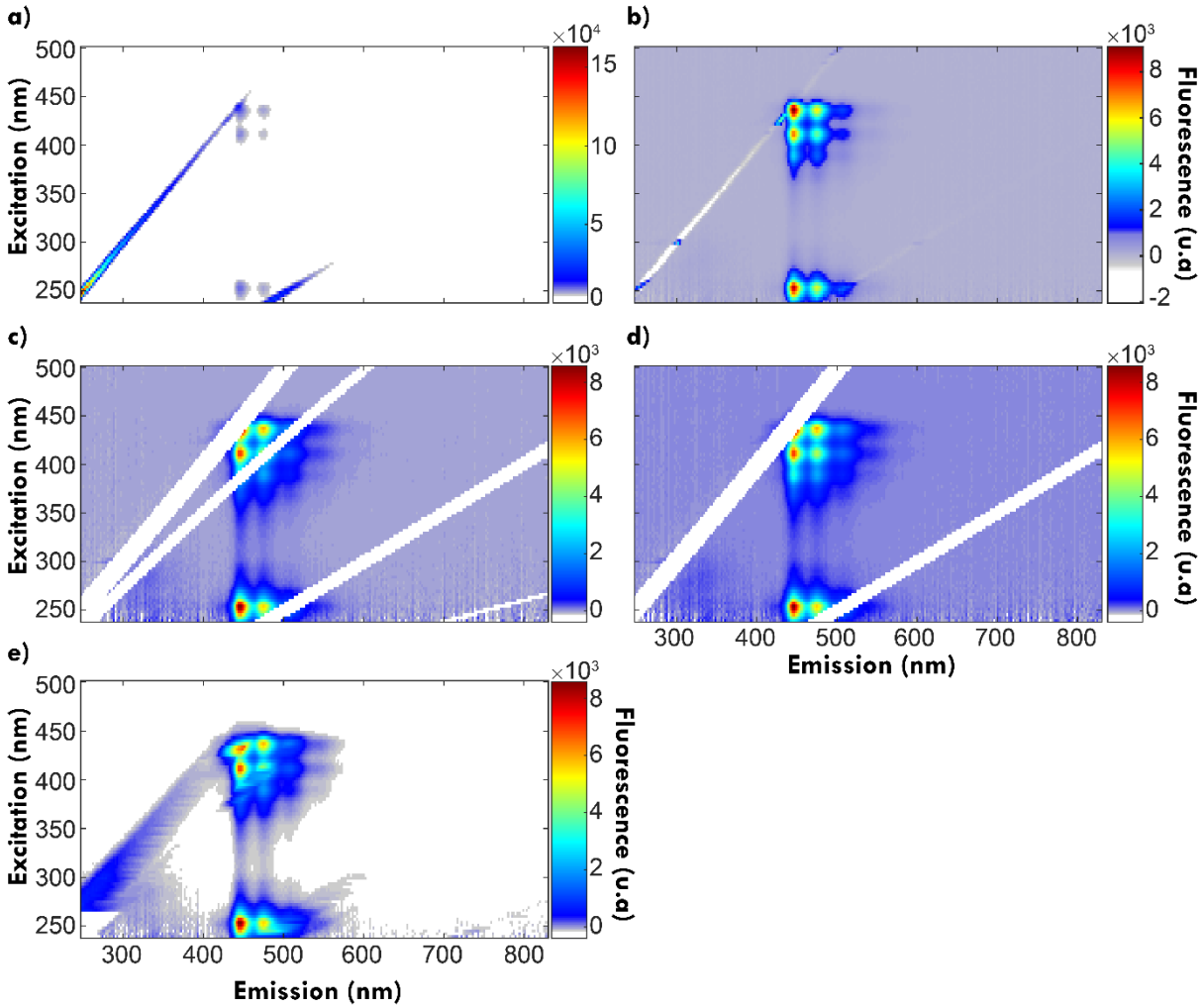


Figure II.4 : Différentes stratégies de correction des effets de diffusion de la lumière. MEEF brute du perylène dans du dichlorométhane (a), MEEF après soustraction du blanc (b), MEEF après insertion de valeurs nulles ou manquante (bandes blanches) à la place du signal Rayleigh de premier ordre et de second ordre ainsi que du signal Raman (c), même cas de figure que (c) à la différence que les valeurs nulles ne sont plus insérées sur la position du signal Raman (d), modélisation des points de données de fluorescence où les effets de diffusion sont observés sur l'EEM (e).

II.3.3 Normalisation des signaux

Une fois les MEEF corrigées de l'effet du filtre interne et des effets de diffusion non linéaires, il convient en chimiométrie d'effectuer au préalable une normalisation des signaux d'un cube de MEEF. Ce prétraitement a pour but de favoriser les variations chimiques d'un échantillon à un autre plutôt que l'amplitude des signaux totaux. En effet, la variabilité des données peut être sévèrement biaisée par l'effet de « levier » en amplitude causé, par exemple, par la réponse prédominante d'un fluorophore à haut rendement de fluorescence ou présentant une concentration élevée[59].

Il existe une multitude de méthodes de normalisation. Il s'agit pour les plus basiques de diviser l'ensemble des spectres par un coefficient ou une métrique sensible à l'effet de levier[75]. Cette opération engendre donc l'annulation de cet effet. Dans le cas des MEEF,

chaque i spectre d'émission et chaque i spectre d'excitation contenant J élements (i.e. vecteurs notés $\mathbf{v}_{i,j}$) est divisé par sa p-norme (i.e. $||\mathbf{v}_i||_p$) qui prend la forme générale :

$$||\mathbf{v}_i||_p = \left[\sum_{j=1}^{J} |\mathbf{v}_{i,j}|^p\right]^{1/p} \quad \text{(II.5)}$$

Avec $p \in \mathbb{Z}$ ou $p = \pm\infty$, i représente le numéro du vecteur (i.e. spectre) et j représente le numéro de la variable (i.e. numéro de la λ_{em} ou de la λ_{ex}).

Ainsi, l'équation de normalisation est[76] :

$$\mathbf{v}_i^{normalisé} = \mathbf{v}_i \cdot ||\mathbf{v}_i||_p^{-1} \quad \text{(II.6)}$$

La valeur de p détermine le type de normalisation. Les trois normalisations les plus courantes sont présentées et décrites dans le Tableau II.1.

Tableau II.1 : Trois normalisations courantes en chimiométrie[76].

Nom	**Description**	**Equation de la p-norme**	**Numéro de l'équation**
Norme L1	$p = 1$. Division de chaque élément j du vecteur $\mathbf{v}_i$ par la somme de la valeur absolue de toutes les variables. Cette normalisation aboutit à des vecteurs normalisés ($\mathbf{v}_i^{normalisé}$) ayant tous une aire sous la courbe égale à 1.	$\|\|\mathbf{v}_i\|\|_1 = \sum_{j=1}^{J} \|\mathbf{v}_{i,j}\|$	(II.7)
Norme L2	$p = 2$. Cette norme est appelée longueur euclidienne du vecteur ou encore amplitude du vecteur. Cette normalisation aboutit à des vecteurs normalisés avec une longueur égale à 1. Cette normalisation pondère significativement les valeurs les plus élevées.	$\|\|\mathbf{v}_i\|\|_2 = \sqrt[2]{\sum_{j=1}^{J} \|\mathbf{v}_{i,j}\|^2}$	(II.8)
Norme ∞	$p = \infty$. Normalisation à la valeur maximale observée pour toutes les variables du vecteur. Elle aboutit à des vecteurs normalisés avec une valeur maximale unitaire.	$\|\|\mathbf{v}_i\|\|_\infty = max_i(\|\mathbf{v}_i\|)$	(II.9)

II.4 Analyse des MEEF

Une fois que les MEEF sont prétraitées, il est possible d'appliquer des méthodes de chimiométrie exploratoire ou de résolution multivariée de courbes. Les méthodes exploratoires (i.e. réduction de dimensionalité) sont des techniques utilisées pour réduire le nombre de variables ou de dimensions d'un jeu de données tout en conservant les informations les plus pertinentes. Ces méthodes sont généralement utilisées lorsque le nombre initial de variables est élevé, ce qui est souvent le cas pour des données de spectroscopie. Parmi ces méthodes, l'analyse en composantes principales est considérée comme un outil exploratoire polyvalent et performant.

II.4.1 Analyse exploratoire des MEEF : réduction de dimensionalité des MEEF

II.4.1.1 Analyse en composantes principales

L'ACP est une technique d'exploration des données multivariées continues ou discrètes. Elle a été développée dès 1901 par Pearson[77] et reformulée par Hotelling en 1933[78]. Cette technique est très largement utilisée en chimiométrie, en particulier pour l'analyse des données spectroscopiques qui sont souvent caractérisées par un nombre important de spectres (i.e. échantillon) et de variables spectrales (e.g. longueurs d'ondes). L'ACP permet en effet, de réduire la dimensionalité des données et de visualiser graphiquement les résultats dans un espace réduit afin de synthétiser l'information contenue dans un jeu de données[79].

Pour illustrer le principe de l'ACP, prenons le cas de notre matrice de données $\mathbf{D}(l,m)$ contenant l spectres d'émission, chacun caractérisé par m λ_{em} (Figure II.1). L'ACP décompose cette matrice (Figure II.5) selon l'équation (II.10) par la construction de variables latentes appelées composantes principales (CP), qui sont des combinaisons linéaires des variables de départ (i.e. λ_{em}), tout en minimisant la somme des carrées des résidus contenus dans $\mathbf{E}$:

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \text{(II.10)}$$

Où $\mathbf{T}(l,r)$, $\mathbf{P}(m,r)$ et $\mathbf{E}(l,m)$ représentent la matrice des '*scores*' (i.e. coordonnées factorielles), la matrice des vecteurs propres ('*loadings*') et la matrice de la variance résiduelle dans chaque échantillon, respectivement. La dimension r représente le nombre de CP retenues. Idéalement, elle correspond au rang de la matrice $\mathbf{D}$ (i.e. $\mathrm{rang}(\mathbf{D}) = r$ où $r \in \mathbb{N}$). Ce rang représente la dimension du sous-espace vectoriel engendré par les vecteurs-ligne ou les vecteurs-colonne de la matrice $\mathbf{D}$. Son maximum est égal à la plus petite dimension de la matrice, mais il est généralement beaucoup plus faible en pratique, car seulement un nombre

réduit de vecteurs (i.e. CP) est nécessaire pour générer tout l'espace vectoriel. C'est là tout l'intérêt des méthodes de réduction de dimensionnalité.

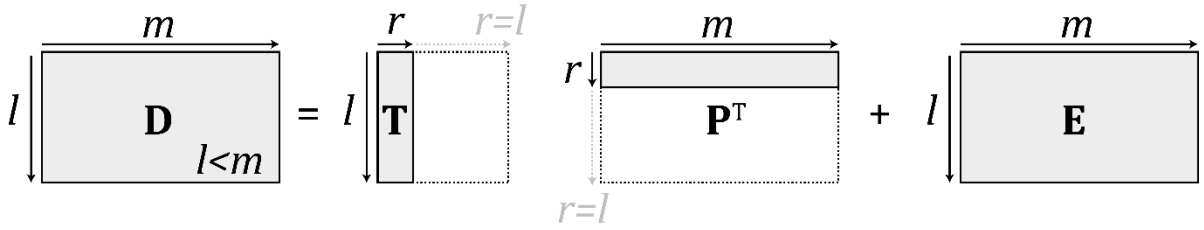


Figure II.5 : Décomposition réalisée en analyse en composantes principales de la matrice $\mathbf{D}$ avec $l < m$ avec $max(r) \leq l$ si $l < m$ et $max(r) \leq m$ si $l > m$.

La matrice des '*scores*' représente les coordonnées des projections des échantillons dans le nouvel espace créé par les CP. Ensuite, la matrice des vecteurs propres (i.e. '*loadings*') est composée des vecteurs CP qui sont des combinaisons linéaires des variables originales. Enfin, la matrice des résidus contient les informations non expliquées par le modèle ACP. Cette matrice devrait normalement être constituée uniquement de bruit lorsque toutes les informations pertinentes ont été capturées par le modèle. De plus, les CP sont orthogonales entre elles (i.e. elles ne partagent pas d'information entre elles) et elles sont classées par ordre décroissant de variance[76].

Mathématiquement, l'algorithme SVD '*Singular Value Decomposition*'[80] et l'algorithme NIPALS '*Non-linear Iterative Partial Least Squares*'[81] sont les plus utilisés pour le calcul du modèle ACP. En pratique, l'algorithme SVD consiste en une décomposition en valeurs et vecteurs singuliers de la matrice $\mathbf{D}_{centrée}$ (i.e. matrice $\mathbf{D}$ centrée sur la moyenne des l spectres qui la composent (équation II.11)). L'algorithme NIPALS quant à lui n'est pas directement appliqué à la matrice des données $\mathbf{D}_{centrée}$, mais plutôt à sa matrice de variance-covariance $\mathbf{C}(m,m)$, qui est calculée par l'équation (II.12)[76] :

$$\mathbf{D}_{centrée} = \mathbf{D} - \text{moy}(\mathbf{D}) \quad \text{(II.11)}$$

$$\mathbf{C} = \text{cov}(\mathbf{D}_{centrée}) = \frac{\mathbf{D}_{centrée}{}^{\mathrm{T}}\mathbf{D}_{centrée}}{l-1} \quad \text{(II.12)}$$

L'algorithme NIPALS consiste en une décomposition en valeurs et vecteurs propres de la matrice de variance-covariance $\mathbf{C}$, qui peut être obtenue en résolvant l'équation[82]:

$$\mathbf{CP} = \mathbf{P}\boldsymbol{\Lambda} \quad \text{(II.13)}$$

Où $\mathbf{P}(m,r)$ est la matrice des vecteurs propres, triés par ordre décroissant de leurs valeurs propres associées, et $\boldsymbol{\Lambda}(r,r)$ est la matrice diagonale des valeurs propres correspondantes. Les vecteurs propres $\mathbf{p}$ contenus dans la matrice $\mathbf{P}$ doivent être orthonormés, ce qui signifie qu'ils doivent satisfaire à la condition que leur norme soit égale à

1 (i.e. $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$ avec $\mathbf{I}(r, r)$ est la matrice identité de dimension r). Dès lors que l'équation (II.13) est résolue, il suffit ensuite de projeter les données centrées $\mathbf{D}_{centrée}$ sur les vecteurs propres $\mathbf{p}$ pour obtenir la matrice des '*scores*' $\mathbf{T}(l, \mathrm{r})$:

$$\mathbf{T} = \mathbf{D}_{centrée}\mathbf{P} \qquad \text{(II.14)}$$

En ce qui concerne les données de fluorescence 3D, il est inapproprié d'appliquer l'ACP sur une seule MEEF car elle contient deux dimensions spectrales et aucune dimension relative aux échantillons, étant donné qu'il s'agit d'un seul échantillon. Par ailleurs, lorsqu'il y a plusieurs MEEF qui, rappelons-le, forment un cube de données, deux options sont disponibles. Soit une ACP est appliquée sur la matrice augmentée du cube de données (i.e. matrices $\mathbf{D}_3$ ou $\mathbf{D}_4$ de la Figure II.3), soit une analyse en composantes principales multivoie (ACPM) mieux adaptée à ce type de données multivoies est utilisée.

II.4.1.2 Analyse en composantes principales multivoie

L'ACPM étend l'analyse en composantes principales aux données multivoies tridimensionnelles et présente les mêmes objectifs et avantages[83]. Pour illustrer son principe, prenons le cas de notre cube $\underline{\mathbf{X}}(l, m, n)$ contenant l MEEF, chacune caractérisée par n λ_{ex} et m λ_{em}. L'ACPM décompose ce cube de données (Figure II.6) en une somme de produits tensoriels externes notés avec le symbole ⊗, entre les vecteurs des '*scores*' $\mathbf{t}_r$ et les matrices des vecteurs propres '*loadings*' $\mathbf{P}_r$. Le produit externe de deux tenseurs est une multiplication de toutes les combinaisons de leurs éléments. Étant donné qu'aucune dimension des tenseurs n'est contractée, le résultat est un tenseur de plus grande dimension :

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{t}_r \otimes \mathbf{P}_r + \underline{\mathbf{E}} \qquad \text{(II.15)}$$

Où $l < \{n, m\}$ et $\underline{\mathbf{E}}$ représente le cube de la variance résiduelle dans chaque échantillon. L'indice r représente quant à lui le nombre de CP.

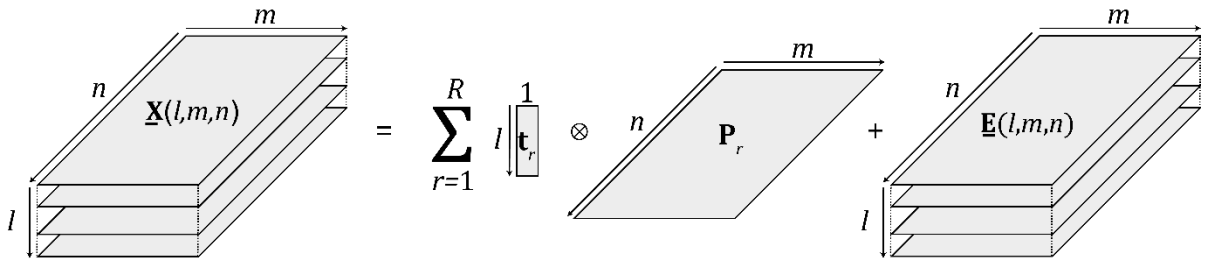


Figure II.6 : Décomposition par analyse en composantes principales multivoie (ACPM) d'un cube de MEEF.

II.4.2 Résolution multivariée de courbes : extraction des signaux sources à partir des MEEF par décomposition spectrale trilinéaire

Les méthodes d'exploration des données sont des outils précieux pour étudier les caractéristiques et les tendances des données. Elles sont généralement les premières à être utilisées pour le traitement des données. À l'inverse, les méthodes de résolution multivariée de courbes ont pour objectif spécifique d'extraire et de séparer les informations relatives aux différentes espèces chimiques qui composent un mélange complexe. Pour analyser les données de fluorescence 3D (i.e. MEEF), la décomposition trilinéaire PARAFAC est l'une des méthodes les plus appropriées.

II.4.2.1 Parallel Factor Analysis, PARAFAC

PARAFAC est une méthode de décomposition trilinéaire proposée indépendamment par Harshman[84] ainsi que par Carroll et Chang[85]. Elle décompose un cube de données en composantes pures trilinéaires correspondant aux trois dimensions (i.e. modes) de ce dernier. Elle implique que chaque composante pure soit composée d'un vecteur de '*scores*' et de deux vecteurs de '*loadings*', plutôt qu'un vecteur de '*scores*' et un seul vecteur de '*loadings*' comme dans le cas de l'ACP (Figure II.7)[63].

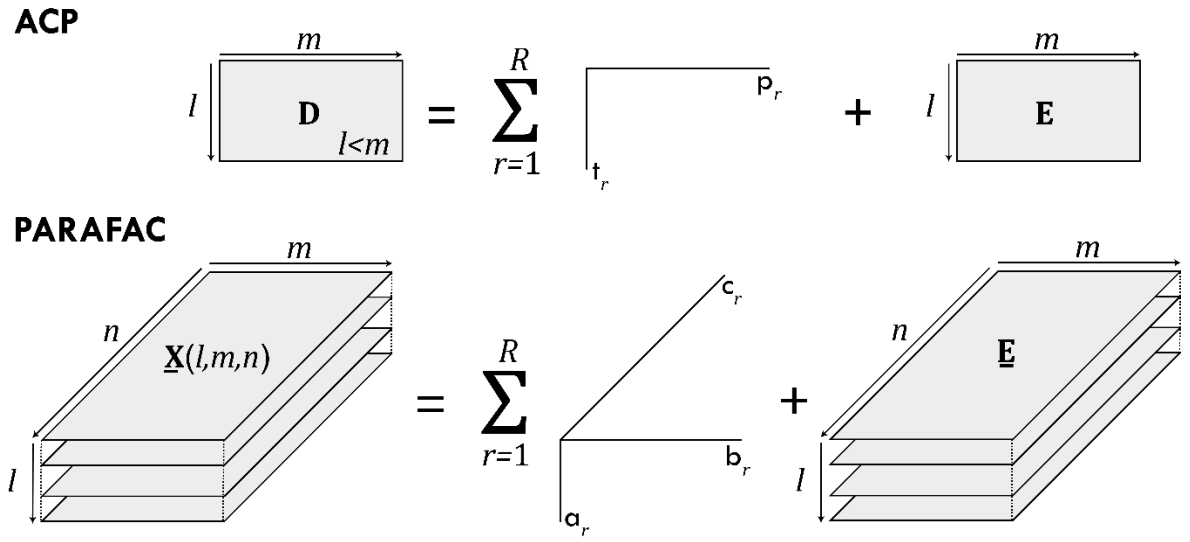


Figure II.7 : Schématisation tensorielle de la décomposition PARAFAC versus la décomposition ACP. t et p correspondent aux deux dimensions des composantes principales de l'ACP. a,b et c correspondent aux trois modes des composantes pures de PARAFAC. r correspond à la dimension du modèle.

Pour illustrer le principe de cette méthode, prenons comme pour l'ACPM, le cas de notre cube $\underline{\mathbf{X}}(l,m,n)$ contenant l MEEF, chacune caractérisée par n longueurs d'excitation λ_{ex} et m longueurs d'émissions λ_{em}. PARAFAC le décompose en une matrice des '*scores*' $\mathbf{A}(l,r)$ et deux matrices de '*loadings*' $\mathbf{B}(m,r)$ et $\mathbf{C}(n,r)$, tout en minimisant la somme des carrés des résidus présents dans le cube $\underline{\mathbf{E}}(l,m,n)$, selon le modèle[86]:

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} + \mathbf{E} \qquad \text{(II.16)}$$

Où $\mathbf{X}(l, m \times n)$ est la matrice réorganisée du cube $\underline{\mathbf{X}}$ et $\mathbf{E}(l, m \times n)$ est la matrice réorganisée du cube des résidus $\underline{\mathbf{E}}$. L'opérateur ⊙ correspond au produit de Khatri-Rao qui est un simple produit matriciel par partition en colonnes des matrices $\mathbf{B}$ et $\mathbf{C}$[87]. La dimension r représente la dimension du modèle (i.e. le nombre de composantes choisies, appelées composantes pures). Chaque colonne de la matrice des '*scores*' $\mathbf{A}$ contient la contribution de chaque composante pure dans les l échantillons. Chaque colonne de la matrice $\mathbf{B}$ contient le profil d'émission (i.e. λ_{em}) de chaque composante pure, tandis que chaque colonne de la matrice $\mathbf{C}$ renferme le profil d'excitation (i.e. λ_{ex}) de chaque composante pure[59] (Figure II.8).

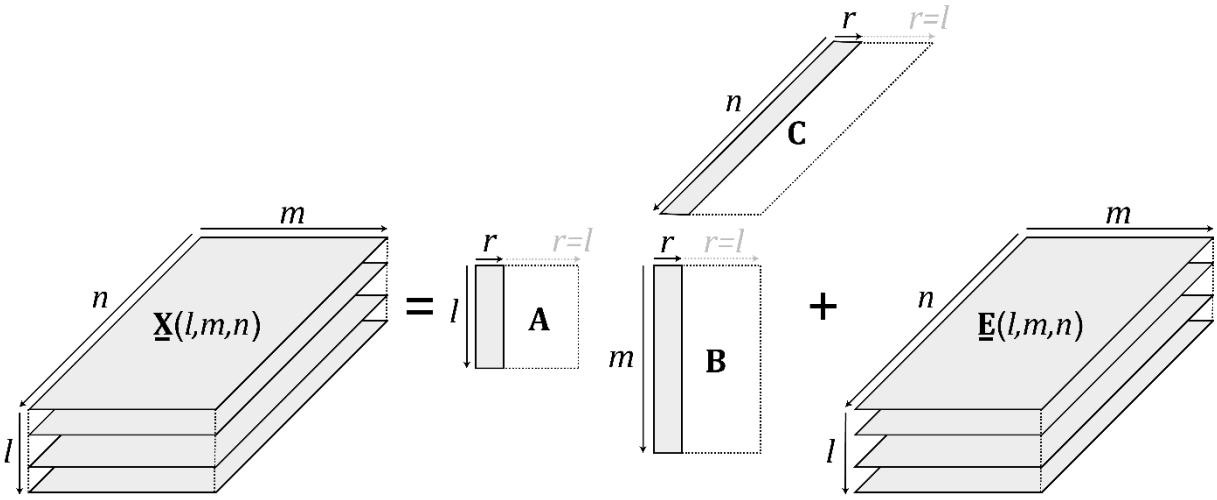


Figure II.8 : Décomposition PARAFAC d'un cube de MEEF.

La méthode PARAFAC peut, d'une certaine manière, être considérée comme une extension de l'ACP aux données multivoies, tout comme l'ACPM. Toutefois, la principale différence entre ces méthodes est que le modèle PARAFAC est exempté de la condition d'orthogonalité pour le calcul des composantes principales, appelées composantes pures dans ce cas[88]. Cette levée de la condition d'orthogonalité améliore l'interprétation chimique des '*loadings*' issus du modèle. En effet, les profils des composantes pures correspondent davantage aux résultats expérimentaux (i.e. spectres) que les profils issus des composantes principales, qui sont des combinaisons linéaires contraintes des variables de départ.

II.4.2.2 Ambiguïté rotationnelle des modèles bilinéaires et unicité du modèle PARAFAC

La levée de cette contrainte d'orthogonalité entraîne toutefois un problème concernant les ambiguïtés rotationnelles des solutions. Ces ambiguïtés peuvent avoir un impact dramatique, en particulier dans le cas des méthodes de décomposition bilinéaires, sur l'interprétation des résultats, à moins de prendre des précautions préalables. Ces ambiguïtés implique l'existence pour une seule réalité chimique de plusieurs solutions mathématiques

différentes (i.e. matrices des '*scores*' et des '*loadings*') pour un modèle ayant le même pourcentage de variance expliquée et les mêmes résidus, à partir des mêmes données de départ[67]. Pour illustrer ce concept d'ambiguïté rotationnelle, prenons le cas de notre matrice $\mathbf{D}(l,m)$ que nous décidons de décomposer bilinéairement en deux matrices $\mathbf{C}$ et $\mathbf{S}$ suivant l'équation (II.2). Il existe ainsi une multitude de matrices non-singulières (i.e. inversibles) notées $\mathbf{R}$ capables de faire subir une rotation aux matrices $\mathbf{C}$ et $\mathbf{S}$:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} = \mathbf{C}\left(\mathbf{R}\mathbf{R}^{-1}\right)\mathbf{S}^{\mathrm{T}} + \mathbf{E} = \mathbf{C}^{*}\mathbf{S}^{*\mathrm{T}} + \mathbf{E} \qquad \text{(II.17)}$$

Où $\mathbf{C}^{*} = \mathbf{C}\mathbf{R}$ et $\mathbf{S}^{*\mathrm{T}} = \mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}$**.** $\mathbf{C}$ avec $\mathbf{C}^{*}$ et $\mathbf{S}^{\mathrm{T}}$ avec $\mathbf{S}^{*\mathrm{T}}$ sont des solutions valables mathématiquement ne différant que par rotation mais ne correspondant, malheureusement, pas toutes à une réalité chimique[67,89]. Pour réduire le risque et les problèmes d'ambiguïté rotationnelle, des contraintes mathématiques peuvent être appliquées pour trier les solutions. Par exemple, en fluorescence 3D, une contrainte de non-négativité sur les '*scores*' et les '*loadings*' peut être introduite en raison de la positivité intrinsèque des concentrations et des MEEF, respectivement[88]. Dans le cas de PARAFAC, cette limitation est atténuée par l'unicité des solutions du modèle trilinéaire sous la condition suffisante, mais non nécessaire, énoncée par Kruskal[90] et dont le formalisme mathématique est reformulé de manière plus intuitive par Stegeman et Sidiropoulos[91] :

$$2r + 2 \leq \mathrm{rang}_k(\mathbf{A}) + \mathrm{rang}_k(\mathbf{B}) + \mathrm{rang}_k(\mathbf{C})$$

Où r représente le nombre de composantes pures choisies. rang_k désigne le rang Kruskal d'une matrice (i.e. nombre maximum de colonnes d'une matrice tel que toute sous-matrice de cette dernière ayant le même nombre de colonnes est de rang plein[92]). Ce dernier est toujours inférieur ou égale au rang matriciel commun (e.g. $\mathrm{rang}_k(\mathbf{A}) \leq \mathrm{rang}(\mathbf{A})$). Dans le cas où $r = 1$, la condition ne peut être vérifiée. Néanmoins, Harshman a prouvé l'unicité du modèle PARAFAC dans ce cas[93]. Dans le cas où $r = 2$ ou $r = 3$, la vérification de la condition est nécessaire pour prouver l'unicité du modèle[91].

Par conséquent, dans le cas de PARAFAC, si les données sont effectivement trilinéaires, les composantes pures reflétant les résultats expérimentaux (i.e. spectres) seront trouvées si le rapport signal/bruit est approprié et si le bon nombre de composantes pures est estimée[63]. En effet, cette précision est nécessaire car les modèles PARAFAC, contrairement aux modèles ACP, ne s'emboitent pas (e.g. pour les mêmes données, le modèle avec deux composantes pures n'englobe pas nécessairement le modèle avec une seule composante pure). Ainsi, cette précision est nécessaire pour fournir un modèle PARAFAC valide et optimal avec une réalité chimique la moins biaisée possible. Malheureusement, il n'y a pas de règle générale pour cela mais, en pratique, ce choix peut être basé sur les connaissances a priori

du système étudié et sur différents critères complémentaires, implémentés à l'algorithme PARAFAC, tels que l'indicateur CORCONDIA pour '*CORE CONsistency DIAgnostics*'[86], la méthode de validation croisée '*split half analysis*'[94] et le pourcentage de variance expliquée par la dernière composante de chaque modèle ou encore la vérification des résidus[59].

II.4.2.3 Algorithme de PARAFAC

Il existe plusieurs méthodes numériques pour calculer un modèle PARAFAC. L'approche la plus couramment utilisée, en raison de sa précision, est basée sur la méthode itérative des moindres carrés alternés '*Alternating Least Squares*' (ALS). Cet algorithme estime la matrice inconnue d'un mode (i.e. une dimension) en supposant que les matrices des deux autres modes sont connues[63,92]. Les étapes de l'algorithme PARAFAC sont :

1 - Fixation de la dimension du modèle r (i.e. nombre de composantes pures). En pratique, ce choix est généralement basé sur les connaissances a priori du système chimique ou sur une estimation obtenue à l'aide de méthodes de réduction de dimensionnalité telles que la SVD, l'ACP ou l'ACPM. Cette estimation peut également être validée en utilisant des indicateurs tels que CORCONDIA ou le pourcentage de variance expliquée par les composantes du modèle.

2 - Initialisation des matrices $\mathbf{B}$ et $\mathbf{C}$ de PARAFAC au moyen d'un algorithme d'initialisation tel que la décomposition trilinéaire directe '*TriLinear Decomposition*' (TLD) qui décompose le cube de données $\underline{\mathbf{X}}$ en la somme du produit tensoriel de trois vecteurs $\mathbf{a}_r$, $\mathbf{b}_r$ et $\mathbf{c}_r$:

$$\underline{\mathbf{X}} = \sum_{\mathrm{r}=1}^{R} \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r + \underline{\mathbf{E}} \qquad \text{(II.18)}$$

Où $\mathbf{a}_r$, $\mathbf{b}_r$ et $\mathbf{c}_r$ sont les $r^{\text{ème}}$ colonnes des matrices initiales de PARAFAC, $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$, respectivement. L'algorithme TLD est fourni, mais aussi détaillé, dans les travaux de E.Sanchez et B.Kowalski[95].

3 - Estimation de la matrice $\mathbf{A}(l,r)$, à partir des matrices $\mathbf{B}(m,r)$, $\mathbf{C}(n,r)$ et $\mathbf{X}(l,m\times n)$ grâce à la résolution du modèle des moindres carrés :

$$\mathbf{X} = \mathbf{AZ} \qquad \text{(II.19)}$$

Où $\mathbf{Z}(r,m\times n)$ est la matrice contenant les vecteurs $\mathbf{z_r}(1,m\times n)$ résultant de l'opération $\mathbf{z_r} = (\mathbf{b_r}\otimes\mathbf{c_r})$.

Ainsi, l'estimation de $\mathbf{A}(l,r)$ (i.e. solution du modèle (II.19)) est obtenue par :

$$\mathbf{A} = \mathbf{XZ}^{\mathrm{T}}(\mathbf{ZZ}^{\mathbf{T}})^{-\mathbf{1}} \qquad \text{(II.20)}$$

4 - Estimation de la matrice $\mathbf{B}$ par résolution du modèle des moindres carrés en utilisant les matrices $\mathbf{A}$ et $\mathbf{C}$ (i.e. calcul identique à celui utilisé lors de l'étape 3 pour le calcul de la matrice $\mathbf{A}$).

5 - Estimation de la matrice $\mathbf{C}$ par résolution du modèle des moindres carrés en utilisant la matrice $\mathbf{A}$ et la nouvelle matrice $\mathbf{B}$ (i.e. calcul identique à celui utilisé lors des étapes 3 et 4 pour le calcul de la matrice $\mathbf{A}$ et de la matrice $\mathbf{B}$, respectivement).

6 - Retour à l'étape 3 et répétition des calculs des matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ pour minimiser le critère des carrés des résidus : $\| \mathbf{X} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} \|^2$, jusqu'à ce que le modèle converge et satisfasse le critère d'arrêt d'ajustement défini par l'utilisateur (i.e. écart d'ajustement très faible entre deux itérations, égal, par exemple à 10^{-6}).

En résumé, la modélisation PARAFAC revient à résoudre un problème inverse définit par l'équation (II.16) par la recherche d'une solution au sens des moindres carrés, en résolvant le modèle définit par l'équation (II.19). Lorsque ce problème inverse est bien posé et que le cube de données est bien conditionné, l'algorithme des moindres carrés alternés améliore l'ajustement du modèle à chaque itération jusqu'à ce qu'il atteigne le critère de convergence. La solution au sens des moindres carrés du modèle PARAFAC est ainsi obtenue. Elle est existante, unique et dépendante des données départ. Dans le cas où le problème inverse est mal posé (i.e. un mauvais dimensionnement du modèle), l'algorithme des moindres carrés alternés peut diverger ou converger vers un minimum local. La solution au sens des moindres carrés du modèle PARAFAC est ainsi soit inexistante soit non représentative de manière optimale de la réalité chimique.

II.5 Méthodes d'apprentissage supervisé : Prédiction d'informations quantitatives par régression linéaire et non linéaire des MEEF

De manière générale, une méthode d'apprentissage supervisé consiste à entraîner un modèle mathématique en utilisant des données d'entrée connues a priori (Figure II.9a). Les données d'entrée correspondent ainsi à une collection de signaux chimiques (e.g. spectres) assimilables à des empreintes digitales des échantillons de départ sur lesquels une connaissance a priori est utilisée. On parlera d'étiquetage des données. Ces étiquettes sont des valeurs qui reflètent une grandeur physico-chimique calculée ou mesurée, telle que la concentration d'un analyte.

Lors de l'apprentissage supervisé, le modèle cherche à établir une relation entre les caractéristiques des spectres et les étiquettes associées. Cette relation est généralement représentée par des paramètres du modèle, qui sont ajustés en minimisant une fonction d'erreur, encore appelée fonction de coût entre les valeurs prédites et ces étiquettes réelles.

Une fois le modèle entraîné, il peut être utilisé pour prédire ces mêmes étiquettes à de nouveaux spectres, permettant ainsi d'estimer des valeurs, initialement inconnues, de cette grandeur physico-chimique ciblée (Figure II.9b). Les modèles mathématiques les plus utilisés sont la régression linéaire (e.g. PLSR, N-PLS et SVR linéaire) et non linéaire (e.g. SVR non linéaire).

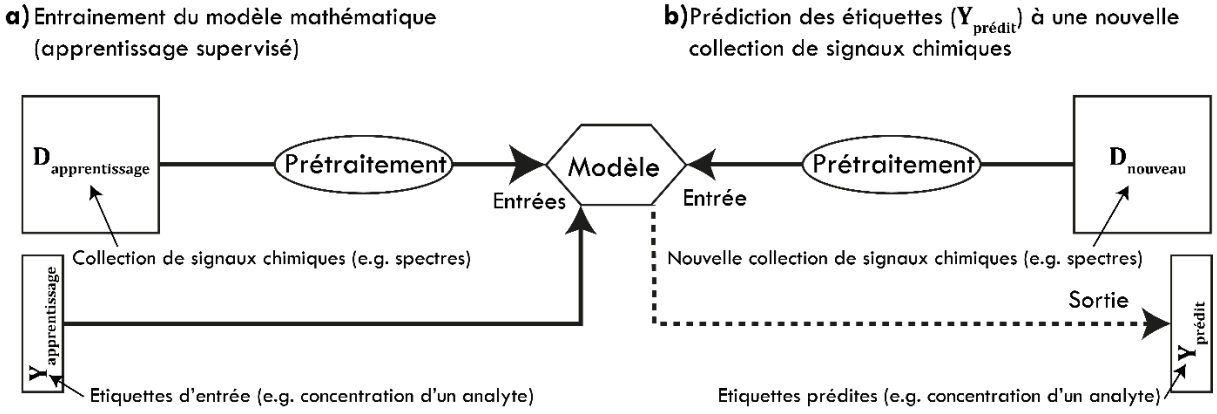


Figure II.9 : Schématisation du concept général d'une méthode d'apprentissage supervisé. Entrainement du modèle (a) et prédiction grâce au modèle (b).

II.5.1 Régression PLSR

Pour illustrer le principe de la PLSR, prenons le cas de notre matrice $\mathbf{D}(l,m)$ contenant l spectres d'émission, chacun caractérisé par m longueurs d'émissions λ_{em} (Figure II.1) et une autre matrice $\mathbf{Y}(l,q)$ contenant l mesures, chacune caractérisée par q variables quantitatives. De façon générale, les modèles de régressions linéaires multivariés (MLR) calculent des coefficients de régression stockés dans une matrice que nous notons $\mathbf{B}_D(m,q)$ par la relation:

$$\mathbf{B}_D = (\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{Y} \qquad \text{(II.21)}$$

Le problème est que cette équation est impossible à résoudre dans le cas où les variables m sont très corrélées entre elles et que $m > l$, ce qui est souvent le cas des données spectroscopiques. Ainsi, la régression sur composantes principales (PCR) fut développée pour surmonter ce problème en se basant sur l'ACP qui réduit la dimensionalité de la matrice $\mathbf{D}$ par la construction de la matrice des '*scores*' et celle des '*loadings*'[96] (cf. §II.4.1.1). La matrice des coefficients de régression $\mathbf{B}_T(r,q)$ est d'abord calculée sur la matrice des '*scores*' de l'ACP $\mathbf{T}(l,r)$ avant d'être projetée sur la matrice des '*loadings*' $\mathbf{P}(m,r)$ pour déduire la matrice des coefficients de régression de $\mathbf{D}$ notée $\mathbf{B}_D(m,q)$:

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \text{(II.22)}$$

$$\mathbf{B}_T = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y} \qquad \text{(II.23)}$$

$$\mathbf{B}_D = \mathbf{P}\mathbf{B}_T \qquad \text{(II.24)}$$

Bien que la PCR ait des avantages, tels que l'orthogonalité des CP et la possibilité de ne retenir qu'un nombre réduit de CP pour expliquer la variance de la matrice $\mathbf{D}$, elle présente également des limitations. Une seule CP peut contenir les contributions de plusieurs phénomènes physico-chimiques (i.e. les CP sont des combinaisons linéaires des variables de départ), ce qui peut rendre les modèles basés sur les variables de départ meilleurs que ceux basés sur les coordonnées factorielles[97]. De plus, la variabilité dans $\mathbf{D}$ permettant de décrire $\mathbf{Y}$ peut être faible (i.e. prédominance dans les spectres de phénomènes indépendants des mesures stockées dans $\mathbf{Y}$), ce qui peut nécessiter la sélection des CP à hauts rangs, une sélection qui n'est pas toujours évidente, car les données expérimentales sont souvent entachées de bruit. Par conséquent, la PLSR fût développée[98] et elle constitue un bon compromis entre la MLR et la PCR. En effet, elle vise à modéliser simultanément la variabilité dans les matrices $\mathbf{D}$ et $\mathbf{Y}$ en calculant des variables latentes (VL), différentes des CP de la PCR, qui maximisent la variance extraite des deux matrices tout en maximisant leur corrélation. Plus précisément, elle diffère de la PCR en utilisant l'information de la matrice $\mathbf{Y}$ pendant la décomposition de la matrice $\mathbf{D}$, plutôt que de décomposer uniquement $\mathbf{D}$ et ensuite de régresser les 'scores' (i.e. $\mathbf{T}$) sur $\mathbf{Y}$**.** Par conséquent, $\mathbf{D}$ et $\mathbf{Y}$ sont toutes les deux décomposées[97] :

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad \text{(II.25)}$$

$$\mathbf{Y} = \mathbf{Q}\mathbf{R}^{\mathrm{T}} + \mathbf{F} \quad \text{(II.26)}$$

Où $\mathbf{Q}(l,r)$ et $\mathbf{R}(q,r)$ représentent les matrices des '*scores*' et '*loadings*' de $\mathbf{Y}$, respectivement. $\mathbf{F}(l,q)$ est la matrice des résidus. La dimension r représente dans ce cas le nombre de VL choisies. Ensuite, une relation linéaire entre les '*scores*' de $\mathbf{D}$ et ceux de $\mathbf{Y}$ est calculée (i.e. modèle de régression). De plus, $\mathbf{Y}$ est projetée selon les '*scores*' de $\mathbf{D}$:

$$\mathbf{Q} = \mathbf{T}\mathbf{G} + \mathbf{H} \quad \text{(II.27)}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{R}^{\mathrm{T}} + \mathbf{F}^{*} \quad \text{(II.28)}$$

Où $\mathbf{G}(r,r)$ est une matrice diagonale. $\mathbf{H}(l,r)$ et $\mathbf{F}^{*}(l,q)$ sont les matrices des résidus. A partir de là, l'algorithme itératif NIPALS[99] est appliqué pour calculer l'ensemble des VL. Au final, une matrice des coefficients de régression $\mathbf{B}_{DY}(m,q)$ de la PLSR est calculée :

$$\mathbf{B}_{DY} = \mathbf{D}^{\mathrm{T}}\mathbf{Q}\,(\mathbf{T}^{\mathbf{T}}\mathbf{D}\mathbf{D}^{\mathrm{T}}\mathbf{Q})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y} \quad \text{(II.29)}$$

A cette étape, le modèle est construit. La prédiction est réalisée ensuite sur un nouveau jeu de données $\mathbf{D}_{nouveau}$ au moyen de l'équation[97] :

$$\mathbf{Y}_{estimé} = \mathbf{D}_{nouveau}\mathbf{B}_{\mathbf{DY}} \quad \text{(II.30)}$$

Où $\mathbf{Y}_{estimé}(l,q)$ est la matrice des variables quantitatives prédites.

II.5.2 Régression N-PLS

La régression N-PLS[65] est une extension de la PLSR aux données multivoies tridimensionnelles et présente les mêmes objectifs et avantages. Cette méthode présente des similitudes avec PARAFAC, néanmoins le modèle N-PLS est ajusté, de sorte à décrire la covariance des variables spectrales et quantitatives[65,76]. La différence majeure entre la PLSR et la N-PLS réside dans le fait que, pour les données tridimensionnelles, la décomposition est trilinéaire plutôt que bilinéaire. Pour illustrer ce point, prenons le cas de notre cube $\underline{\mathbf{X}}(l,m,n)$ contenant l MEEF, chacune caractérisée par n λ_{ex} et m λ_{em} et de notre matrice $\mathbf{Y}(l,q)$ contenant l mesures, chacune caractérisée par q variables quantitatives. N-PLS décompose $\underline{\mathbf{X}}$ suivant le modèle trilinéaire, qui d'ailleurs ressemble beaucoup à celui de PARAFAC :

$$\mathbf{X} = \mathbf{T}(\mathbf{P^m} \odot \mathbf{P^n})^{\mathrm{T}} + \mathbf{E} \qquad \text{(II.31)}$$

Avec $\mathbf{X}(l, m \times n)$ la matrice réorganisée du cube $\underline{\mathbf{X}}$, $\mathbf{T}(l,r)$ la matrice des '*scores*' et $\mathbf{P^m}(m,r)$ et $\mathbf{P^n}(n,r)$ les matrices des '*loadings*' suivant le mode m et le mode n, respectivement. $\mathbf{E}(l, m \times n)$ est la matrice réorganisée des résidus. La dimension r représente le nombre de variables latentes choisies. La matrice $\mathbf{Y}$ est décomposée suivant le modèle bilinéaire comme en PLSR (N.B. Les données quantitatives peuvent également se présenter sous une forme tridimensionnelle et être décomposé trilinéairement) :

$$\mathbf{Y} = \mathbf{QR}^{\mathrm{T}} + \mathbf{F} \qquad \text{(II.32)}$$

Ainsi, le modèle de régression N-PLS reliant les deux décompositions est[88] :

$$\mathbf{Q} = \mathbf{TG} + \mathbf{H} \qquad \text{(II.33)}$$

A partir de là, le raisonnement est le même que pour la PLSR.

Malheureusement, la relation entre les variables spectrales et les variables quantitatives n'est pas toujours linéaire de par la condition de la loi de Beer-Lambert non vérifiée. Il existe une multitude de facteurs contribuant à cette non linéarité. Certains d'entre eux, tels que la concentration élevée en analytes, peuvent être communs à plusieurs techniques analytiques. En revanche, d'autres facteurs peuvent être inhérent à une seule technique tels que les effets de la diffusion de la lumière et l'effet du filtre interne en fluorescence 3D. La PLSR est réputée pour sa capacité à gérer des non-linéarités faibles, en particulier lorsque le nombre de VL augmente. Cependant, cette approche n'est pas toujours suffisante et peut entraîner une baisse de l'ajustement du modèle en raison de l'utilisation de VL à haut niveau de bruit. Par conséquent, il est nécessaire dans ce cas de recourir à des

techniques de chimiométrie capables de gérer les non linéarités dans les données comme la régression SVR[100].

II.5.3 Régression SVR

La régression par vecteurs de support (SVR) est une méthode relativement récente et qui est une extension des techniques des machines à vecteurs de support (SVM)[101] pour résoudre des problèmes de régression. À l'origine, les SVM sont des méthodes d'apprentissage automatique basées sur la théorie de l'apprentissage statistique. Elles ont été essentiellement développées pour la classification des données. De façon générale, leur principe repose sur la recherche d'un hyperplan (i.e. une fonction mathématique généralisant le concept de plan en dimension supérieure à deux) qui sépare de façon optimale (i.e. Bonne performance et bonne généralisation du modèle) les données via la maximisation de la marge entre l'hyperplan (appelé aussi frontière de décision) et les observations[66]. Par exemple, lorsque nous choisissons un hyperplan capable de séparer deux classes mais que la marge n'est pas maximisée (Figure II.10a), le résultat est un modèle qui présente de bonnes performances sur le jeu de données d'entraînement, mais une mauvaise généralisation aux nouveaux échantillons qui ne sont pas inclus dans l'ensemble d'entraînement (Figure II.10a'). En revanche, en maximisant la marge (Figure II.10b), nous obtenons non seulement de bonnes performances sur le jeu d'entraînement, mais également une généralisation optimale aux nouveaux échantillons qui ne sont pas inclus dans l'ensemble d'entraînement (Figure II.10b').

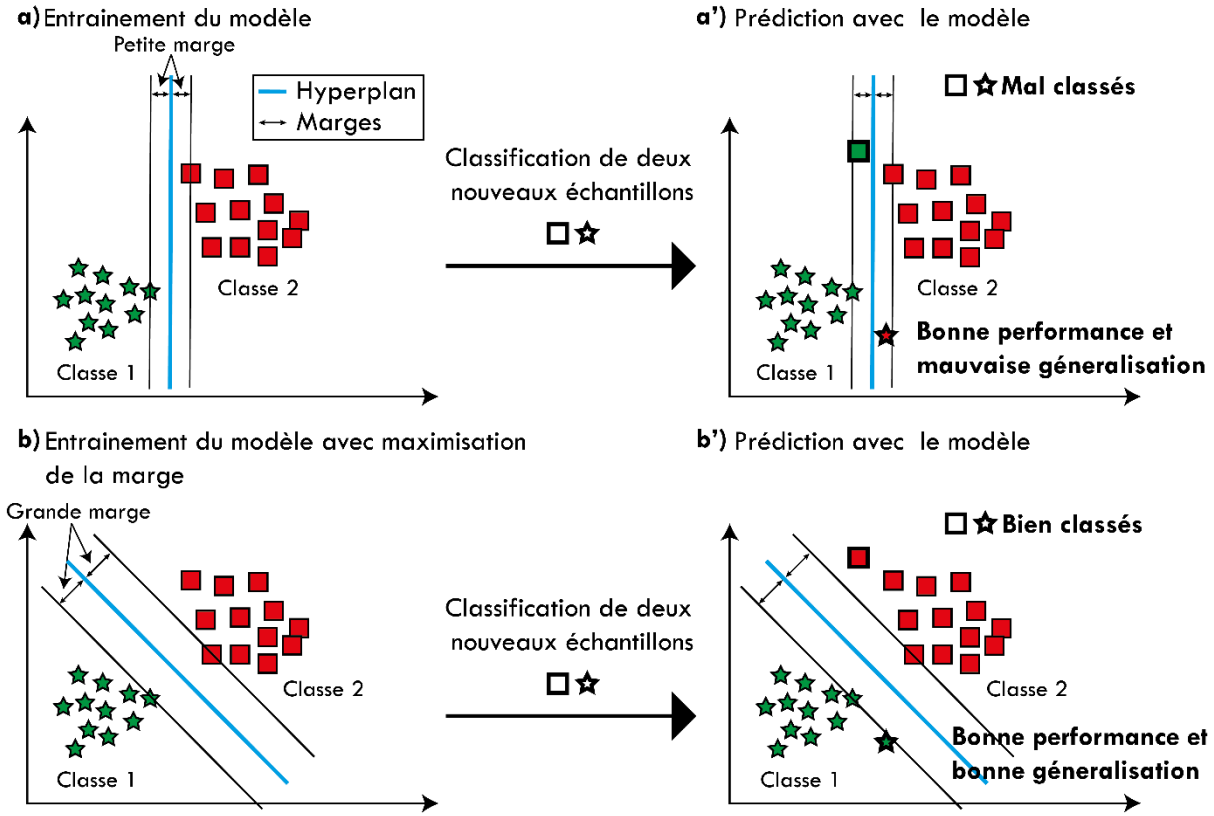


Figure II.10 : Principe de la classification par machines à vecteurs de support (SVM). Exemple de l'impact de la maximisation de la marge dans le cas de la classification de deux échantillons non compris dans les données d'apprentissage du modèle SVM. La maximisation de la marge permet , en plus d'une bonne performance, une bonne généralisation du modèle SVM.

La méthode SVM a la capacité de traiter des données séparables soit linéairement, soit non linéairement. Dans le premier cas, elle recherche directement un hyperplan sur les données d'entrée. Pour traiter les données qui ne peuvent pas être séparées linéairement, elle utilise des fonctions de type noyau '*kernel*' (i.e. fonctions mathématiques reliant les données entre elles en exprimant leurs similarités) pour projeter les données dans un nouvel espace de dimension supérieure. Cette projection permet le basculement d'un problème non linéaire vers un problème linéaire dans cet espace de plus haute dimension. Ce problème linéaire peut ensuite être résolu par la SVM. La différence mathématique entre la classification SVM et la régression SVR réside dans la formulation de la fonction objectif (i.e. fonction mathématique utilisée pour trouver l'hyperplan optimal), qui vise soit à maximiser la marge entre les classes pour la classification, soit à minimiser l'erreur de l'apprentissage pour la régression[102]. Pour illustrer le principe de la SVR, prenons le cas de notre matrice $\mathbf{D}(l,m)$ contenant l spectres d'émission (i.e. l vecteurs $\mathbf{d}_i(1,m)$ avec $i = 1,\dots,l$) et un vecteur $\mathbf{y}(l,1)$ contenant l mesures y_i d'une variable quantitative. Le set d'apprentissage pris en compte en SVR se note donc :

$$\{\mathbf{d}_i, y_i \mid \mathbf{d}_i \in \mathbb{R}^m \mid y_i \in \mathbb{R}\}$$

De manière générale, le but d'une régression est de trouver une fonction f modélisant la relation entre les variables spectrales et la variable quantitative :

$$y_i = f(\mathbf{d}_i) \tag{II.34}$$

La régression SVR présente le même objectif qu'une régression classique à la différence qu'elle ne tient pas compte des erreurs faibles (i.e. observations se trouvant à une distance prédéfinie, appelée marge et notée ε, de la vraie valeur).

II.5.3.1 Régression SVR linéaire

Dans le cas d'une modélisation linéaire des données, la fonction SVR se note sous la forme d'un produit scalaire :

$$f(\mathbf{d}) = \boldsymbol{\omega} \cdot \mathbf{d} + \beta \tag{II.35}$$

Où l'opérateur · signifie le produit scalaire. Le vecteur $\boldsymbol{\omega}$ est le vecteur poids qui définit la direction perpendiculaire à l'hyperplan et β est une constante relative au biais. La fonction SVR doit être la plus plate possible pour assurer une bonne généralisation du modèle. Ainsi, le critère utilisé pour la platitude de la fonction est la minimisation de la norme euclidienne du vecteur poids $\|\boldsymbol{\omega}\|$ (i.e. minimisation de la longueur de ce vecteur perpendiculaire à l'hyperplan et donc maximisation de la platitude de l'hyperplan). Ce problème de minimisation peut être formulé sous la forme primale d'un problème quadratique convexe respectant les contraintes liées à la marge[64,66] :

$$\text{argmin } (\frac{1}{2}\|\boldsymbol{\omega}\|^2) \text{ sous contraintes:} \begin{cases} y_i - f(\mathbf{d}_i) \leq \varepsilon \\ f(\mathbf{d}_i) - y_i \leq \varepsilon \end{cases} \tag{II.36}$$

Cependant, dans certains cas, une fonction ajustant les données avec une précision ε (i.e. contraintes citées ci-dessus) n'existe pas. Par conséquent, de nouvelles variables de relâchement, appelées variables d'écart et notées ξ_i^+ et ξ_i^- sont introduites. Ces dernières confèrent au modèle SVR un caractère à la fois flexible, par le relâchement des contraintes sur la marge, et à la fois rigoureux, par la pénalisation des erreurs importantes, ce qui diminue la sensibilité du modèle aux points aberrants[64,102]. Ce critère d'assouplissement est à l'origine de la fonction de coût ε-SVR dite ε-insensible[66] (Figure II.11) :

$$|\xi|_\varepsilon = \begin{cases} 0, & si \ \ |f(\mathbf{d}) - y| \leq \varepsilon \\ |f(\mathbf{d}) - y| - \varepsilon, & sinon \end{cases} \tag{II.37}$$

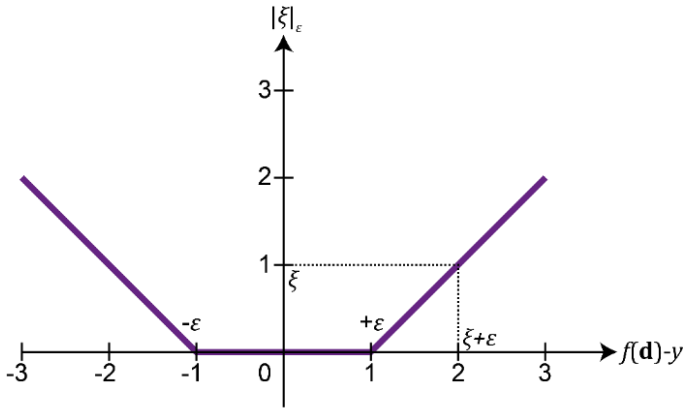


Figure II.11 : La fonction de coût ε-SVR, avec $\varepsilon = \pm 1$.

Ainsi, le problème de minimisation (II.36) devient :

$$\begin{aligned} &\text{argmin}\left(\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{l}(\xi_i^+ + \xi_i^-)\right) \\ &\text{Sous contraintes:} \begin{cases} y_i - f(\mathbf{d}_i) \leq \varepsilon + \xi_i^+ \\ f(\mathbf{d}_i) - y_i \leq \varepsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \,;\, C > 0 \end{cases} \end{aligned} \tag{II.38}$$

Le C est un méta-paramètre qui s'optimise lors du processus d'apprentissage. Il traduit un compromis entre la platitude de la fonction (i.e. généralisation du modèle) et l'erreur de l'apprentissage (i.e. performance du modèle). Ce problème d'optimisation convexe peut être résolu, dans sa forme duale[66], en utilisant les multiplicateurs de Lagrange α_i et α_i^*, avec $\alpha_i \geq 0$ et $\alpha_i^* \leq C$, pour aboutir aux valeurs du vecteur poids :

$$\boldsymbol{\omega} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\mathbf{d}_i \tag{II.39}$$

Seuls les coefficients $(\alpha_i - \alpha_i^*)$ non nuls sont retenus pour la détermination de la courbe de régression. Les vecteurs $\mathbf{d}_i$ y correspondants sont appelés vecteurs de support. Ils se retrouvent sur ou à l'extérieur de la marge ε (Figure II.12). Enfin, l'équation finale du modèle SVR linéaire est obtenue en remplaçant dans l'équation (II.35) le terme $\boldsymbol{\omega}$ par son expression (II.39) :

$$f(\mathbf{d}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\mathbf{d} \cdot \mathbf{d}_i) + \beta \tag{II.40}$$

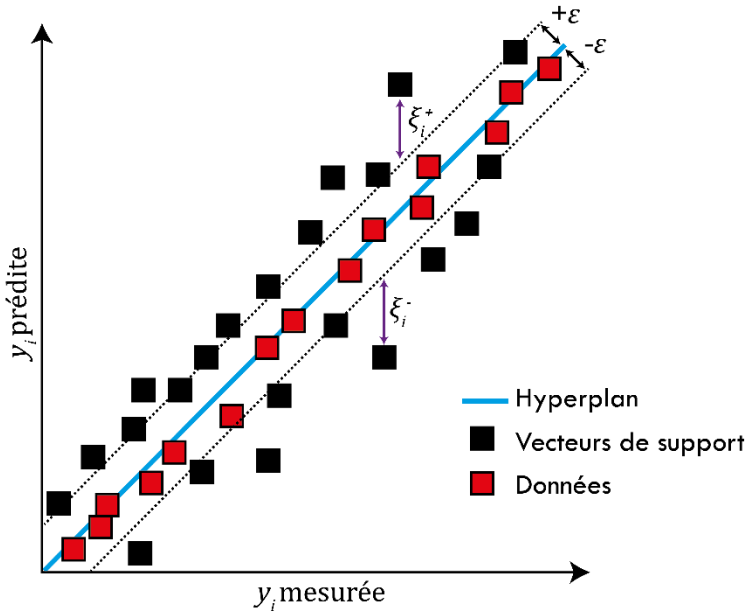


Figure II.12 : Illustration de la courbe de régression SVR. Les vecteurs de support se retrouvent sur ou à l'extérieur de la marge ε. ξ_i^+ et ξ_i^- sont les variables d'écart du modèle.

II.5.3.2 Régression SVR non-linéaire

Comme pour la SVM, la régression SVR utilise, dans le cas d'une modélisation non-linéaire des données, une projection des données d'entrée situées dans l'espace $\mathbb{R}^m$ vers un espace de plus grande dimension noté E :

$$\mathbf{d}_i \rightarrow \phi(\mathbf{d}_i) \mid \mathbf{d}_i \in \mathbb{R}^m \text{ et } \phi(\mathbf{d}_i) \in E$$

Ainsi, la fonction SVR non linéaire implique une expression implicite du vecteur $\boldsymbol{\omega}$ dans l'espace E au lieu de l'espace $\mathbb{R}^m$ [103]. Elle se note :

$$f(\mathbf{d}) = \boldsymbol{\omega} \cdot \phi(\mathbf{d}) + \beta \qquad \text{(II.41)}$$

En résolvant ce problème de la même façon que la fonction SVR linéaire, l'équation finale du modèle SVR non linéaire est[66] :

$$f(\mathbf{d}) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\phi(\mathbf{d}) \cdot \phi(\mathbf{d}_i)) + \beta \qquad \text{(II.42)}$$

Le problème avec ce type de transformation est, qu'à mesure que la dimension m de l'espace de départ $\mathbb{R}^m$ augmente, la dimension de l'espace E s'accroit de façon conséquente. Ainsi, il est parfois numériquement impossible de résoudre le problème. Par conséquent, une fonction noyau $K(\mathbf{d}, \mathbf{d}_i)$ est introduite pour remplacer le produit $\phi(\mathbf{d}) \cdot \phi(\mathbf{d}_i)$ dans l'espace de haute dimension **:**

$$f(\mathbf{d}) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(\mathbf{d}, \mathbf{d}_i) + \beta \quad \text{(II.43)}$$

Les fonctions noyau les plus couramment utilisées sont celles basées sur la fonction linéaire, polynomiale, sigmoïde ou gaussienne (Tableau II.2). Le noyau gaussien '*Radial basis function*' (RBF) est très largement utilisé en chimiométrie grâce à sa flexibilité due au paramètre γ qui est à optimiser, à l'instar de ε et C, lors de l'apprentissage[104].

Tableau II.2 : Les fonctions noyau[102].

Nom de la fonction noyau	Formule mathématique	Numéro de l'équation	Détails
Noyau linéaire	$K(\mathbf{d}, \mathbf{d}_i) = \mathbf{d} \cdot \mathbf{d}_i$	(II.44)	
Noyau polynomial	$K(\mathbf{d}, \mathbf{d}_i) = (\mathbf{d} \cdot \mathbf{d}_i)^d$	(II.45)	d est le degré du polynôme.
Noyau sigmoïde	$K(\mathbf{d}, \mathbf{d}_i) = \tanh((\mathbf{d} \cdot \mathbf{d}_i) + 1)$	(II.46)	
Noyau RBF	$K(\mathbf{d}, \mathbf{d}_i) = \exp\left(-\frac{\|\mathbf{d} \cdot \mathbf{d}_i\|^2}{2\sigma^2}\right) = \exp(-\gamma \|\mathbf{d} \cdot \mathbf{d}_i\|^2)$	(II.47)	$\gamma = \frac{1}{2\sigma^2}$ représente la largeur du noyau RBF.

II.6 Contextualisation de la thèse et conclusion

II.6.1 Contextualisation de la thèse

Les différentes techniques analytiques génèrent souvent d'importantes quantités de données protéiformes. Pour les gérer efficacement et en extraire le maximum d'informations chimiques pertinentes, l'utilisation de la chimiométrie s'avère essentielle. Elle intervient tant en amont des mesures en élaborant des plans d'expériences qu'en aval des mesures grâce à la mise en œuvre d'approches de traitement de données.

Dans la littérature, différents travaux en chimiométrie ont été réalisés pour résoudre des problèmes complexes liés à l'identification, à la quantification et à la caractérisation des CAP dans les différents compartiments environnementaux. Plusieurs travaux ont été sélectionnés et sont brièvement présentés pour illustrer cette combinaison chimiométrie/analyse ciblant la problématique des CAP. Des stratégies de plans d'expériences ont été élaborées, pour étudier par exemple la dépendance des variables opérationnelles de l'extraction par solvant sur la récupération des 16 HAP réglementaires à partir de sols contaminés[105]. En outre, la matrice de Box-Behnken a été utilisée pour optimiser des expériences de pyrolyse couplée à la chromatographie gazeuse et la spectrométrie de masse

(Pyr-GC-MS) pour obtenir des performances optimales en séparation chromatographique des HAP présents dans un sol contaminé, sans extraction préalable[106]. Par ailleurs, des méthodes exploratoires d'analyse de variance, telles que l'ACP, ont été déployées pour la classification d'échantillons environnementaux marins basée sur l'analyse chromatographique des hydrocarbures, notamment les HAP[107]. Aussi, la présence de HAP, leurs concentrations et leur évolution temporelle en milieu marin ont été étudiés en couplant données GC-MS et traitements par ACP[108]. Enfin, des méthodes exploratoires de résolution multivariée de courbes, telles que PARAFAC, ainsi que des méthodes d'apprentissage quantitatif, telles que la PLSR, ont été utilisées, d'une part, pour extraire les signaux sources relatifs aux CAP à partir de données spectroscopiques[109–111], et d'autre part, pour modéliser l'aspect quantitatif de ces derniers[112–114]. De manière générale, la littérature démontre la pertinence des outils de chimiométrie, ainsi que les progrès constants dans leur application à la problématique environnementale de pollution aux CAP et particulièrement aux HAP réglementaires. Cependant, des progrès restent encore à faire, notamment pour la résolution multivariée des courbes et la modélisation quantitative des données, en adoptant une réflexion sur des stratégies de modélisation différentes et en abordant notamment le domaine de la fusion de données issues de différentes techniques analytiques complémentaires. Ceci dans le but d'élargir les études à d'autres CAP tout aussi essentiels à étudier que les HAP réglementaires.

Avant l'utilisation d'algorithmes de chimiométrie pour le traitement de données, la démarche en chimiométrie consiste à comprendre d'abord la nature même du signal mesuré et la correction éventuellement nécessaire à réaliser en prétraitant les données. Certains algorithmes de prétraitement peuvent être communs à des données provenant de différentes techniques analytiques, tels que les algorithmes de correction de la ligne de base. Par exemple, l'algorithme itératif 'Asymmetric Least Squares' basé sur le filtre de Whittaker (WLS), détaillé en annexe C, est appliqué avec succès à des données de spectroscopie infrarouge[115], à des données de spectroscopie Raman[104] mais aussi à des données de spectroscopie de résonnance magnétique nucléaire[116]. D'autres algorithmes peuvent être spécifiques aux données d'une technique analytique particulière, comme ceux permettant d'atténuer les effets de diffusion de la lumière dans les MEEF en spectroscopie de fluorescence 3D. Dans le chapitre III, nous détaillerons un nouvel algorithme qui a été développé, dans le cadre de cette thèse, pour cet objectif.

II.6.2 Conclusion

Dans ce chapitre d'état de l'art, nous avons évoqué et expliqué le caractère trilinéaire des signaux de fluorescence 3D ainsi que les méthodes de chimiométrie pouvant être appliquées. Il ne fait aucun doute que le potentiel de ces méthodes dans l'exploitation des

MEEF, de leur correction et de leur analyse pour la recherche d'informations pertinentes, en passant par l'apprentissage, sont de véritables atouts pour identifier et caractériser des échantillons environnementaux complexes.

Néanmoins, la nature même des contaminants issus de sols étudiés dans cette thèse ont une incidence sur les résultats des méthodes de chimiométrie présentées. A titre d'exemple, le rapport signal sur bruit relativement faible de certaines MEEF peut engendrer une caractérisation biaisée de la réalité chimique. De même, le choix du rang matriciel (i.e. de la dimension d'un modèle) s'avère complexe lorsqu'un échantillon est composé de plusieurs espèces chimiques conduisant à un recouvrement spectral lors de l'utilisation de méthodes de résolution multivariée de courbe, type PARAFAC.

Bien évidemment, différentes stratégies expérimentales et mathématiques existent pour tenter de lever ou du moins minimiser ces inconvénients. Ainsi, dans le prochain chapitre, nous présenterons un nouvel algorithme appelé MT-SVD pour '*Multi-Truncated Singular Value Decomposition*' permettant de corriger efficacement les effets indésirables de la mesure et de donner une estimation relativement précise du rang d'une matrice.

CHAPITRE III

DEVELOPPEMENT D'UN ALGORITHME POUR LE PRETRAITEMENT DES MATRICES D'EXCITATION-EMISSION EN FLUORESCENCE 3D : **MT-SVD**

'***M****ULTI-****T****RUNCATED ****S****INGULAR ****V****ALUE ****D****ECOMPOSITION*'

Un article sur l'algorithme MT-SVD est publié sur https://doi.org/10.1021/acsomega.2c02256

III.1 Introduction

Comme précisé dans le chapitre II, déterminer la dimension optimale d'un modèle en chimiométrie n'est pas une tâche facile. Ce choix a un impact direct sur la pertinence chimique du modèle, tout comme le prétraitement en amont des données spectrales, tels que les MEEF. C'est une des raisons pour lesquelles un nouvel algorithme appelée MT-SVD ('*Multi-truncated Singular Value Decomposition*') a été développé pour le prétraitement des MEEF. Cet algorithme permet la gestion des effets de diffusion de la lumière, du bruit instrumental et des déficiences du rang matriciel. Il est basé sur la méthode de décomposition en valeurs singulières '*Singular Value Decomposition*' (SVD) qui est l'un des algorithmes de factorisation matricielle les plus utilisés en algèbre linéaire pour la réduction en dimension de données ou l'estimation du rang matriciel. La SVD est présente dans bien d'autres domaines d'applications tels que la compression d'images[117]. Pour illustrer son principe, prenons le cas d'une matrice rectangulaire $\mathbf{X}(n,m)$ contenant n lignes chacune caractérisée par m colonnes avec $n < m$. La SVD décompose cette matrice en un produit matriciel :

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{III.1}$$

Où $\mathbf{U}(n,n)$ et $\mathbf{V}(m,m)$ sont respectivement, la matrice des vecteurs singuliers de gauche et la matrice des vecteurs singuliers de droite. Ces deux matrices carrées sont unitaires car les vecteurs colonnes (i.e. vecteurs singuliers) qui les composent (i.e. $\mathbf{u}$ et $\mathbf{v}$, respectivement) sont orthonormés. En d'autres termes, ces deux matrices vérifient respectivement, les égalités : $\mathbf{UU}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{\boldsymbol{n}}$ et $\mathbf{VV}^{\mathrm{T}} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{\boldsymbol{m}}$ avec $\mathbf{I}_{\boldsymbol{n}}$ et $\mathbf{I}_{\boldsymbol{m}}$, les matrices d'identité de dimension (n,n) et (m,m), respectivement. $\mathbf{S}(n,m)$ est la matrice diagonale des valeurs singulières σ_i pour $i = 1,..,n$. Ces valeurs singulières sont toujours réelles, non négatives et classées dans l'ordre décroissant. Leur nombre coïncide avec la plus petite dimension de la matrice de départ $\mathbf{X}$, dans notre cas avec la dimension n (Figure III.1).

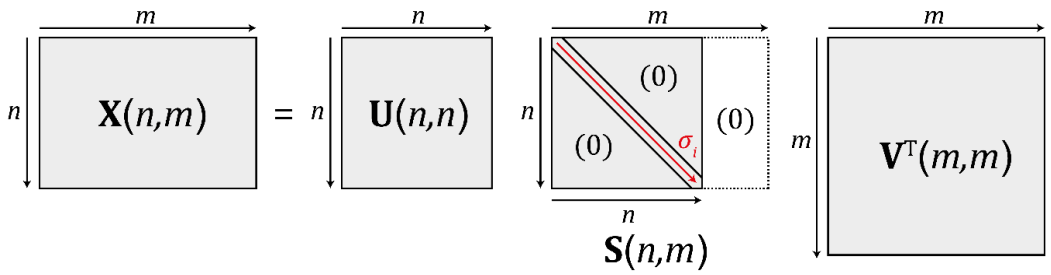


Figure III.1 : Schématisation du principe de la factorisation SVD.

Après factorisation SVD, il est possible de reconstruire, avec l'équation III.2, une matrice estimée de $\mathbf{X}$, notée $\widehat{\mathbf{X}}$ qui ne contient que les informations potentiellement pertinentes, tout en conservant les dimensions de la matrice $\mathbf{X}$ de départ :

$$\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{S}}\widehat{\mathbf{V}}^{\mathrm{T}} \qquad \text{(III.2)}$$

Cette opération de troncature SVD se réalise en préservant uniquement les k valeurs singulières les plus significatives pour reconstruire la matrice $\widehat{\mathbf{X}}$. Les valeurs singulières sont notées σ_i avec $i = 1,..,k$, et leurs vecteurs respectifs sont contenus dans les matrices $\widehat{\mathbf{U}}(n,\mathrm{k})$ et $\widehat{\mathbf{V}}(k,m)$. La matrice $\widehat{\mathbf{S}}(k,k)$ est la matrice diagonale des k valeurs singulières les plus significatives (Figure III.2). Le défi avec la stratégie de troncature SVD est la détermination du nombre des k valeurs singulières les plus significatives. Cette information n'est, en effet, pas toujours connue a priori. Toutefois, les valeurs singulières sont définies comme les racines carrées des valeurs propres. Ces dernières sont directement liées au pourcentage de variance capturée par un vecteur singulier[118].

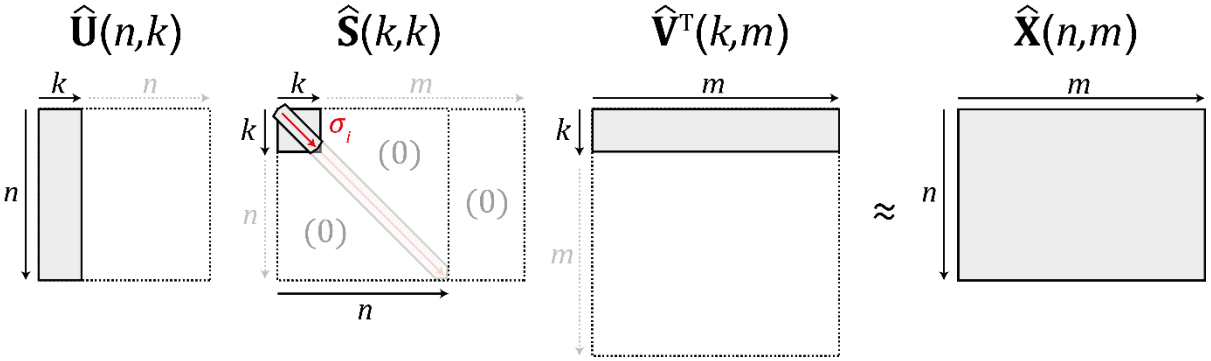


Figure III.2 : Schématisation du principe de l'opération de troncature SVD.

Ainsi, pour tronquer $\mathbf{X}$ et reconstruire notre matrice estimée $\widehat{\mathbf{X}}$, nous estimons souvent le nombre k par une évaluation visuelle de la courbe d'évolution des valeurs singulières. Ce nombre est en réalité une estimation du rang matriciel global de la matrice de départ $\mathbf{X}$ qui pour rappel, représente la dimension du sous-espace vectoriel engendré par les vecteurs-ligne ou les vecteurs-colonne de cette matrice. Cette estimation est permise car le rang matriciel (i.e. mathématique) est analogue au rang chimique du système (i.e. nombre d'espèces chimiques impliquées dans le système étudié). Ce fait est vrai dans le cas d'un système chimique idéal, non bruitée et obéissant à une loi mathématique linéaire, puisque les signaux mesurés sont alors des combinaisons linéaires de signaux sources.

Cependant, ce cas idéal est éloigné de la réalité et l'estimation du rang matriciel global n'est pas toujours évidente à cause de la présence systématique du bruit dans les données expérimentales, mais aussi à cause de phénomènes de colinéarités, ce qui génère souvent une déficience du rang matriciel. Cette déficience se traduit par une sous-estimation du rang mathématique par rapport à la réalité chimique. Ainsi, la sélection des k valeurs singulières les plus significatives (i.e. rang matriciel global de $\mathbf{X}$) via une évaluation visuelle de la courbe d'évolution des valeurs singulières est très délicate et nécessite la détermination d'un seuil à

partir duquel les informations portées par les valeurs singulières, et leurs vecteurs respectifs sont considérées comme du bruit. Pour illustrer ce point, prenons l'exemple d'une matrice de données $\mathbf{X}$ simulée sans bruit. La détermination du rang global de cette matrice est facile car il correspond au nombre de valeurs singulières non nulles, qui dans ce cas est de deux (Figure III.3a). Maintenant, considérons la même matrice à laquelle nous ajoutons du bruit blanc simulé. Nous observons que les valeurs singulières diminuent mais ne s'annulent pas. Ainsi, nous sommes contraints d'établir un seuil à partir duquel nous considérons que les valeurs singulières et leurs vecteurs correspondants portent des informations relatives au bruit (Figure III.3b). Cette estimation du rang matriciel est importante dans le cas de la reconstruction de notre matrice estimée $\hat{\mathbf{X}}$ par la stratégie de troncature SVD, mais elle est également cruciale pour la pertinence des résultats pour plusieurs algorithmes de chimiométrie de réduction de dimensionnalité et de résolution multivariée de courbes.

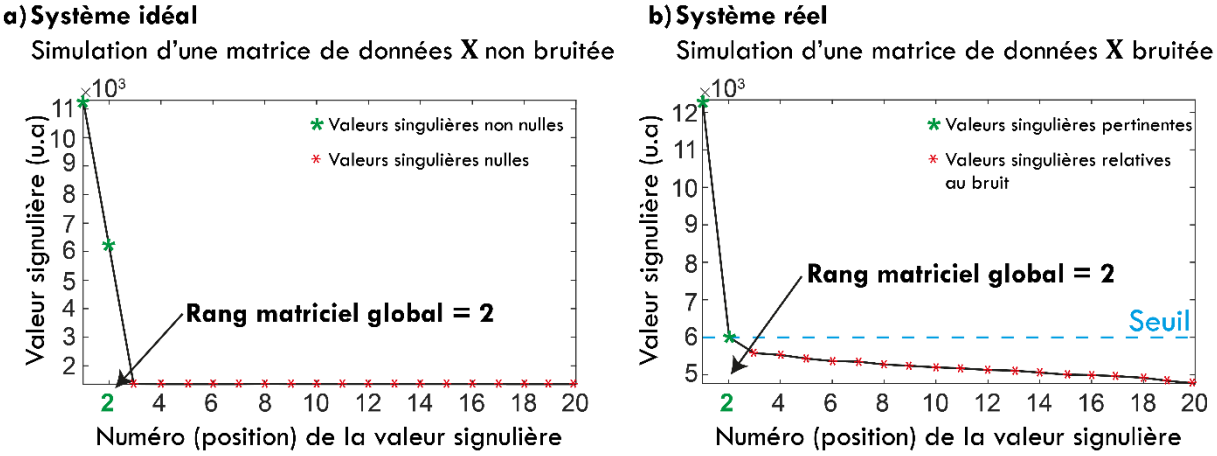


Figure III.3 : Evaluation visuelle de la courbe d'évolution des valeurs singulières dans un système idéal (a) et dans un système bruité (b).

Dans la littérature, d'autres approches sont proposées[119,120] pour tenter d'estimer, dans des systèmes complexes et de manière plus précise, la valeur du rang matriciel la plus juste. L'algorithme MT-SVD explicité dans ce chapitre propose une nouvelle approche pour déduire le rang matriciel global à partir de l'étude des rangs locaux. L'objectif est de comprendre la nature même de l'information de la matrice de donnée de départ $\mathbf{X}$ et ainsi proposer la détection la plus fine possible de son rang global. Cette opération nous permet également la reconstruction d'une matrice de donnée corrigée des effets de diffusion de la lumière et du bruit instrumental.

III.2 Algorithme MT-SVD

L'algorithme MT-SVD est structuré en trois étapes principales : (a) le formatage des données, (b) la recherche d'un ensemble optimal de valeurs singulières à partir d'une stratégie de troncature avancée de la SVD, et (c) la reconstruction de la MEEF prétraitée. Les étapes

de l'algorithme MT-SVD sont résumées dans un logigramme pour faciliter sa compréhension en Figure III.5.

(a) Étape #1 : formatage des données

Cette étape permet de préparer les données pour la factorisation par SVD. Dans le cas où un nombre de l MEEF sont acquises (i.e. l échantillons), il est possible de réaliser un redimensionnement via l'augmentation matriciel du cube de données $\underline{\mathbf{X}}(l, m, n)$ pour basculer de l'espace 3D à l'espace 2D. De plus, une opération de réduction de taille manuelle peut s'avérer utile pour minimiser l'impact des informations chimiques non pertinentes sur une large MEEF en sélectionnant une région d'intérêt notée $\mathbf{X}_{cropped}$. En outre, une contrainte de non-négativité est appliquée, après soustraction du signal du blanc, pour supprimer les intensités négatives et ne conserver que les données relatives à l'information spectrale :

$$\mathbf{X} = \max\{\mathbf{X}, 0\} \tag{III.3}$$

Pour une meilleure compréhension, les étapes suivantes de l'algorithme sont présentées pour une matrice de données brute $\mathbf{X}(n, m)$, où $n < m$. Cette matrice peut être interprétée de deux manières : soit comme une collection de spectres d'excitation et de spectres d'émission, soit comme une image de dimensions $(excitation \times émission)$ où chaque pixel représente l'intensité de fluorescence.

(b) Étape #2 : recherche d'un set optimal de valeurs et vecteurs singuliers à partir d'une stratégie avancée de troncature SVD

Cette étape de l'algorithme débute par une factorisation SVD de la matrice $\mathbf{X}(n, m)$ suivant l'équation (III.1). Ensuite, $\sum_{i=1}^{n} \sigma_i$ est considéré et il correspond au maximum d'information pertinente ou non dans les données de départ. Les fréquences cumulées f_i (en %) des valeurs singulières σ_i sont alors calculées avec la formule suivante :

$$f_i = \frac{\sum_i \sigma_i}{\sum_{i=1}^{n} \sigma_i} \times 100 \tag{III.4}$$

Les valeurs f_i obtenues fournissent alors un pourcentage d'informations dans les données de départ (similaire à une variance cumulée) ajoutée à chaque étape de 1 à n. Ensuite, nous déterminons les rangs locaux r_k pour un k allant de 1 à 100 comme étant un nombre de σ_i capturant un pourcentage de 1% à 100% de la fréquence cumulée f_i définit avec le critère suivant :

$$r_k = \min_{i=1,n}(\mathbf{f_i} > k/100) \text{ sous contrainte : } \{k = 1, \dots, 100\} \tag{III.5}$$

Dans ce travail de thèse, les valeurs de k sont considérées avec un pas de 1, mais elles peuvent être modifiées par l'utilisateur. Ce pas peut être vu comme étant un seuil pour lequel l'utilisateur souhaite visualiser les informations de la matrice de départ. Par conséquent, nous calculons ici 100 nouvelles matrices $\widehat{\mathbf{X}}_k$ suivant les valeurs r_k associées et avec l'équation suivante :

$$\widehat{\mathbf{X}}_k = \widehat{\mathbf{U}}_{r_k} \widehat{\mathbf{S}}_{r_k} \widehat{\mathbf{V}}_{r_k}^{\mathbf{T}} \tag{III.6}$$

Les matrices $\widehat{\mathbf{U}}(n, r_k)$ et $\mathbf{V}(m, r_k)$ sont respectivement, la matrice des vecteurs singuliers de gauche et la matrice des vecteurs singuliers de droite. La matrice $\mathbf{S}(r_k, r_k)$ est la matrice diagonale des valeurs singulières correspondant aux rangs locaux r_k. Cette stratégie de troncature est une approche originale se démarquant de l'approche de la SVD classique. Le choix du rang global, afin de reconstruire les données non biaisées, est obtenu par investigation de l'information ajoutée entre les valeurs r_k grâce au calcul des matrices $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ avec $j = 1, \dots, k-1$:

$$\widehat{\mathbf{X}}_j^{\mathrm{ADD}} = \max\{\widehat{\mathbf{X}}_{j+1} - \widehat{\mathbf{X}}_j, 0\} \tag{III.7}$$

L'information résiduelle qui en découle est aussi analysée. Elle est obtenue grâce au calcul des $\widehat{\mathbf{X}}_k^{\mathrm{residual}}$ définit par l'équation suivante :

$$\widehat{\mathbf{X}}_k^{\mathrm{residual}} = \max\{\mathbf{X} - \widehat{\mathbf{X}}_k, 0\} \tag{III.8}$$

Une fois ces calculs effectués, la recherche du rang global optimal se réalise suivant 3 étapes intermédiaires d'analyse d'images :

- Une première sélection par un test binaire est effectuée sur le cube $\underline{\widehat{\mathbf{X}}}_j^{\mathrm{ADD}}(n, m, j)$. L'objectif est de réduire les dimensions de ce cube de données en identifiant des matrices nulles. En d'autres termes, il s'agit de repérer les matrices où aucune information chimique et spatiale n'est ajoutée entre deux $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ successives.

- Une deuxième sélection est effectuée en étudiant les distributions des valeurs des pixels calculées suivant l'allure des histogrammes relatifs à ces pixels pour chacune des matrices $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ précédemment sélectionnées. L'aire sous chaque courbe de distribution est alors calculée puis, la sélection des cartes à conserver est effectuée en fonction de leurs valeurs (i.e. notion de seuillage). En effet, plus la valeur de ce paramètre est faible, plus il est probable que le $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ soit un artefact, un bruit ou encore un signal Rayleigh faible. De plus, si plusieurs $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ peuvent avoir les mêmes valeurs d'aires, cela met en évidence qu'il n'y a pas d'ajout de nouvelle information dans le

signal de fluorescence. Ainsi, plus le nombre de $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ est grand, plus la redondance des informations ajoutées est grande. A ce stade, les matrices $\widehat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ sont déterminées à partir de l'ensemble des $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ comme celles ne contenant que de l'information pertinente.

- Un algorithme de segmentation d'images par régions[121] est exécuté sur chaque $\widehat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ pour extraire les bords extérieurs de régions contenus dans l'image analysée afin de les superposer sur la carte $\widehat{\mathbf{X}}_k$ associée. Cette opération commence par la conversion de chaque $\widehat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ en une image de niveaux de gris. Ensuite, cette image est convertie en une image binaire, où les valeurs des pixels sont soit 0 soit 1, en utilisant un seuil global déterminé par la méthode d'Otsu[122]. Cette méthode vise à minimiser la variance intra-classe des pixels seuillés en noir et blanc en utilisant l'intensité moyenne locale du voisinage de chaque pixel. Les pixels ayant une valeur de 1 correspondent aux régions porteuses d'informations, et leurs extrémités dans toutes les directions de l'espace représentent les bords extérieurs des régions contenues dans l'image analysée. L'objectif de cette étape intermédiaire est de comprendre la nature du signal ajouté (i.e. les signaux de fluorescence, la diffusion Rayleigh ou le bruit) en étudiant sa localisation dans le plan excitation-émission. Dans le même temps, les cartes résiduelles $\widehat{\mathbf{X}}_k^{\mathrm{residual}}$ correspondantes sont également tracées pour s'assurer que toutes les informations chimiques pertinentes de fluorescence sont bien capturées.

A l'issue de cette analyse d'image, les $\widehat{\mathbf{X}}_{j_{\mathrm{accepted}}}^{\mathrm{ADD}}$ sont choisies à partir des $\widehat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$. Les matrices $\widehat{\mathbf{X}}_{j_{\mathrm{accepted}}}^{\mathrm{ADD}}$ capturent donc les informations chimiques pertinentes liées aux valeurs des r_k associés, qui eux-mêmes correspondent à des valeurs singulières sélectionnées σ_i. Ainsi, les r_k et les σ_i correspondant ne sont pas nécessairement successifs. Ceci constitut l'atout majeur de notre approche puisqu'elle permet de visualiser s'il y a lieu une ou plusieurs déficiences de rangs locaux observés, pour à la fin de la sélection obtenir une approximation du rang global moins biaisé.

(c) Étape #3 : Reconstruction de la carte chimique optimale

La reconstruction de la matrice multi-tronquée notée $\widehat{\mathbf{X}}_{\mathrm{truncated}}(n,m)$ s'effectue à partir de l'ensemble des σ_i sélectionnées, noté Σ (équation III.9), en association avec les vecteurs singuliers correspondants (Figure III.4) :

$$\Sigma = \{\sigma_i | i \in [1,n]\} \tag{III.9}$$

$$\hat{\mathbf{X}}_{\text{truncated}} = \hat{\mathbf{U}}_{\boldsymbol{\Sigma}}\hat{\mathbf{S}}_{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{\boldsymbol{\Sigma}}^{\mathbf{T}} \quad \text{(III.10)}$$

L'approximation optimale du rang global se lit donc comme le nombre total de rangs locaux r_k retenus (i.e. nombre de σ_i chimiquement pertinents).

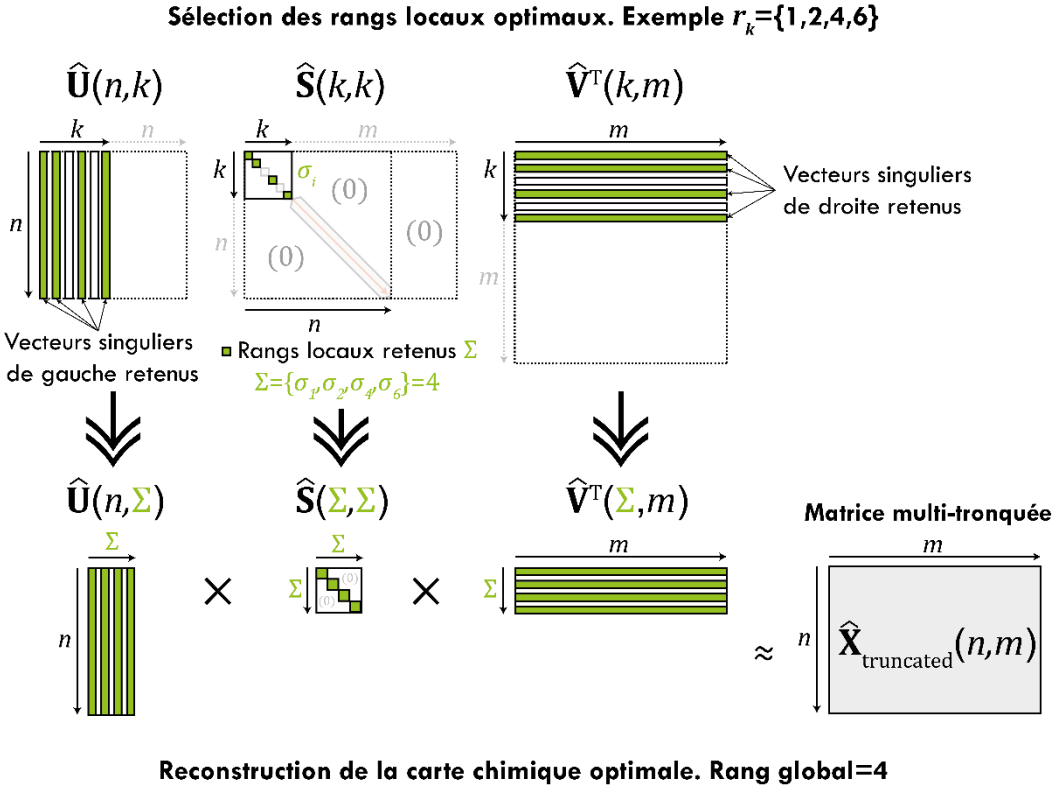


Figure III.4 : Reconstruction MT-SVD de la MEEF optimale après sélection des rangs locaux optimaux.

Enfin, il est possible lors de cette étape, de recadrer automatiquement la matrice $\hat{\mathbf{X}}_{\text{truncated}}$ avec le même algorithme de segmentation par régions que précédemment[121] pour se focaliser uniquement sur l'information chimique pertinente. Cette réduction de dimensions peut s'avérer utile ultérieurement pour accélérer les calculs lors de l'utilisation d'autres algorithmes de chimiométrie tels que ceux de la résolution multivariée de courbes.

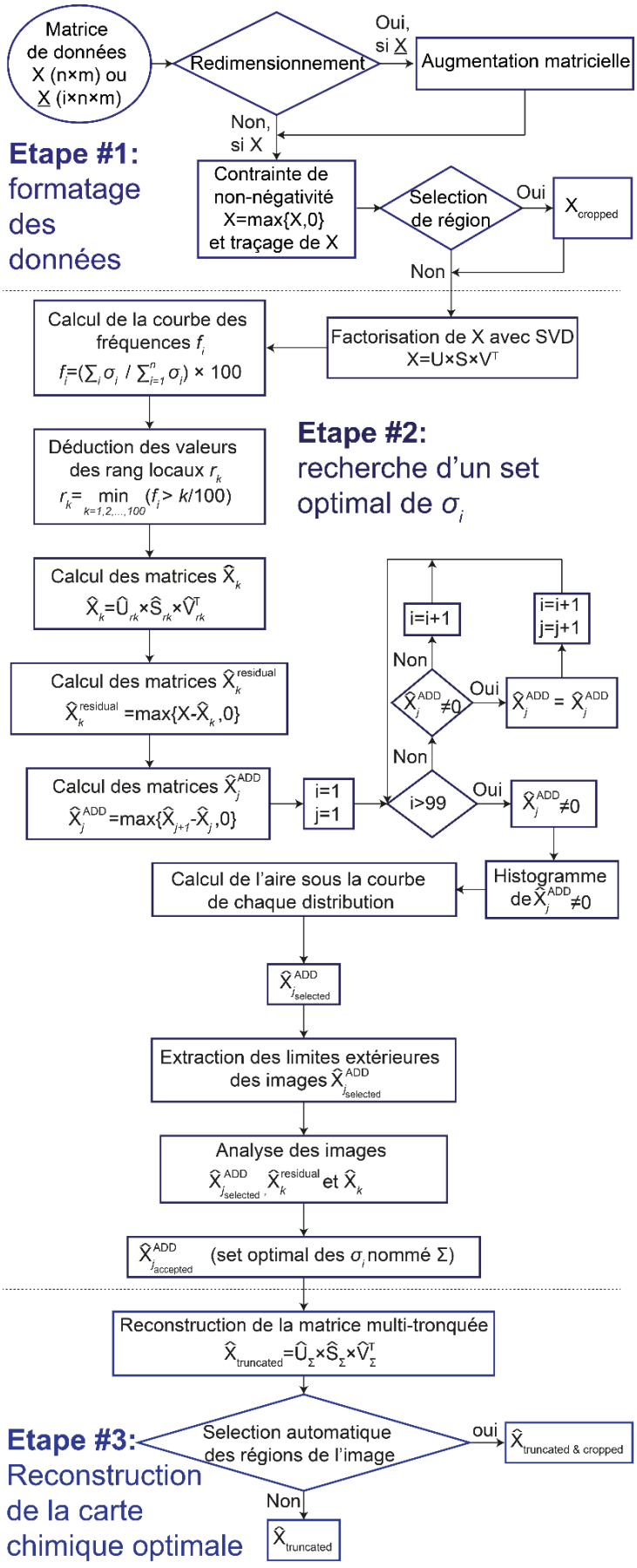

Matrice de données X (n×m) ou X (i×n×m)
Redimensionnement
Oui, si X
Augmentation matricielle
Non, si X
Etape #1: formatage des données
Contrainte de non-négativité X=max{X,0} et traçage de X
Selection de région
Oui
X cropped
Non
Calcul de la courbe des fréquences f_i
$f_i=(\sum_i \sigma_i / \sum_{i=1}^{n} \sigma_i) \times 100$
Factorisation de X avec SVD
$X=U\times S\times V^T$
Déduction des valeurs des rang locaux r_k
$r_k=\min_{k=1,2,...,100}(f_i > k/100)$
Etape #2: recherche d'un set optimal de σ_i
Calcul des matrices $\hat{X}_k$
$\hat{X}_k=\hat{U}_{rk}\times\hat{S}_{rk}\times\hat{V}^T_{rk}$
Calcul des matrices $\hat{X}_k^{residual}$
$\hat{X}_k^{residual}=\max\{X-\hat{X}_k,0\}$
Calcul des matrices $\hat{X}_j^{ADD}$
$\hat{X}_j^{ADD}=\max\{\hat{X}_{j+1}-\hat{X}_j,0\}$
i=1 j=1
i>99
Non
Oui
$\hat{X}_j^{ADD}\neq 0$
$\hat{X}_j^{ADD}=\hat{X}_j^{ADD}$
i=i+1
i=i+1 j=j+1
Histogramme de $\hat{X}_j^{ADD}\neq 0$
Calcul de l'aire sous la courbe de chaque distribution
$\hat{X}_{j_{selected}}^{ADD}$
Extraction des limites extérieures des images $\hat{X}_{j_{selected}}^{ADD}$
Analyse des images $\hat{X}_{j_{selected}}^{ADD}$, $\hat{X}_k^{residual}$ et $\hat{X}_k$
$\hat{X}_{j_{accepted}}^{ADD}$ (set optimal des σ_i nommé Σ)
Reconstruction de la matrice multi-tronquée
$\hat{X}_{truncated}=\hat{U}_\Sigma\times\hat{S}_\Sigma\times\hat{V}^T_\Sigma$
Etape #3: Reconstruction de la carte chimique optimale
Selection automatique des régions de l'image
oui
$\hat{X}_{truncated\ \&\ cropped}$
Non
$\hat{X}_{truncated}$

Figure III.5 : Logigramme de l'algorithme MT-SVD.

III.3 Mise en œuvre et preuves de concept de l'algorithme MT-SVD

Pour illustrer les différentes étapes de l'algorithme et son intérêt en prétraitement, des MEEF de quatre HAP purs considérés comme « modèles » ont été acquises pour différentes conditions expérimentales constituant une base de données.

III.3.1 Matériels et méthodes

III.3.1.1 Instrumentation

Le spectrofluorimètre 3D (Aqualog ®) est utilisé pour acquérir les MEEF. Il est équipé d'un détecteur à dispositif de couplage de charge (CCD) réglé sur un gain moyen et une intégration temporelle de 1 seconde. La source de lumière continue utilisée est une lampe à arc au xénon sans ozone de 150 Watts, qui est couplée à un monochromateur d'excitation. Les échantillons sont excités à l'aide d'une plage de longueurs d'onde d'excitation comprises entre 239 et 800 nm avec un pas de 3 nm. L'émission de fluorescence est collectée dans une plage de longueurs d'onde comprises entre 248.27 et 829.32 nm avec une résolution de 4 pixels (i.e. 2.33 nm). Toutes les données brutes des MEEF ont donc la même taille de 188 × 250 pixels (i.e. 188 spectres d'émission et 250 spectres d'excitation). Une cuve en quartz SUPRASIL® d'une longueur de trajet lumineux de 10 mm est utilisée pour l'acquisition de chaque MEEF.

III.3.1.2 Préparation des échantillons et construction de la base de données

Des MEEF de 35 échantillons constitués à partir de quatre HAP : naphtalène (NPH), benz[a]anthracène (BaA), anthracène (ANT) et pyrène (PYR) ont été acquises et utilisées pour construire une base de données répartie en deux ensembles de données distincts. L'ensemble de données #1 contient les MEEF individuels des quatre HAP et l'ensemble de données #2 contient des MEEF de mélanges des quatre espèces utilisées. Le choix s'est porté sur ces HAP en raison de la proximité des domaines de longueurs d'onde entre eux et de la possibilité de recouvrement spectral en fonction de leurs concentrations (i.e. problèmes de sélectivité). De plus, ces HAP, qui possèdent un nombre de cycles benzéniques compris entre 2 et 4, sont représentatifs de la majorité des HAP contenus dans la liste de l'US-EPA (cf. §Introduction). Le Tableau III.1 présente cette base de données test.

(a) Ensemble de données #1

Pour chacun des quatre HAP, des MEEF ont été acquises à six concentrations différentes (0.05, 0.1, 0.25, 1, 10, 20 mg.L^{-1}). Pour ce faire, des solutions mères ont été préparées dans du dichlorométhane de qualité GC-MS (Carlo Erba) à une concentration de 1

mg.mL^{-1}. Ensuite, elles ont été diluées dans ce même solvant à des concentrations différentes (Tableau III.1). Ces solutions ont été stockées à -20 °C et ramenées à température ambiante avant d'être analysées.

(b) Ensemble de données #2

Les 11 échantillons de concentrations variables (Tableau III.1) des quatre fluorophores i.e. NPH, BaA, PYR et ANT, ont été préparés dans le même solvant en utilisant les solutions de l'ensemble de données #1.

Tableau III.1 : Base de données des MEEF « modèles » ; chiffres en rouge : Saturation du signal ; chiffres en gras : concentration optimale.

	Ensemble de données #1				**Ensemble de données #2**			
	NPH (mg. L^{-1})	BaA (mg. L^{-1})	ANT (mg. L^{-1})	PYR (mg. L^{-1})	NPH (mg. L^{-1})	BaA (mg. L^{-1})	ANT (mg. L^{-1})	PYR (mg. L^{-1})
Echantillon 1	0.05				0.100	0.100	0.100	0.100
Echantillon 2	0.10				0.050	0.050	0.050	0.050
Echantillon 3	0.25				0.005	0.005	0.005	0.005
Echantillon 4	**1.00**				1.000	0.010	0.025	0.100
Echantillon 5	10.00				1.000	0.005	0.100	0.025
Echantillon 6	20.00				1.000	0.010	0.100	0.005
Echantillon 7		0.05			1.000	0.005	0.025	0.010
Echantillon 8		**0.10**			1.000	0.100	0.025	0.010
Echantillon 9		0.25			1.000	0.050	0.100	0.250
Echantillon 10		1.00			1.000	0.010	0.010	0.010
Echantillon 11		10.00			0.100	0.005	0.010	0.025
Echantillon 12		20.00						
Echantillon 13			0.05					
Echantillon 14			**0.10**					
Echantillon 15			0.25					
Echantillon 16			1.00					
Echantillon 17			10.00					
Echantillon 18			20.00					
Echantillon 19				0.05				
Echantillon 20				0.10				
Echantillon 21				**0.25**				
Echantillon 22				1.00				
Echantillon 23				10.00				
Echantillon 24				20.00				

Avant l'analyse, les échantillons ont été soumis à une sonication pendant 15 minutes. Pour chaque acquisition, une réponse du solvant est acquise et seul le signal de la diffusion Raman est efficacement éliminé en soustrayant la matrice de réponse du dichlorométhane

des données. Toutes les analyses ont été réalisées dans une salle climatisée afin de limiter l'impact des variations de température sur la réponse instrumentale.

III.3.2 Résultats et discussions

Cette section présente les résultats obtenus par MT-SVD à travers une série de cas « modèles » que l'on retrouve régulièrement sur des problèmes en chimiométrie. Les deux premiers cas sont relativement simples sur le plan de leur chimie, car ils consistent à appliquer la correction MT-SVD sur les MEEF d'une seule espèce chimique pure (i.e. naphtalène). La différence entre eux réside dans les conditions d'acquisition des MEEF, qui sont optimales pour le premier cas (i.e. un rapport signal/bruit élevé grâce à une concentration optimale) et non optimales pour le deuxième cas (i.e. un rapport signal/bruit faible à cause d'une concentration faible). L'objectif de ces deux premiers cas est de démontrer que la MT-SVD permet de corriger efficacement les effets de diffusion de la lumière et le bruit pouvant être observé dans les MEEF sans accumuler les problèmes liés à la chimie (e.g. recouvrement spectral). Les cas 3 et 4 portent sur la correction MT-SVD d'une MEEF résultant d'un mélange des quatre HAP purs, sans et avec ajout de bruit blanc simulé, respectivement. L'objectif du cas 3 est double. Il vise à illustrer comment la MT-SVD permet de corriger les effets de diffusion et de bruit blanc dans une MEEF, tout en offrant la possibilité de visualiser et de potentiellement résoudre les problèmes liés aux déficiences du rang matriciel. Le cas 4 met en évidence l'efficacité de la correction MT-SVD dans la gestion du bruit de mesure, même lorsque celui-ci est intense. Enfin, le cas 5 concerne la correction MT-SVD de plusieurs MEEF (i.e. plusieurs mélanges) et la décomposition par PARAFAC du cube de données corrigé afin d'extraire les signatures pures de chaque espèce chimique constituant les mélanges.

III.3.2.1 Cas 1 : Prétraitement MT-SVD de la MEEF d'une seule espèce chimique pure acquise à une concentration optimale

L'approche proposée, basée sur une stratégie avancée de troncature SVD, ne nécessite aucune information préalable sur les signaux de diffusion. La SVD de l'étape #2 cherche, en effet, à extraire les informations importantes des données, qui se caractérisent par un rang de faible dimension (i.e. réduction de dimension). Puisque les signaux Rayleigh sont indépendants les uns des autres (i.e. une paire de longueurs d'onde d'émission et d'excitation différente pour chaque signal), ils sont considérés comme une succession de signaux à faible importance par rapport aux signaux de fluorescence et ils sont classés par la SVD dans les rangs de plus grande dimension que ceux des signaux de fluorescence. En effet, les signaux de fluorescence se caractérisent, pour un fluorophore, par des spectres d'émission avec une allure invariante (variation d'intensité uniquement), peu importe la longueur d'onde d'excitation (i.e. loi de Kasha, cf. §I.3.2). Ce phénomène de redondance

engendre ainsi une dominance de ces signaux et un classement dans les rangs de plus faibles dimensions par la SVD. Ce cas est vrai pour les MEEF au-dessus d'un certain seuil de rapport signal sur bruit, et où la déficience du rang matriciel n'est pas observée. Cependant, il s'agit généralement de situations simples sans problèmes de sélectivité entre les composés chimiques et avec un rapport signal sur bruit élevé, ce qui diffère considérablement des échantillons réels.

L'échantillon 4 de l'ensemble de données #1 constitue un bon exemple de ces cas simples, car il s'agit d'une MEEF d'une seule espèce chimique pure (i.e. naphtalène) préparée à une concentration optimale de 1 mg.mL^{-1}. Les signaux chimiques de fluorescence partagent les mêmes longueurs d'onde d'émission à différentes longueurs d'onde d'excitation tandis que pour chaque paire de longueurs d'onde d'émission et d'excitation identiques, il n'y a qu'un signal Rayleigh de 1er ordre (Figure III.6a). Le résultat du prétraitement avec la MT-SVD de cet échantillon illustre que l'information chimique est conservée intacte tandis que les signaux de diffusion ont été supprimés (Figure III.6b).

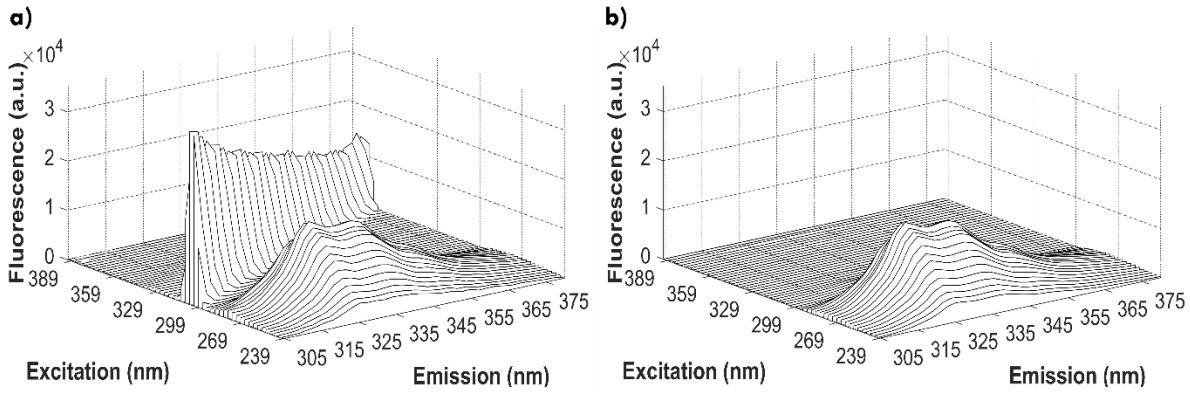


Figure III.6 : MEEF de l'échantillon 4 de l'ensemble de données #1 (a) et résultat après son prétraitement par MT-SVD (b). La carte a été découpé au préalable pour améliorer la visibilité quant au caractère indépendant de chaque signal Rayleigh.

III.3.2.2 Cas 2 : Prétraitement MT-SVD de la MEEF d'une seule espèce chimique pure acquise à une concentration très faible

Dans le cas où les signaux Rayleigh sont nettement plus intenses que les signaux de fluorescence dans une MEEF (i.e. une MEEF avec un rapport signal sur bruit très faible), cela entraîne une prédominance de l'information non pertinente provenant des signaux de diffusion de la lumière par rapport à l'information pertinente provenant des signaux de fluorescence. Cela induit une inversion dans le sens du tri de l'information par la SVD et les signaux de fluorescence sont alors classés dans les rangs de plus grande dimension que ceux des signaux de diffusion. Ce phénomène est dû au fait qu'on se rapproche des limites de détection de l'appareil de fluorescence. Il survient, dans le cas où, la concentration ou le rendement du fluorophore est très faible.

L'échantillon 1 de l'ensemble de données #1 constitue un bon exemple de ce genre de cas avec un rapport signal sur bruit très faible. En effet, le naphtalène présente naturellement un faible rendement de fluorescence par rapport aux autres espèces utilisées dans cette étude en raison de sa structure chimique (i.e. seulement deux cycles aromatiques, cf. §Introduction). De plus, dans cet échantillon, sa concentration est très faible (i.e. 0.05 $mg.L^{-1}$). Dans la MEEF obtenue, la signature chimique de fluorescence est à peine visible par rapport au signaux Rayleigh du 1er ordre, même après soustraction du blanc (Figure III.7).

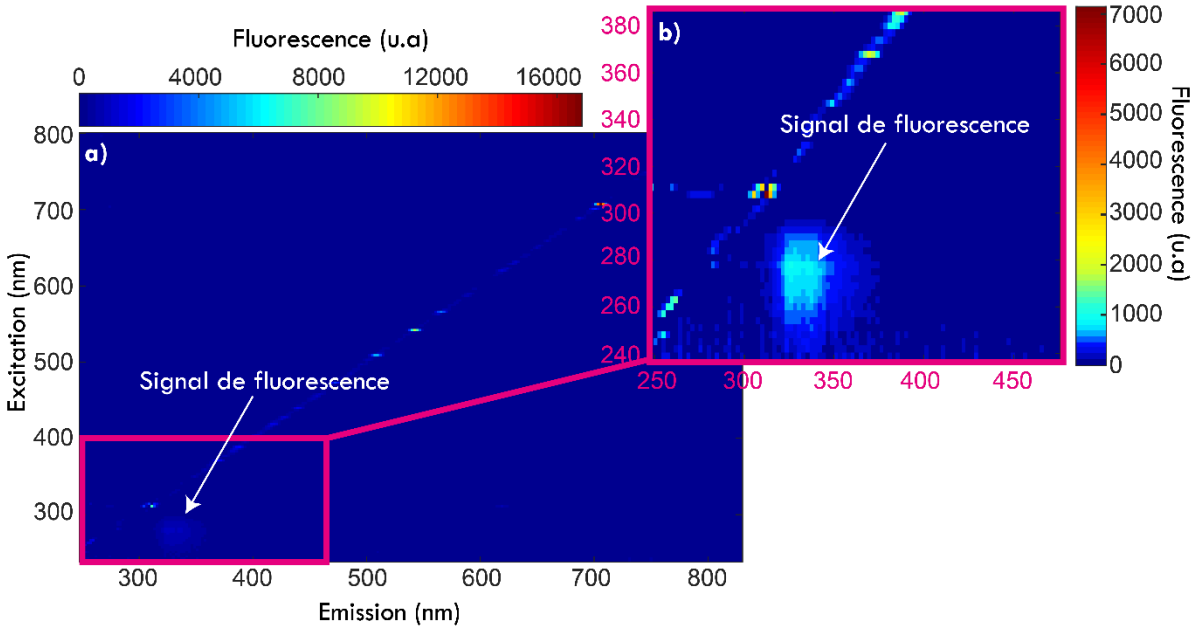


Figure III.7 : MEEF de l'échantillon 1 de l'ensemble de données #1 (a) avec un zoom sur la zone du signal de fluorescence (b).

Dans ce type de cas, il est possible d'utiliser deux approches proposées dans l'étape #1 de l'algorithme (cf. §III.2). En première intention, nous pouvons utiliser l'augmentation matricielle qui permet d'enrichir les données avec des informations spectrales pertinentes. Si l'augmentation matricielle n'est pas réalisable en raison d'un manque de temps d'acquisition de données, nous pouvons opter pour l'opération de réduction de taille qui permet un gain qualitatif des données en diminuant la quantité d'informations provenant des signaux de diffusion de la lumière. Le résultat MT-SVD obtenu sur cette MEEF (Figure III.8a) après augmentation matricielle suivant la dimension d'excitation et au moyen des quatre premiers échantillons de l'ensemble de données #1 (Figure III.8b) montrent clairement que cette stratégie est efficace lorsque le signal de fluorescence du fluorophore est bien plus faible que le signal de diffusion. Nous notons en effet, la faible émission de fluorescence du NPH dans cet échantillon qui est inférieure à 800 u.a (Figure III.8b'). Dans le cas où nous optons pour la stratégie de réduction de taille de la MEEF vers la zone de fluorescence d'intérêt (Figure III.8c), nous constatons également que cette approche est efficace. En effet, les signaux interférents

et le bruit blanc instrumental restants malgré la réduction de taille ont correctement été filtrés par la MT-SVD (Figure III.8c').

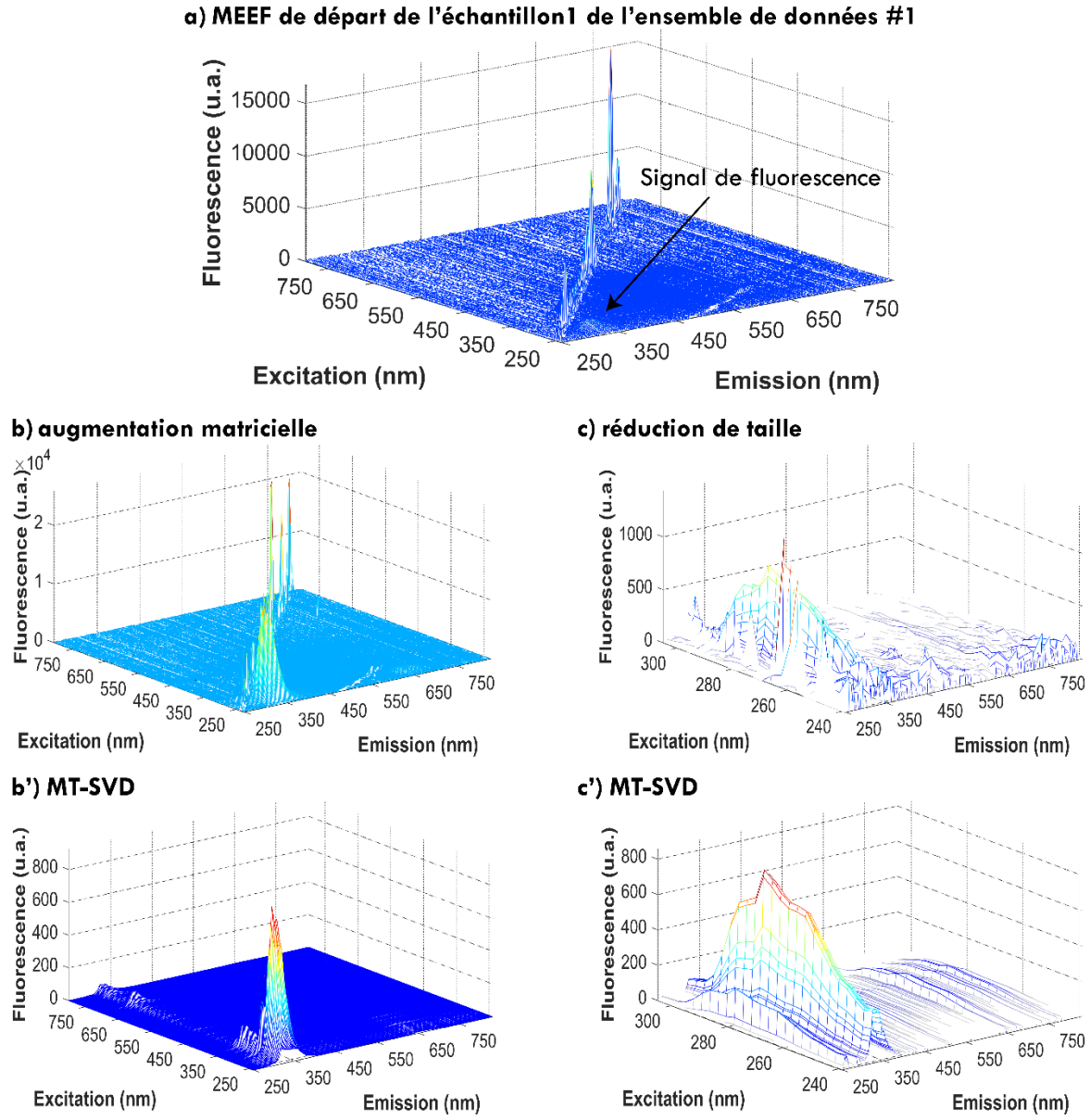


Figure III.8 : MEEF de départ de l'échantillon 1 de l'ensemble de données #1 (a). Augmentation matricielle suivant la dimension d'excitation et au moyen des quatre premiers échantillons de l'ensemble de données #1 (b). Résultats de la correction MT-SVD, après augmentation matricielle, de la MEEF de l'échantillon 1 de l'ensemble de données #1 (b'). Réduction de la taille de la MEEF de départ de l'échantillon 1 de l'ensemble de données #1 vers la zone de fluorescence (c). Résultats de la correction MT-SVD, après réduction de la taille, de la MEEF de l'échantillon 1 de l'ensemble de données #1 (c').

Toutefois, ces deux approches sont efficaces pour ce cas spécifique, qui, malgré un rapport signal sur bruit faible, reste relativement simple sur le plan chimique. En effet, il n'y a qu'une seule espèce chimique impliquée et aucun problème de recouvrement spectral n'est observé.

III.3.2.3 Cas 3 : Prétraitement MT-SVD de la MEEF d'un mélange de quatre espèce chimique pures, acquise à des concentrations relativement faibles

L'objectif avec le cas 3, à savoir, l'échantillon 11 de l'ensemble de données #2 est de montrer que la MT-SVD permet de corriger les effets de diffusion de la lumière et du bruit dans une MEEF, mais aussi et surtout, elle permet de visualiser et gérer les déficiences du rang matriciel.

A partir de l'étape #1 de l'algorithme, les données de départ sont formatées avec seulement l'application de la contrainte de non-négativité (Figure III.9a). Sur la MEEF, le rapport signal sur bruit est acceptable cependant, les effets de diffusion de la lumière sont toujours sur la diagonale. Ensuite, l'étape #2 de l'algorithme recherche l'ensemble optimal Σ de valeurs singulières σ_i avec, la construction des $\widehat{\mathbf{X}}_k$, $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ et $\widehat{\mathbf{X}}_k^{\mathrm{residual}}$. Tout d'abord, une sélection est effectuée sur le cube formé par les $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ pour réduire sa dimension en identifiant les matrices nulles (Figure III.9b). Deuxièmement, l'étude des valeurs des aires calculées sous les courbes de distribution des valeurs des pixels des $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ sélectionnés est effectuée. Le critère de seuillage visuel est choisi égal à 0.6×10^5 (a.u). Les $\widehat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ en vert sont ceux sélectionnés (Figure III.9c). En d'autres termes, lorsque les valeurs d'aire sont faibles, de nombreux $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ ont des valeurs équivalentes symbolisées par des plateaux rouges reflétant une redondance d'informations. Cela implique que les distributions des pixels sont similaires et caractéristiques d'informations ajoutées lorsqu'elles se rapprochent du niveau de bruit de l'instrument ou des effets de diffusion de la lumière relativement faible. En effet, lorsque $\widehat{\mathbf{X}}_j^{\mathrm{ADD}}$ n'a plus d'informations chimiques pertinentes, les distributions des valeurs de pixels deviennent plus petites et plus minces, reflétant ainsi des classes similaires de valeurs de pixels et expliquant que les valeurs des aires soient faibles.

A ce stade, les j_{selected} sont $\{44,51,58,63,66,69,71,73,74\}$ et sont liés à leurs valeurs k correspondantes : $\{45,52,59,64,67,70,72,74,75\}$. Les valeurs de rang locaux sont alors déduites suivant la relation (III.5) (cf. §III.2) avec $r_k = \{2,3,4,5,6,7,8,9,10\}$. Le $r_k = 1$ est systématiquement inclus pour l'analyse d'image car il représente l'information la plus pertinente dans les données puisqu'il est relatif à la première valeur singulière. Enfin, $r_k = \{1,2,3,4,5,6,7,8,9,10\}$ est considéré dans cet exemple et représenté avec les informations capturées par chaque valeur singulière σ_i sélectionnée (Figure III.9d).

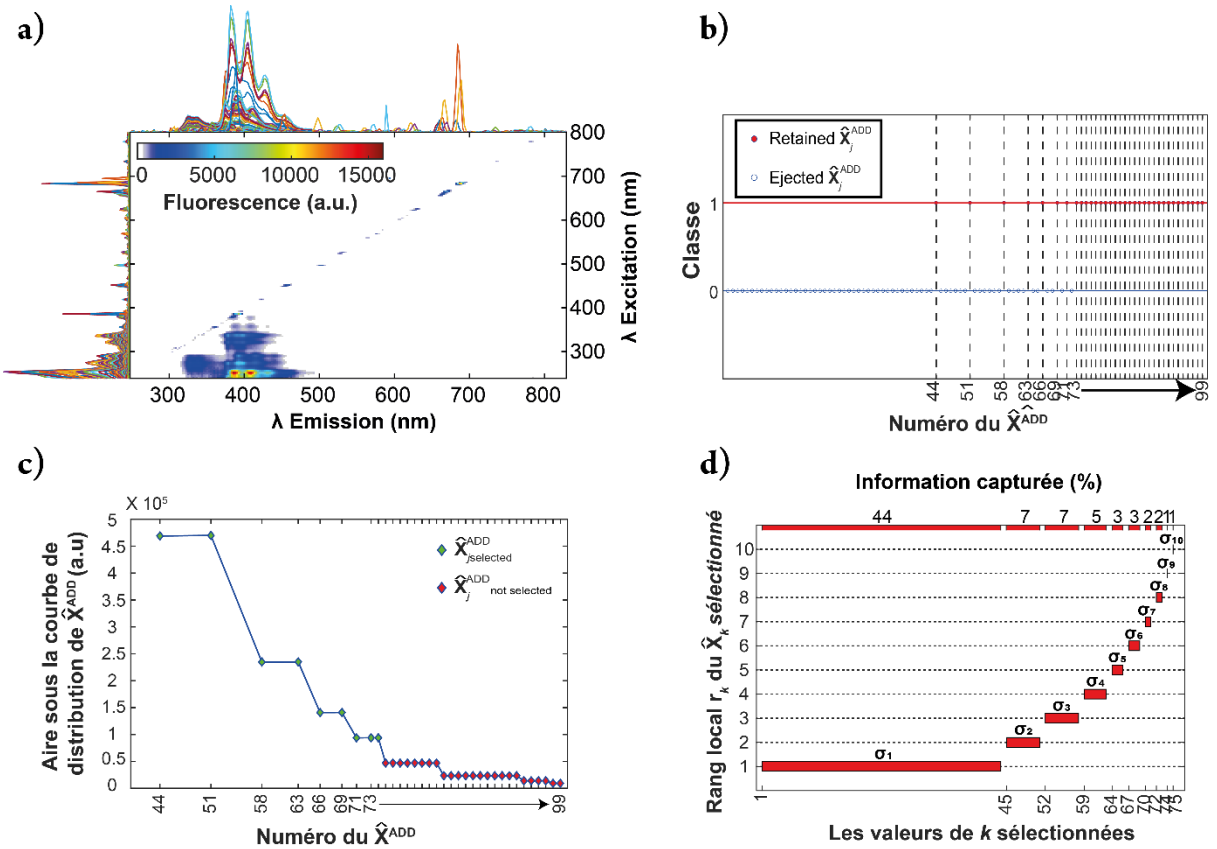


Figure III.9 : MEEF et spectres de fluorescence de l'échantillon 11 de l'ensemble de données #2 : données de départ avec contrainte de non négativité (a) ; Identification des matrices $\hat{X}_j^{ADD}$ nulles qui symbolisent une absence d'information ajoutée entre deux $\hat{X}_k$ successifs. Quand $\hat{X}_j^{ADD} \neq 0$ alors la matrice est en classe 1. Par contre, si $\hat{X}_j^{ADD} = 0$, la matrice est en classe 0 (b) ; Représentation des aires sous les courbes de distribution des pixels des $\hat{X}_j^{ADD}$ associés (c) ; les valeurs k sélectionnées en fonction de leurs rangs locaux r_k. Chaque rang local seul est représenté par son σ_i et le pourcentage d'informations qu'il capture (d).

Une fois les calculs de l'étape #2 effectués et que les deux premières étapes intermédiaires de l'analyse d'images réalisées, l'objectif est alors de comprendre le type d'informations contenues à chacune des k valeurs sélectionnées. Pour cela, les différentes cartes $\hat{\mathbf{X}}_k$, $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{selected}}}$ et $\hat{\mathbf{X}}^{\mathrm{residual}}_k$ sont tracées (Figure III.10). Une analyse d'image minutieuse du point de vue des longueurs d'onde peut permettre de sélectionner les matrices $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{accepted}}}$ qui correspondent à l'ensemble des σ_i de la factorisation SVD (étape #2) ne contenant que les informations chimiques de fluorescence pertinentes. Pour faciliter l'observation des informations chimiques de fluorescence, l'algorithme de segmentation des images en régions est utilisé pour extraire les bords extérieurs des $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{selected}}}$ afin de les superposer aux matrices $\hat{\mathbf{X}}_k$ correspondantes. La position du signal ajouté sur chaque carte est étudiée pour ressortir sa particularité. On observe ainsi, que pour $\sigma_2, \sigma_4, \sigma_8$, la majeure partie du signal ajouté est éloignée de la région de diffusion Rayleigh. Concernant σ_1, l'information qu'il contient est clairement un signal de fluorescence. Par conséquent, seuls $\sigma_1, \sigma_2, \sigma_4, \sigma_8$ possèdent des informations pertinentes sur les composés chimiques fluorescents. En effet, les $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{accepted}}}$ avec

des contours violets sur la Figure III.10 montre l'addition d'informations chimiques entre les rangs locaux. Le signal ajouté est toujours situé en bas à gauche des matrices $\widehat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{accepted}}}$ contrairement aux autres, pour lesquelles les signaux ajoutés sont plus fins et/ou dispersés, se trouvant tantôt sur la diagonale (effets de diffusion, par exemple $\widehat{\mathbf{X}}^{\mathrm{ADD}}_{51}$) ou tantôt ailleurs sur la carte (artefacts ou effets de bruit, par exemple $\widehat{\mathbf{X}}^{\mathrm{ADD}}_{74}$). Le résidu $\widehat{\mathbf{X}}^{\mathrm{residual}}_{72}$ confirme également que toute l'information chimique de fluorescence a été considérée avec le choix de l'ensemble $\{\sigma_1, \sigma_2, \sigma_4, \sigma_8\}$. En effet, la matrice résiduelle $\widehat{\mathbf{X}}^{\mathrm{residual}}_{72}$ ne présente que l'effet de diffusion Rayleigh répartis aléatoirement sur la diagonale puisque la dernière information chimique est ajoutée avec $\widehat{\mathbf{X}}^{\mathrm{ADD}}_{71}$ (motif en bas à gauche) et qui est donc présente dans la carte $\widehat{\mathbf{X}}_{72}$. Ce n'est plus vrai pour $k = \{74{,}75\}$.

En conséquence, l'ensemble optimal de valeurs singulières est $\Sigma = \{\sigma_1, \sigma_2, \sigma_4, \sigma_8\}$. Il reflète, de par la discontinuité dans les numéros des valeurs singulières sélectionnées, des déficiences de rang matriciel dues aux interférences. Le rang global déduit est alors égal à 4 et correspond à un rang global « idéal » puisqu'il reflète le nombre exact de HAP dans l'échantillon de départ. Avec une SVD classique, il est impossible d'observer cela. En effet, les valeurs singulières sont triées par ordre décroissant avec les fréquences cumulées et dépendent du rapport signal sur bruit. La MT-SVD trouve σ_8 qui capture autour de 2% de l'information totale (Figure III.9d). De par sa faible contribution au signal total, σ_8 peut être facilement confondu avec du bruit de mesure sur une SVD classique. Le risque est alors soit (i) de surestimer le rang de la matrice et donc d'extraire par des méthodes multivariées des composantes non représentatives de la réalité chimique soit (ii) ou au contraire de sous-estimer le rang et donc d'obtenir une caractérisation biaisée de la réalité chimique de l'échantillon au départ. La visualisation et l'analyse des informations avec la MT-SVD portées par chacune des σ_i permet de repousser les déficiences de rang avec, en fin de processus, une reconstruction des données de départ moins biaisées. Pour résumer, le pourcentage de l'information chimique portée par le rang global trouvé par la MT-SVD représente 58% de l'information contenue dans la MEEF de l'échantillon 11 de l'ensemble de données #2 (soit pour les rang locaux, σ_1 avec 44%, σ_2 avec 7%, σ_4 avec 5% et σ_8 avec 2%, Figure III.9d).

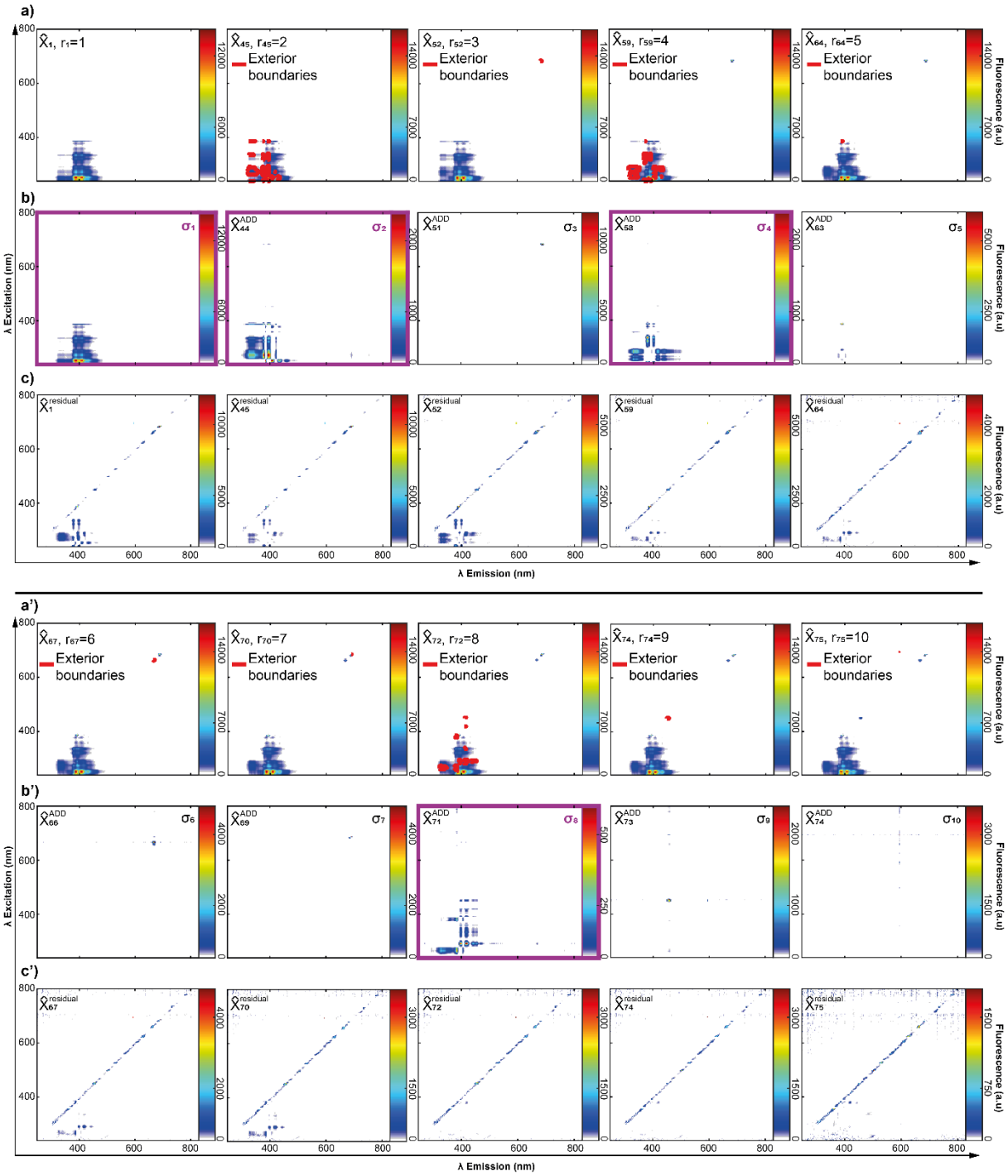


Figure III.10 : Analyse d'image de l'étape #2 de la MT-SVD : cartes $\widehat{\boldsymbol{X}}_k$(a-a') ; cartes $\widehat{\boldsymbol{X}}_{j_{selected}}^{ADD}$(b-b') et cartes $\widehat{\boldsymbol{X}}_k^{residual}$(c-c') respectivement, pour $k = \{1,45,52,59,64,67,70,72,74,75\}$ et $j_{selected} = \{44,51,58,63,66,69,71,73,74\}$ pour étudier les signaux de fluorescence. De plus, les limites extérieures trouvées par la MT-SVD avec les cartes $\widehat{\boldsymbol{X}}_{j_{selected}}^{ADD}$ sont tracées en rouge sur chaque carte $\widehat{\boldsymbol{X}}_k$. Les $\widehat{\boldsymbol{X}}_{j_{selected}}^{ADD}$ avec des contours violets correspondent aux $\widehat{\boldsymbol{X}}_{j_{accepted}}^{ADD}$ et montrent l'addition d'information chimique entre les rangs locaux.

La reconstruction de la matrice multi-tronquée notée $\widehat{\mathbf{X}}_{\text{truncated}}$ à l'étape #3 est effectuée avec l'ensemble « optimal » de σ_i noté Σ et les vecteurs singuliers correspondants. D'un point de vue de l'image ou d'un point de vue spectral, l'information chimique est conservée intacte alors que les signaux de diffusion ont été supprimés (Figure III.11). De plus,

l'algorithme de segmentation en régions des images, utilisé pour un recadrage automatique de $\widehat{\mathbf{X}}_{\text{truncated}}$ montre de bonnes performances. En effet, les longueurs d'onde d'excitation et d'émission maximales sélectionnées automatiquement sont respectivement, de 422 nm et de 511 nm (Figure III.11). Cette sélection automatique de région permet une réduction de la taille des données et donc une réduction du temps de traitement ultérieure, avec par exemple, un algorithme de décomposition spectrale.

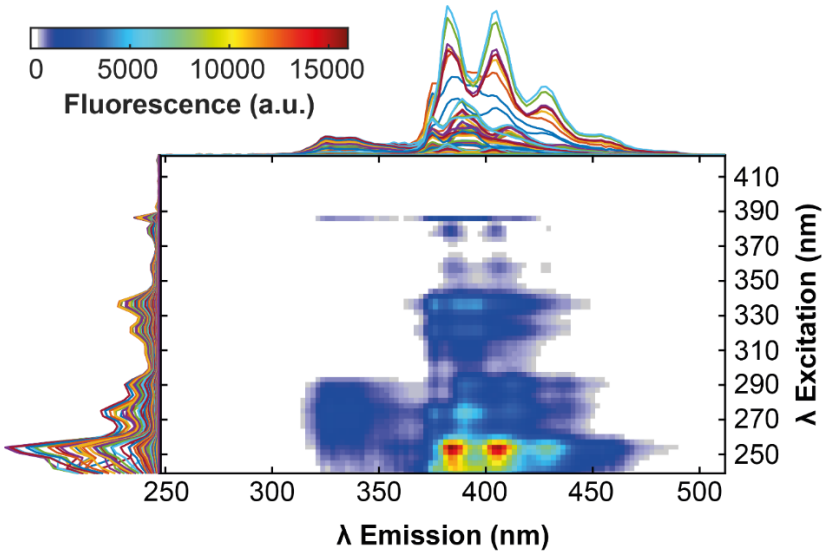


Figure III.11 : Le résultat du prétraitement MT-SVD de la MEEF de l'échantillon 11 de l'ensemble de données #2. Les spectres d'émission sont placés au-dessus de la carte et les profils d'excitation à sa gauche.

III.3.2.4 Cas 4 : Prétraitement MT-SVD de la MEEF d'un mélange de quatre espèce chimique pures, acquise à des concentrations relativement faibles et à laquelle, du bruit blanc simulé est rajouté

Une simulation de bruit blanc à haut niveau (moyenne = 0 et amplitude = 500) a été réalisée et ajoutée à la MEEF de départ de l'échantillon 11 de l'ensemble de données #2 (Figure III.12a). En suivant le même processus décrit dans le cas 3, le nombre de σ_i pertinents trouvés sans a priori par la MT-SVD est égal au nombre de HAP présents dans les mélanges de départ (i.e. quatre), avec $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_5\}$. Malgré l'ajout de bruit blanc, la déficience du rang matriciel a été corrigée et les résultats obtenus sont donc satisfaisants (Figure III.12b).

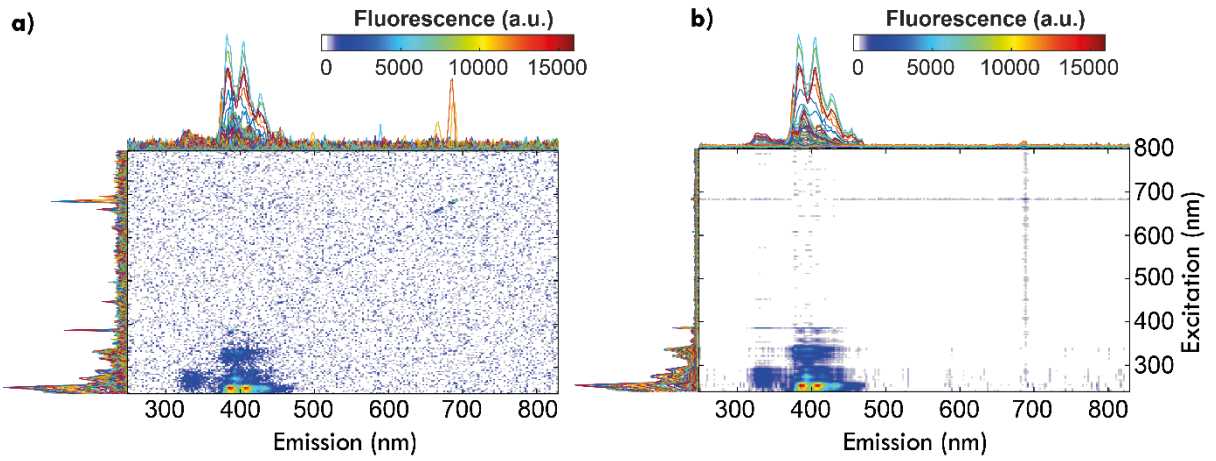


Figure III.12 : MEEF de l'échantillon 11 de l'ensemble de données #2 : données brutes auxquelles est ajoutées du bruit blanc à haut niveau et sur lesquelles une contrainte de non négativité est appliquée (a). Résultat après la correction par MT-SVD (b).

III.3.2.5 Cas 5 : Prétraitement MT-SVD après augmentation matriciels de plusieurs MEEF et décomposition PARAFAC

Dans cet exemple, les 11 matrices de l'ensemble de données #2 sont prétraitées ensemble par MT-SVD en utilisant l'augmentation matricielle en colonnes (i.e. dimension émission augmentée), puis elles sont décomposées par PARAFAC afin d'extraire les signatures pures de chaque espèce chimique constituant les mélanges et leurs quantités relatives. L'augmentation matricielle est une opération plus flexible que la construction d'un cube de données car elle permet l'analyse simultanée de matrices de données qui n'ont pas nécessairement la même taille dans toutes les directions. Aussi, elle ne nécessite pas que les profils obtenus dans la direction augmentée soient identiques en forme et/ou en nature chimique. L'accumulation de données augmente non seulement la quantité d'informations utilisées, mais peut conduire également à un gain qualitatif[62]. L'objectif ici est de combiner les avantages de la MT-SVD, discutés dans les cas précédents avec l'augmentation matricielle[123] afin d'avoir la meilleure caractérisation possible lors d'une décomposition PARAFAC non supervisée (i.e. sans utilisation de MEEF de références dans le calcul du modèle) de ces mélanges relativement complexes. La plupart du temps, PARAFAC est utilisé en combinant les matrices de mélanges et celles de références. La Figure III.13 illustre les résultats de l'augmentation matricielle avant et après MT-SVD. Dans ce cas, aucune déficience de rang n'est trouvée et la MT-SVD montre que $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$ contenait l'information chimique non biaisée.

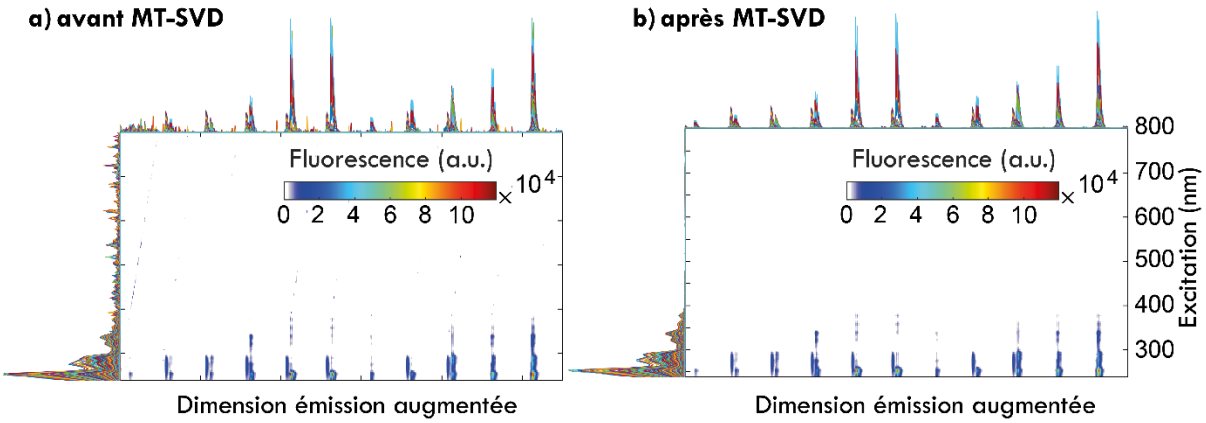


Figure III.13 : Matrice augmentée des 11 échantillons de l'ensemble de données #2 (a). Résultats MT-SVD sur la matrice augmentée (b).

Une fois que la matrice augmentée est prétraitée et repliée sous forme 3D grâce à une opération de redimensionnement (i.e. $188{,}250{,}11$), l'algorithme de décomposition PARAFAC est ensuite employé. Les valeurs des différents critères (cf. §II.4.2.2) utilisés pour choisir un modèle PARAFAC confirment la première estimation de la valeur du rang global par MT-SVD (Tableau III.2). Compte tenu de tous ces indicateurs et après analyse visuelle des matrices résiduelles, le modèle à quatre composantes est toujours choisi comme modèle valide pour PARAFAC, ce qui correspond à notre connaissance a priori des échantillons chimiques complexes idéaux formés expérimentalement, mais également à l'estimation de l'algorithme MT-SVD, en l'occurrence 4 HAP pour une estimation du rang matriciel égale à 4.

Tableau III.2 : Les résultats des différents critères utilisés pour choisir le modèle PARAFAC valide. CP : composante pure.

Nombre de CP du modèle PARAFAC	**Variance de la dernière composante du modèle (%)**	**Core Consistency (%)**	**Mesure de la similarité entre le modèle de chaque sous-ensemble et le modèle global (%) | 3 mesures**		
1	84.18	100.00	88.20	88.20	88.20
2	10.91	100.00	48.50	48.60	48.50
3	9.21	99.00	68.10	68.20	68.10
4	3.17	100.00	99.00	99.00	99.00
5	0.05	<0.00	0.00	0.00	0.00
6	0.06	<0.00	0.00	0.00	0.00

Les résultats de la modélisation PARAFAC de l'ensemble de données #2, avec 4 composantes pures, sont présentés en termes de '*loadings*' (i.e. les profils purs reconstruits à partir des profils d'émission et d'excitation estimés, Figure III.14). Aucune contrainte n'a été appliquée puisque le modèle est stable et interprétable selon les critères du Tableau III.2. En effet, un modèle PARAFAC est une solution mathématiquement unique et ne nécessite pas systématiquement l'application de contraintes pour obtenir une solution chimiquement valide (cf. §II.4.2.2). Les résultats de la décomposition PARAFAC sont satisfaisants du point de vue qualitatif grâce à la comparaison avec les références (i.e. MEEF de l'ensemble de données

#1). De plus, la validation croisée '*split half analysis*' est effectuée trois fois et la mesure de similarité des '*loadings*' résultants (i.e. ceux du modèle global et ceux de deux modèles de deux moitiés indépendantes) est calculée par une corrélation avec 99,00% de similarité pour les trois fois. Le résultat du critère de validation '*Core Consistency*' égal à 100.00% pour le modèle à quatre composantes et <0% pour les modèles à cinq et six composantes confirme que le modèle à quatre composantes est celui qui est susceptible de se rapprocher le plus de la réalité chimique. En outre, la variance de la dernière composante de chaque modèle (en %) nous permet de voir quelles composantes contribuent de manière significative à la décomposition des données prétraitées. C'est le cas de la quatrième composante du modèle PARAFAC puisque sa contribution est de 3.17%. Cette contribution chute autour de 0% pour la cinquième composante du modèle PARAFAC. Pour cette modélisation, les échantillons de référence de l'ensemble de données #1 ne sont pas utilisés dans la construction du modèle et un recouvrement spectral est observé sur les données de départ (ensemble de données #2), notamment entre ANT, BaA et PYR, ce qui aurait pu perturber la modélisation et le choix du rang global. D'autant plus que ces trois composés chimiques émettent à des longueurs d'onde d'émission très proches (i.e. entre 370 et 470 nm) et seules les longueurs d'onde d'excitation permettent leur distinction.

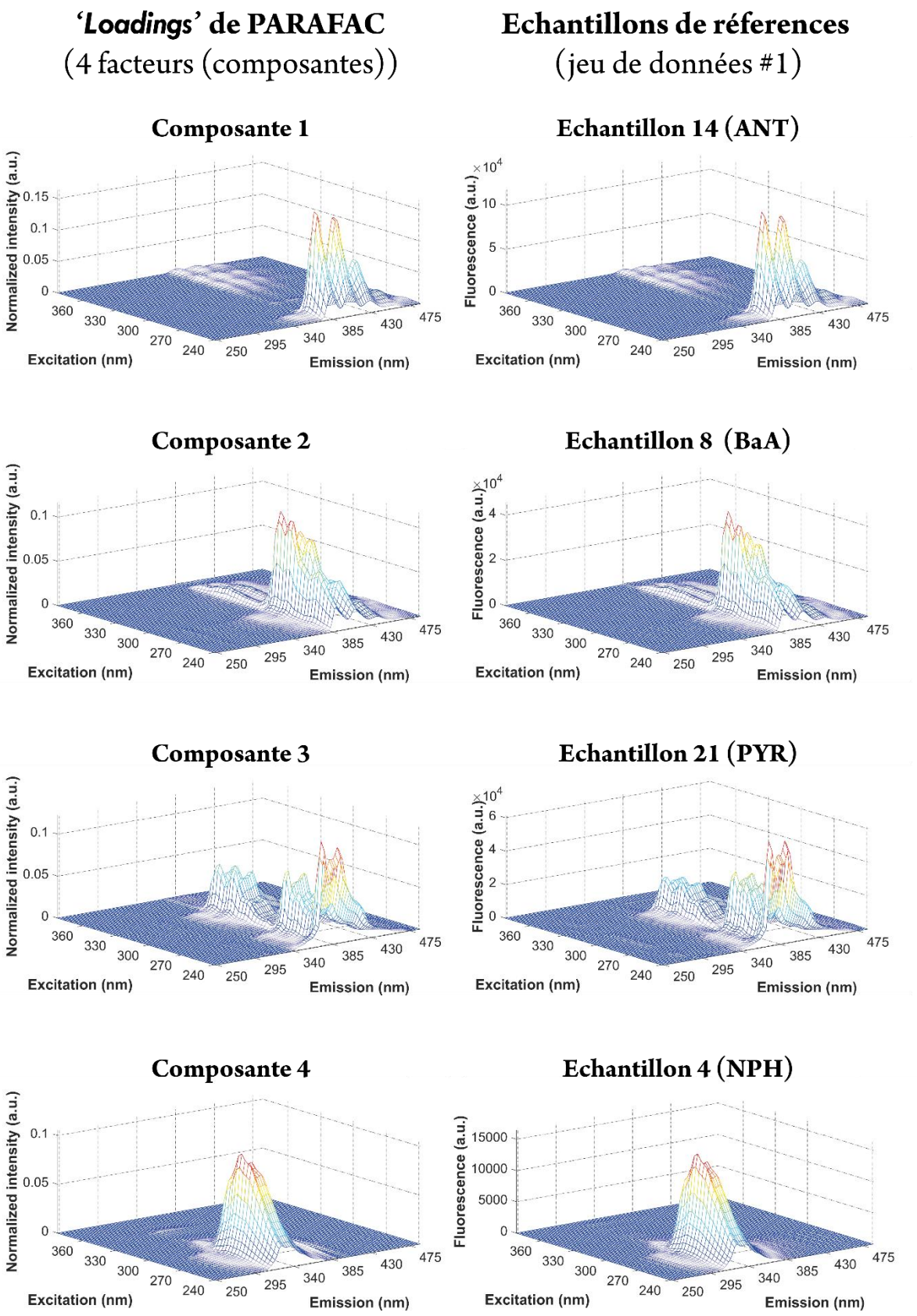


Figure III.14 : Résultats de la décomposition PARAFAC en termes de 'loadings', avec 4 composantes, de l'ensemble de données #2.

Par ailleurs, un ajustement linéaire avec une ordonnée à l'origine égale à zéro a été établi entre les contributions factorielles (i.e. '*scores*') des quatre composantes et les concentrations réelles des HAP correspondants dans les mélanges (Figure III.15). Les résultats obtenus sont relativement satisfaisants, car les valeurs des coefficients de détermination (i.e. R^2) observées se situent entre 0.97% et 0.99%. De plus, la sensibilité du modèle est correcte malgré le recouvrement spectral et les faibles concentrations de fluorophores dans certains échantillons (e.g. concentration du BaA dans l'échantillon 11 égale à 0.005 mg.L^{-1}). Cela traduit la capacité de la MT-SVD à conserver intacte qualitativement et quantitativement l'information chimique contenue dans les données brutes.

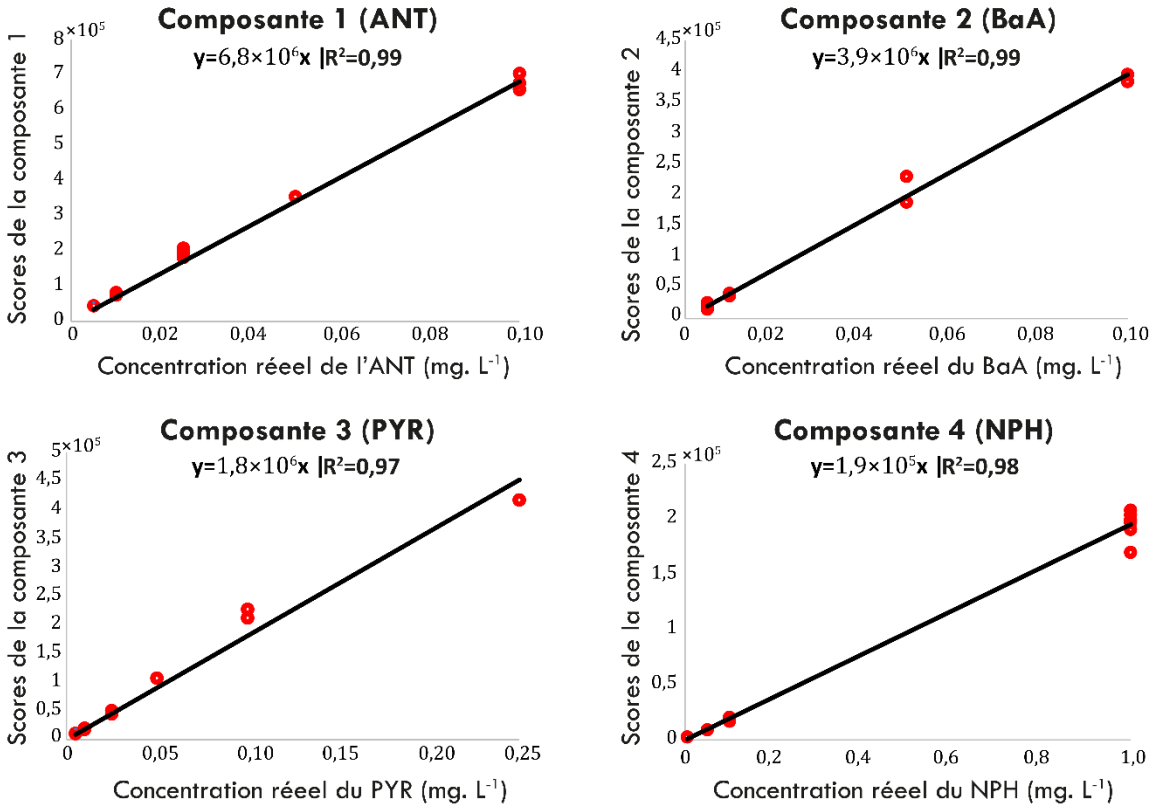


Figure III.15 : Ajustement linéaire avec une ordonnée à l'origine égale à zéro entre les 'scores' des quatre composantes et les concentrations réelles des HAP correspondants dans les mélanges.

Cette section du chapitre nous a permis d'illustrer les différentes étapes de l'algorithme MT-SVD ainsi que son intérêt en prétraitement, à travers son application à des MEEF de quatre HAP purs « modèles ». Le prochain et dernier cas abordera une analyse MT-SVD d'une MEEF complexe d'un échantillon de goudron de houille, un sous-produit de la cokéfaction.

III.4 Mise en pratique de la MT-SVD via l'analyse d'un échantillon réel. Cas 6 : Prétraitement MT-SVD de la MEEF d'un goudron de houille

Pour ce cas, l'échantillon de goudron de houille sélectionné provient du centre de pyrolyse de Marienau situé dans la région du Grand-Est en France. Il a été préparé et soumis à une analyse par fluorescence 3D, en utilisant le même solvant (i.e. dichlorométhane) et les mêmes conditions expérimentales que celles utilisées pour la construction de la base de données des HAP modèles (cf. §III.3.1). Une analyse par spectroscopie UV-Visible a également été réalisée simultanément. Les spectres d'absorption ont été acquis sur la même plage spectrale que les spectres d'excitation (i.e. longueurs d'onde allant de 239 à 800 nm avec un pas de 3 nm). Pour faciliter le prélèvement, l'échantillon a d'abord été soumis à une sonication pendant 30 minutes afin de réduire sa viscosité. Ensuite, un prélèvement a été effectué et des solutions de différentes concentrations (de 0.05 à 300 $mg.L^{-1}$) ont été préparées dans du dichlorométhane avant d'être analysées.

La MEEF de la solution présentant une concentration de 0.6 $mg.L^{-1}$ a été retenue, car elle présentait le meilleur rapport signal sur bruit. Au-delà de cette concentration, on observe une saturation du signal de fluorescence. De plus, son spectre d'absorbance présente une absorbance maximale de 0.063 (u.a) située à la longueur d'onde de 239 nm (Figure III.16). Cette valeur n'est pas très éloignée de la valeur recommandée dans la littérature pour respecter les conditions de linéarité (i.e. une absorbance de 0.05 u.a).

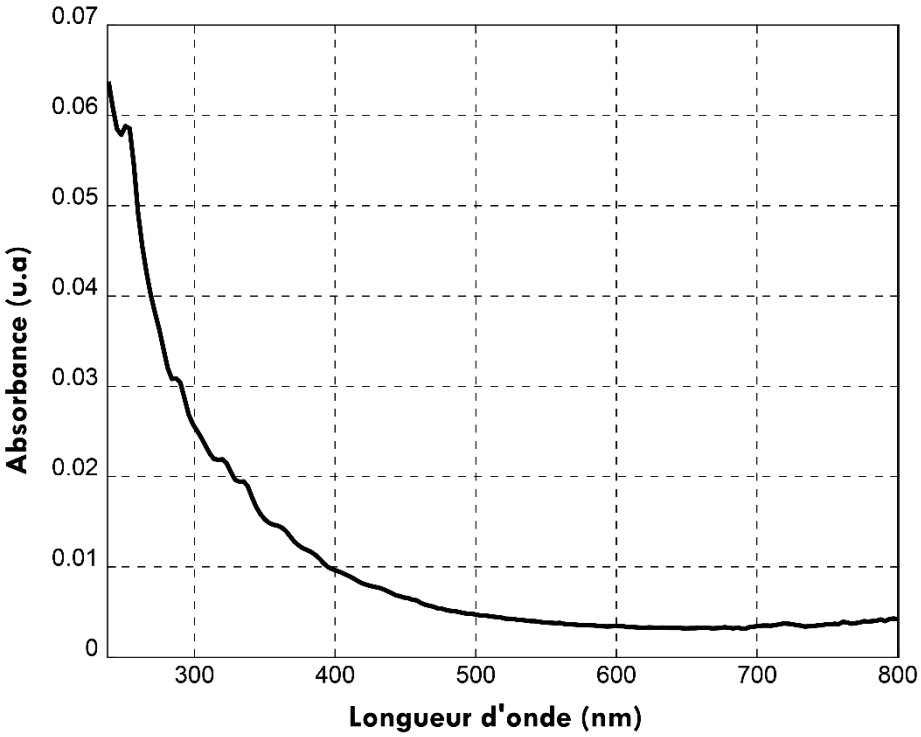


Figure III.16 : Spectre d'absorbance de la solution de goudron de houilles à 0.6 $mg.L^{-1}$.

L'algorithme MT-SVD, dans le cadre de l'analyse de la MEEF de ce goudron de houille, fournit, en plus du débruitage, une estimation du rang global qui est égale à 8 avec $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7, \sigma_8\}$ (Figure III.17).

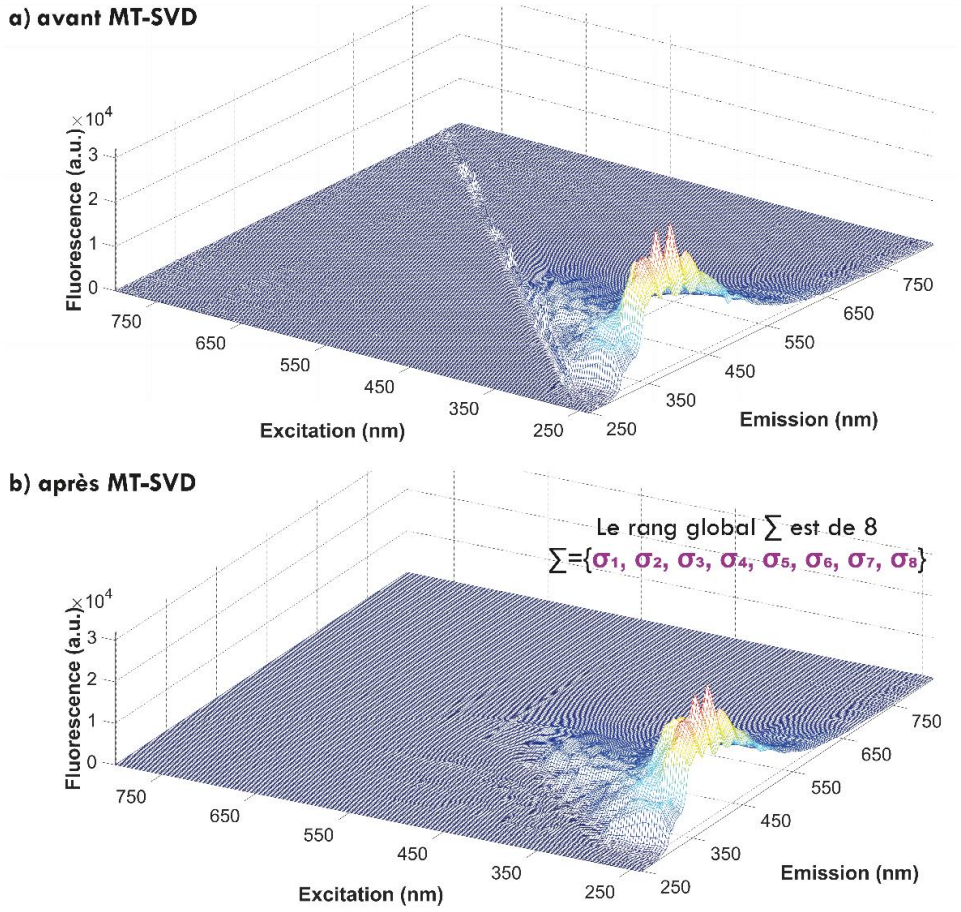


Figure III.17 : MEEF du goudron de houille avant (a) et après prétraitement MT-SVD (b).

Il est indéniable que le goudron de houille n'est pas composé uniquement de 8 CAP. Contrairement au cas 5 où chaque rang local correspond à une espèce chimique spécifique (i.e. HAP), dans le cas des échantillons réels, très complexe au niveau de leur chimie, chaque rang local ne correspond pas nécessairement à une seule espèce chimique. Cela est dû aux interactions chimiques complexes entre les différents composés et aux phénomènes physico-chimiques qui affectent le signal de fluorescence. Ces interactions et ces phénomènes entraînent des problèmes de sélectivité et de recouvrement spectral, ce qui se traduit par des colinéarités entre les signaux provenant des différentes espèces et rend difficile la distinction entre elles. Ainsi, l'estimation du rang mathématique de la MEEF peut être faible. Elle ne reflète pas nécessairement le nombre d'espèces chimiques contenues dans un mélange complexe, mais plutôt et souvent les combinaisons entre celles-ci. Pour obtenir une estimation aussi précise que possible de la composition en CAP dans un système complexe tel que le goudron de houille, et en cherchant à repousser les limites de la décomposition trilinéaire classique des

données, il est nécessaire d'adopter une approche innovante qui combine les connaissances sur le système étudié a priori, le prétraitement MT-SVD et la décomposition PARAFAC. Les détails de ce travail méthodologique et les résultats obtenus seront discutés dans le prochain chapitre.

III.5 Conclusion

L'algorithme MT-SVD décrit dans ce chapitre représente une approche nouvelle et originale pour la gestion des effets de diffusion de la lumière, du bruit et des phénomènes pouvant être observé de déficience du rang matriciel dans les MEEF. Ce prétraitement de chimiométrie est basé sur la décomposition en valeurs singulières, qui est l'un des algorithmes les plus couramment utilisés en algèbre linéaire, avec une valeur ajoutée. En effet, la MT-SVD extrait les informations chimiques les plus pertinentes en déduisant un rang global optimal à partir du calcul de rangs locaux. Ces rangs locaux sont déterminés en utilisant un pourcentage de fréquence seuillé et couplé à une analyse d'image.

Les échantillons « modèles » étudiés sont représentatifs des effets de diffusion de la lumière que l'on trouve généralement en fluorescence. Ces effets ont été éliminés des matrices de départ, et l'information chimique pertinente a été conservée intacte. De plus, l'ajout d'un bruit blanc important a eu une faible influence sur la capacité de l'algorithme à filtrer les MEEF. Au-delà de cela, la MT-SVD a permis de visualiser et de surmonter les déficiences du rang matriciel, en particulier lorsqu'il y a un problème de sélectivité spectrale. À la fin du prétraitement, la nouvelle matrice de données est prête à être analysée par des méthodes de chimiométrie (e.g. décomposition bilinéaire ou trilinéaire) pour la séparation de sources. Par ailleurs, la MT-SVD est une méthode flexible pouvant être appliquée à tout type de MEEF mais surtout, à d'autres techniques d'imagerie (e.g. imagerie hyperspectrale Raman). Cette approche présente donc de nombreux avantages.

Enfin, il a été observé, dans ce chapitre, que les interactions chimiques complexes qui se produisent entre les composés d'un échantillon complexe tel que le goudron de houille et les phénomènes physico-chimiques qui affectent le signal de fluorescence conduisent à une sous-estimation du rang des MEEF, même lorsque les conditions de linéarité sont respectées et que des prétraitements avancés tels que la MT-SVD sont appliqués aux MEEF. Cela limite l'efficacité des méthodes de décomposition spectrale dans leur approche classique et nécessite une prudence lors de l'interprétation des '*loadings*' générés. Par conséquent, il a été essentiel, dans ce travail de thèse, de développer une stratégie visant à surmonter, du moins partiellement, cette difficulté. Cette approche sera présentée en détail dans le prochain chapitre.

CHAPITRE IV

CARACTERISATION QUALITATIVE DES HYDROCARBURES AROMATIQUES POLYCYCLIQUES PRESENTS DANS DES SOLS INDUSTRIELS CONTAMINES PAR DECOMPOSITION PARAFAC DE MEEF D'EXTRAITS ORGANIQUES

IV.1 Introduction

Dans ce chapitre de thèse, une méthode d'identification de 16 HAP est proposée pour caractériser des sols industriels anthropisés. Ces 16 HAP ciblés dans cette étude sont ceux inclus dans la liste des polluants prioritaires établie par l'US-EPA[4]. Cependant, l'acénaphtylène, bien qu'il soit répertorié dans la liste, n'a pas été inclus dans le modèle en raison de son rendement de fluorescence très faible. Il a finalement été remplacé par un autre HAP, qui est le pérylène (Tableau IV.1). Ce dernier est un marqueur d'intérêt de l'altération biologique de la matière organique en conditions anoxiques[124,125].

Tableau IV.1 : Liste des 16 HAP ciblés dans l'étude.

	Nom du HAP	**Abréviation**
1	Naphtalène	NPH
2	Anthracène	ANT
3	Benz[α]anthracène	BaA
4	Pyrène	PYR
5	Fluoranthène	FA
6	Phénanthrène	PHE
7	Fluorène	FLR
8	Benz[α]pyrène	BaP
9	Benz(*g,h,i*)perylène	BghiP
10	Acénaphtène	AC
11	Benz[*b*]fluoranthène	BbF
12	Dibenz(*a,h*)anthracène	DahA
13	Benz[*k*]fluoranthène	BkF
14	Indénopyrène	IP
15	Chrysène	CHR
16	Pérylène	PER

La méthodologie proposée d'identification de ces 16 HAP cibles s'appuie sur un modèle PARAFAC. Dans un premier temps, ce modèle est calibré, optimisé et validé à partir d'échantillons de référence analysés en fluorescence 3D. Dans un deuxième temps, il est appliqué à différents extraits organiques de sols industriels complexes pour caractériser la nature de cette pollution.

Habituellement, un modèle PARAFAC est construit en décomposant les MEEF des échantillons complexes, puis les résultats sont interprétés sous forme de '*loadings*' et de '*scores*' pour tenter de déduire les espèces pures (i.e. espèces chimiques seules ou mélanges d'espèces chimiques) qui les composent ainsi que leurs proportions respectives. Cependant, dans ce travail, une approche légèrement différente a été adoptée. Le modèle est construit en décomposant le cube des MEEF des espèces de référence (i.e. 16 HAP), qui sont les espèces

ciblées que nous souhaitons détecter (Figure IV.1a). Une fois le modèle construit et validé, les MEEF des échantillons complexes sont projetées dans ce modèle préexistant. Cela permet d'estimer la présence ou l'absence des espèces de référence (espèces que nous ciblons) dans ces échantillons complexes en analysant la matrice des '*scores*' prédits. En effet, avec cette approche, lors de la projection, les deux matrices de '*loadings*' obtenues lors de la calibration du modèle, qui sont liées à l'identité des 16 HAP, restent inchangées. En revanche, les '*scores*' liés à la proportion de ces 16 HAP dans les extraits de sols varient, permettant ainsi d'estimer leur présence ou leur absence dans ces derniers (Figure IV.1b). Il est important de souligner que cette approche restreint la recherche des espèces pures uniquement à celles qui ont été utilisées pour la construction du modèle.

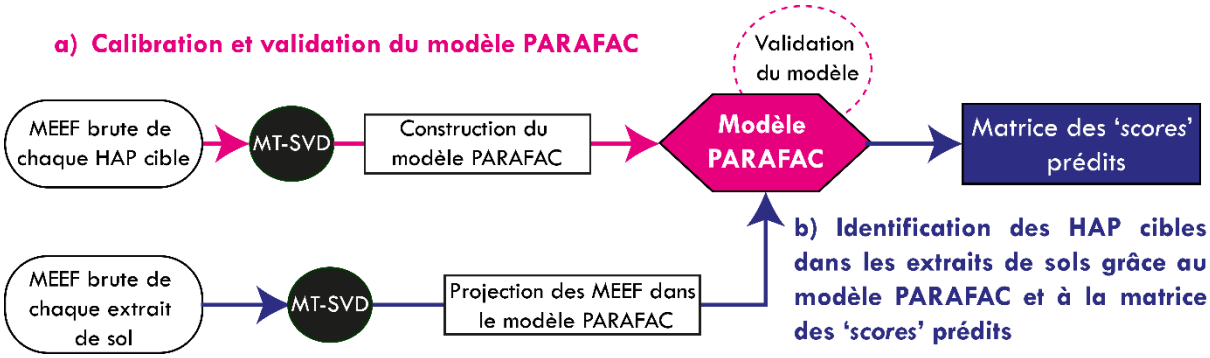


Figure IV.1 : Calibration et validation du modèle PARAFAC (a). Caractérisation qualitative des HAP cibles dans les extraits de sols grâce au modèle PARAFAC (b).

Les deux premières sections de ce chapitre décrivent les différentes étapes de construction et de validation du modèle PARAFAC. La dernière section présente les résultats de l'application de ce modèle PARAFAC à l'analyse des MEEF des extraits de sols pollués par les HAP.

IV.2 Calibration du modèle PARAFAC

IV.2.1 Base de données de l'ensemble de calibration

Pour la construction de la base de données des 16 HAP (i.e. ensemble de calibration du modèle PARAFAC), la même instrumentation que celle décrite dans le chapitre III (cf. §III.3.1) a été utilisée, à la différence que les échantillons ont été excités à l'aide d'une plage de longueurs d'onde d'excitation plus restreinte. Cette plage est comprise entre 239 et 500 nm avec un pas de 3 nm. L'émission de fluorescence est quant à elle collectée dans la même plage de longueurs d'onde allant de 248.27 à 829.32 nm avec une résolution de 2.33 nm. Ainsi, toutes les MEEF présentent une taille de 88 × 250 pixels, c'est-à-dire 88 spectres d'émission et 250 spectres d'excitation.

La MEEF optimale pour chaque molécule a été recherchée et identifiée. Elle dépend à la fois de la concentration dans le dichlorométhane et du rendement de fluorescence de chaque molécule. En d'autres termes, il s'agit d'identifier pour chaque HAP, la MEEF qui présente le meilleur rapport signal sur bruit. Par exemple, la MEEF du naphtalène (NPH) a été acquise à la plus haute concentration de l'étude, qui est de 1 $mg.L^{-1}$, en raison du rendement de fluorescence faible de cette molécule. En revanche, pour un composé présentant un rendement de fluorescence important comme le pérylène (PER), la concentration la plus faible de 0.007 $mg.L^{-1}$ a été utilisée (Tableau IV.2).

En outre, le spectre d'absorbance de chaque espèce a été mesuré sur la même plage spectrale que les spectres d'excitation afin de vérifier que l'absorbance de chaque espèce à la concentration choisie respecte bien la condition de linéarité de la loi de Beer-Lambert. Pour rappel, dans la littérature, il est recommandé d'avoir une absorbance qui n'excède pas la valeur de 0.05 u.a afin d'obtenir un écart de la linéarité entre l'intensité de fluorescence et la concentration acceptable, autour des 5.5% (Figure I.6 du chapitre I). Ainsi, la majorité des absorbances mesurées sont inférieures à 0.05 u.a respectant ainsi cette condition. Par exemple, pour le phénanthrène (PHE) à 0.1 $mg.L^{-1}$, l'absorbance maximale a été observée à 251 nm avec une valeur de 0.033 u.a. Néanmoins, seules trois espèces présentent des absorbances supérieures à 0.05 u.a. L'anthracène (ANT) à 0.1 $mg.L^{-1}$ présente une absorbance maximale de 0.060 u.a à 254 nm. Le Benz[α]anthracène (BaA), également à 0.1 $mg.L^{-1}$, présente une absorbance maximale de 0.074 u.a entre 278 et 290 nm. Enfin, le chrysène (CHR) présente à la même concentration, une absorbance maximale de 0.068 u.a à 269 nm (Figure IV.2). Ainsi, pour ces espèces, l'écart de linéarité est estimé inférieur à 8%, ce qui reste relativement acceptable.

L'algorithme MT-SVD est utilisée, pour le prétraitement individuel des MEEF de références de nos 16 HAP et ainsi vérifier que le rang matriciel de chaque MEEF est bien égal à 1 (Figure IV.3). Enfin, chacune des MEEF est normalisée suivant sa norme L1.

Tableau IV.2 : Base de données des 16 HAP de l'étude qui sert d'ensemble de calibration du modèle PARAFAC. L'abréviation E signifie échantillon.

		NPH (mg.L^{-1})	ANT (mg.L^{-1})	BaA (mg.L^{-1})	PYR (mg.L^{-1})	FA (mg.L^{-1})	PHE (mg.L^{-1})	FLR (mg.L^{-1})	BaP (mg.L^{-1})	BghiP (mg.L^{-1})	AC (mg.L^{-1})	BbF (mg.L^{-1})	DBahA (mg.L^{-1})	BkF (mg.L^{-1})	IP (mg.L^{-1})	PER (mg.L^{-1})	CHR (mg.L^{-1})
Ensemble de calibration du modèle PARAFAC	E1	1	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	E2	/	0.100	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	E3	/	/	0.100	/	/	/	/	/	/	/	/	/	/	/	/	/
	E4	/	/	/	0.100	/	/	/	/	/	/	/	/	/	/	/	/
	E5	/	/	/	/	0.200	/	/	/	/	/	/	/	/	/	/	/
	E6	/	/	/	/	/	0.100	/	/	/	/	/	/	/	/	/	/
	E7	/	/	/	/	/	/	0.010	/	/	/	/	/	/	/	/	/
	E8	/	/	/	/	/	/	/	0.020	/	/	/	/	/	/	/	/
	E9	/	/	/	/	/	/	/	/	0.050	/	/	/	/	/	/	/
	E10	/	/	/	/	/	/	/	/	/	0.200	/	/	/	/	/	/
	E11	/	/	/	/	/	/	/	/	/	/	0.100	/	/	/	/	/
	E12	/	/	/	/	/	/	/	/	/	/	/	0.050	/	/	/	/
	E13	/	/	/	/	/	/	/	/	/	/	/	/	0.010	/	/	/
	E14	/	/	/	/	/	/	/	/	/	/	/	/	/	0.200	/	/
	E15	/	/	/	/	/	/	/	/	/	/	/	/	/	/	0.007	/
	E16	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	0.100

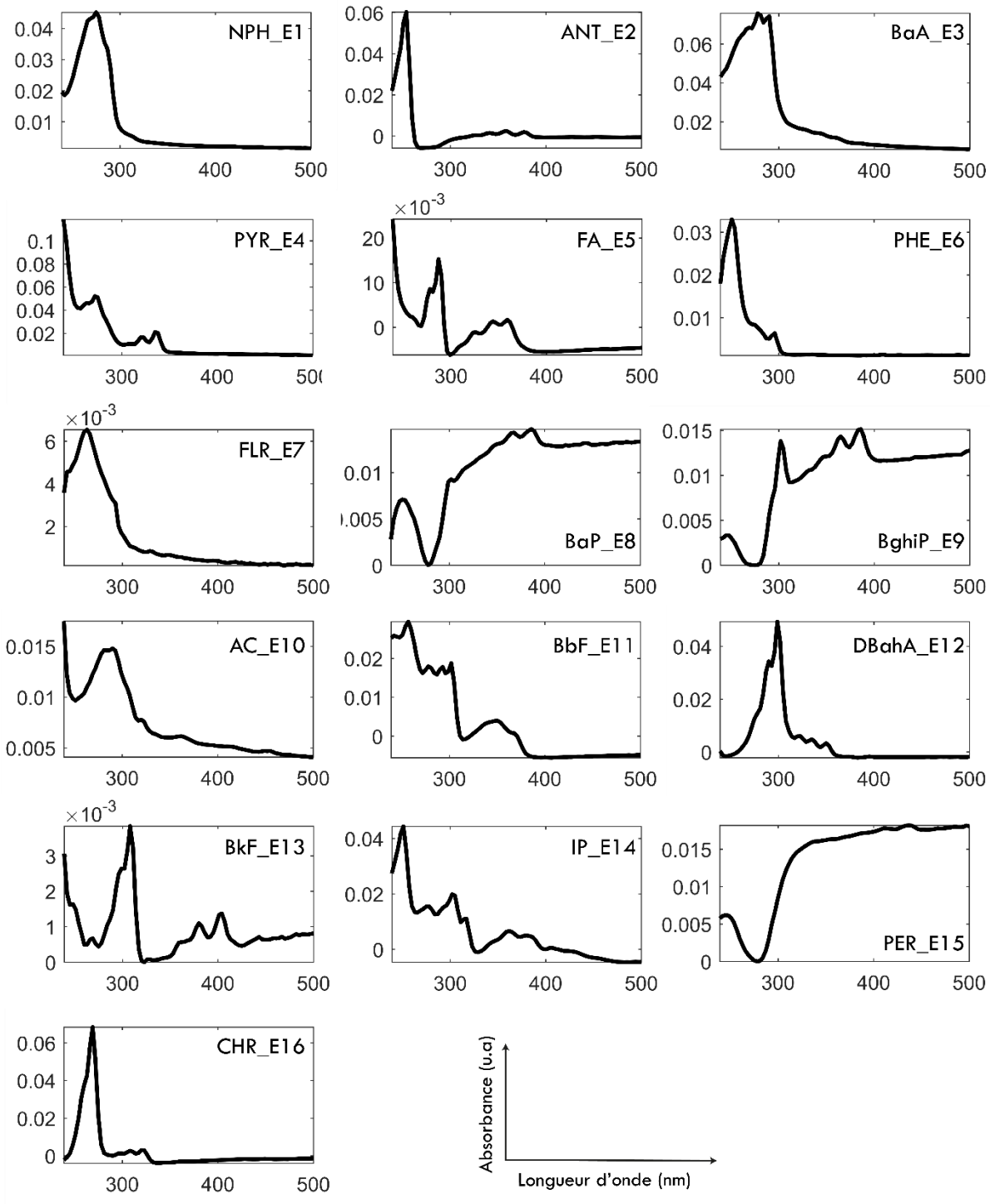


Figure IV.2 : Spectres d'absorbance des 16 échantillons (i.e. HAP) de la base des données. L'abréviation E signifie échantillon.

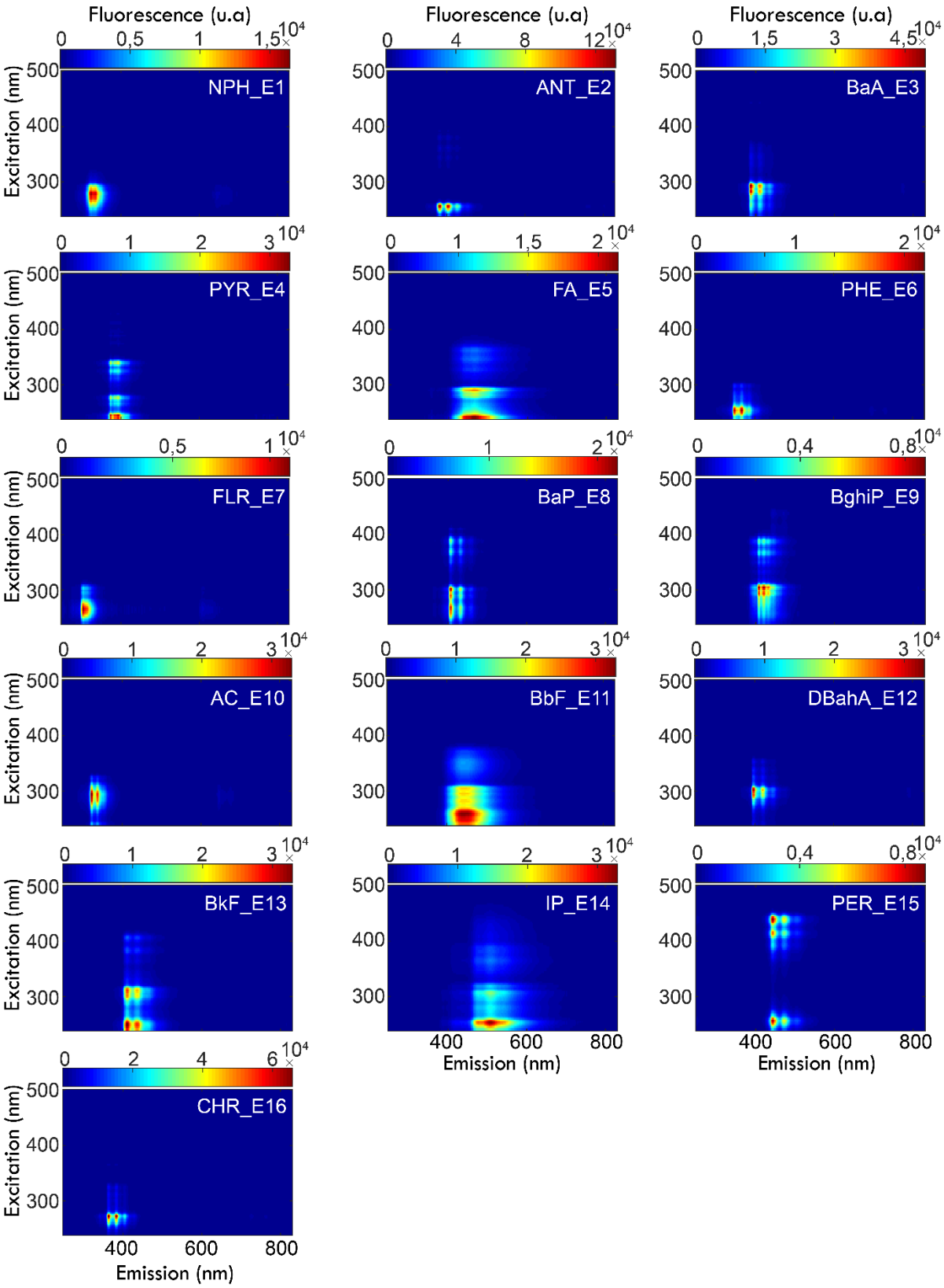


Figure IV.3 : MEEF des 16 échantillons HAP de la base de données après prétraitement par MT-SVD. L'abréviation E signifie échantillon.

IV.2.2 Construction du modèle PARAFAC

La construction du modèle PARAFAC s'effectue sur un cube de données de taille $(16,88,250)$ pour 16 HAP, 88 longueurs d'ondes d'excitation et 250 longueurs d'ondes d'émission. Une contrainte de non négativité est appliqué aux trois modes (i.e. '*scores*' et '*loadings*') pour améliorer la stabilité du modèle[92]. La décomposition TLD est utilisée lors de l'initialisation de l'algorithme. Elle est appliquée pour décomposer le cube des MEEF et estimer les matrices initiales $\mathbf{B}$ et $\mathbf{C}$. Ces matrices sont ensuite introduites dans le modèle PARAFAC pour l'ajustement des moindres carrés alternés lors de la minimisation de la somme des carrés des résidus tout en respectant la contrainte de non négativité. Plusieurs modèles ont été construits, comprenant un nombre croissant de composantes pures, allant de 1 à 20. Deux critères ont été utilisés comme indicateurs pour valider le choix approprié du nombre de composantes. La variance individuelle, c'est-à-dire celle qui est associée à la dernière composante de chaque modèle, a été examiné en parallèle du critère de '*core consistency'*. La variance individuelle indique le poids de la contribution de la dernière composante de chaque modèle. Plus sa valeur est élevée, plus la contribution au modèle de la dernière composante est significative. Le '*core consistency'* indique le degré de la cohérence mathématique de chaque modèle. La cohérence maximale est de 100%, indiquant une bonne adéquation mathématique du modèle. Des cohérences inférieures à 70% ou à 90% suggèrent soit une utilisation excessive de composantes, soit une instabilité du modèle. Une cohérence négative indique que le modèle est inadéquat mathématiquement.

Dans notre cas, la variance individuelle diminue progressivement jusqu'à atteindre des valeurs très faibles inférieures à 1% et ce à partir du modèle à 17 composantes. Cela indique une très faible contribution de chacune des dernières composantes des modèles allant de 17 à 20 composantes (Tableau IV.3). Le '*core consistency*' varie très légèrement du modèle à 1 composante jusqu'au modèle à 16 composantes, mais une chute brutale de cet indicateur est observée pour les modèles de 17 à 20 composantes. Cela indique que le modèle à 16 composantes est celui qui est le plus susceptible de représenter notre cube de données de départ construit à partir de connaissance a priori (Tableau IV.3). En outre, la convergence des modèles de 17 à 20 composantes est très lente, ce qui suggère une possible dégénérescence des solutions associées à ces modèles PARAFAC. Par conséquent, le nombre de composantes pures est fixé à 16 et correspond bien à notre connaissance a priori du système étudié.

Tableau IV.3 : Résultats des deux critères utilisés pour choisir le modèle PARAFAC valide. Chiffres en gras : valeurs des critères pour le modèle choisi.

Nombre de composantes du modèle PARAFAC	**Variance individuelle (%)**	**Core Consistency (%)**
1	27.50	100
2	18.00	99
3	17.06	96
4	14.81	93
5	6.25	92
6	6.24	95
7	5.04	90
8	3.62	90
9	4.91	98
10	3.62	94
11	3.62	97
12	3.41	97
13	1.30	98
14	1.31	99
15	1.31	99
16	**1.30**	**100**
17	0.41	<0
18	0.01	<0
19	0.01	<0
20	0.00	<0

IV.2.3 Identification des '*loadings*'

Une fois que le modèle PARAFAC est construit, l'attribution des '*loadings*' aux MEEF de références correspondantes est effectuée. Bien que la modélisation soit réalisée sur le cube de MEEF des espèces de référence, l'ordre dans lequel les MEEF sont introduites dans le modèle n'est pas nécessairement le même que celui des '*loadings*'. C'est pourquoi il est nécessaire de réaliser cette opération d'attribution. Habituellement, cette tâche est réalisée par une comparaison visuelle entre la MEEF du '*loading*' et des MEEF de références stockées dans une base de données. Cependant, dans notre cas, une interface graphique utilisateur '*Graphical User Interface*' (GUI) (Figure IV.4), a été développée sous l'environnement MATLAB® pour réaliser le calcul de distances métriques et d'angles spectraux entre celles-ci, à savoir (i) la distance euclidienne, (ii) le '*Spectral Angle Mapper*' (SAM) et (iii) le '*Hit Quality Index*' (HQI). Cette GUI est un outil interactif qui permet un gain de temps notable en automatisant cette tâche. Elle offre également d'autres possibilités que l'identification des '*loadings*', notamment un accès visuel rapide aux données de référence. De plus, elle est évolutive, car sa base de données peut être enrichie avec de nouvelles espèces, et le

programme peut être étendu pour accueillir d'autres fonctionnalités ainsi que d'autres types de données (e.g. des spectres Raman). Une présentation complète et une explication détaillée de la GUI ainsi que, la liste des CAP inclus dans la base de données sont disponibles dans l'annexe B.

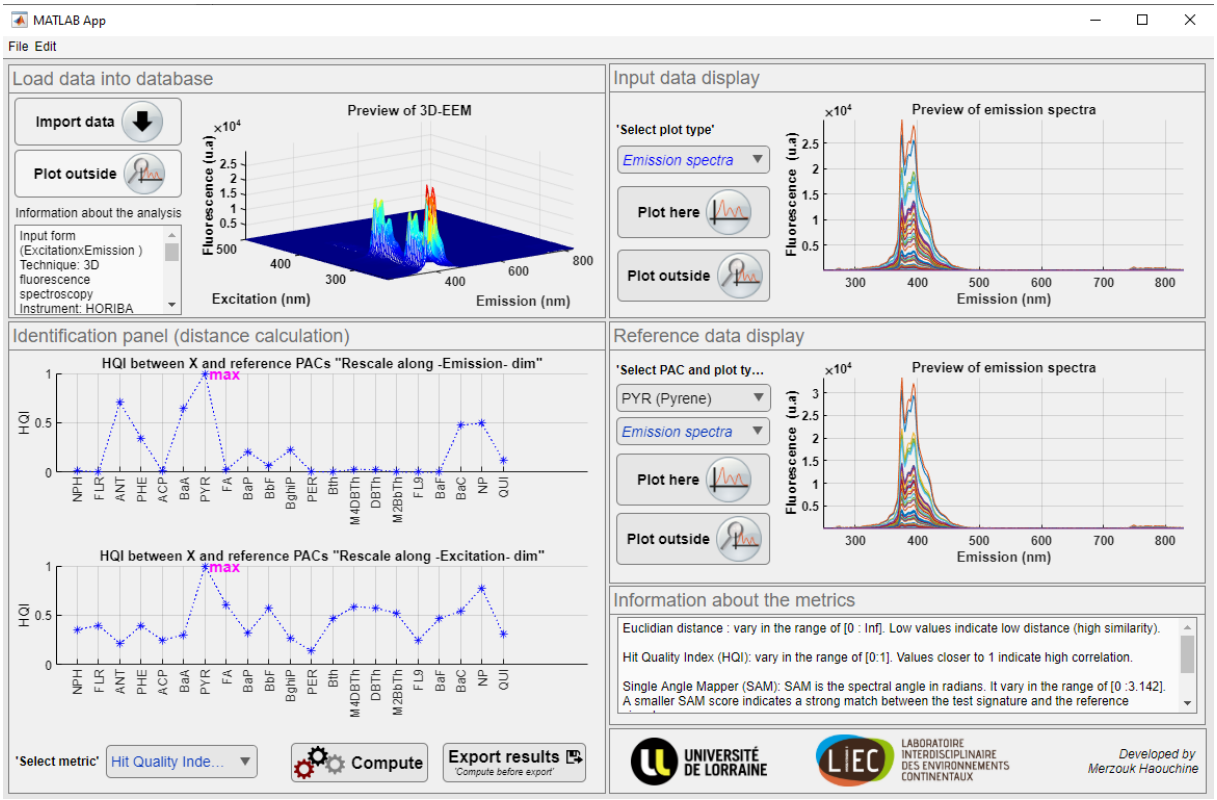


Figure IV.4 : Interface graphique utilisateur d'identification des cartes 3D des 'loadings'. Exemple d'identification, au moyen d'une des distances métrique (i.e. HQI), d'une carte 3D de 'loadings' qui correspond à la MEEF du pyrène.

Ainsi, les résultats de la modélisation PARAFAC avec 16 composantes sont présentés en termes de '*loadings*' et de leurs espèces pures correspondantes (Figure IV.5).

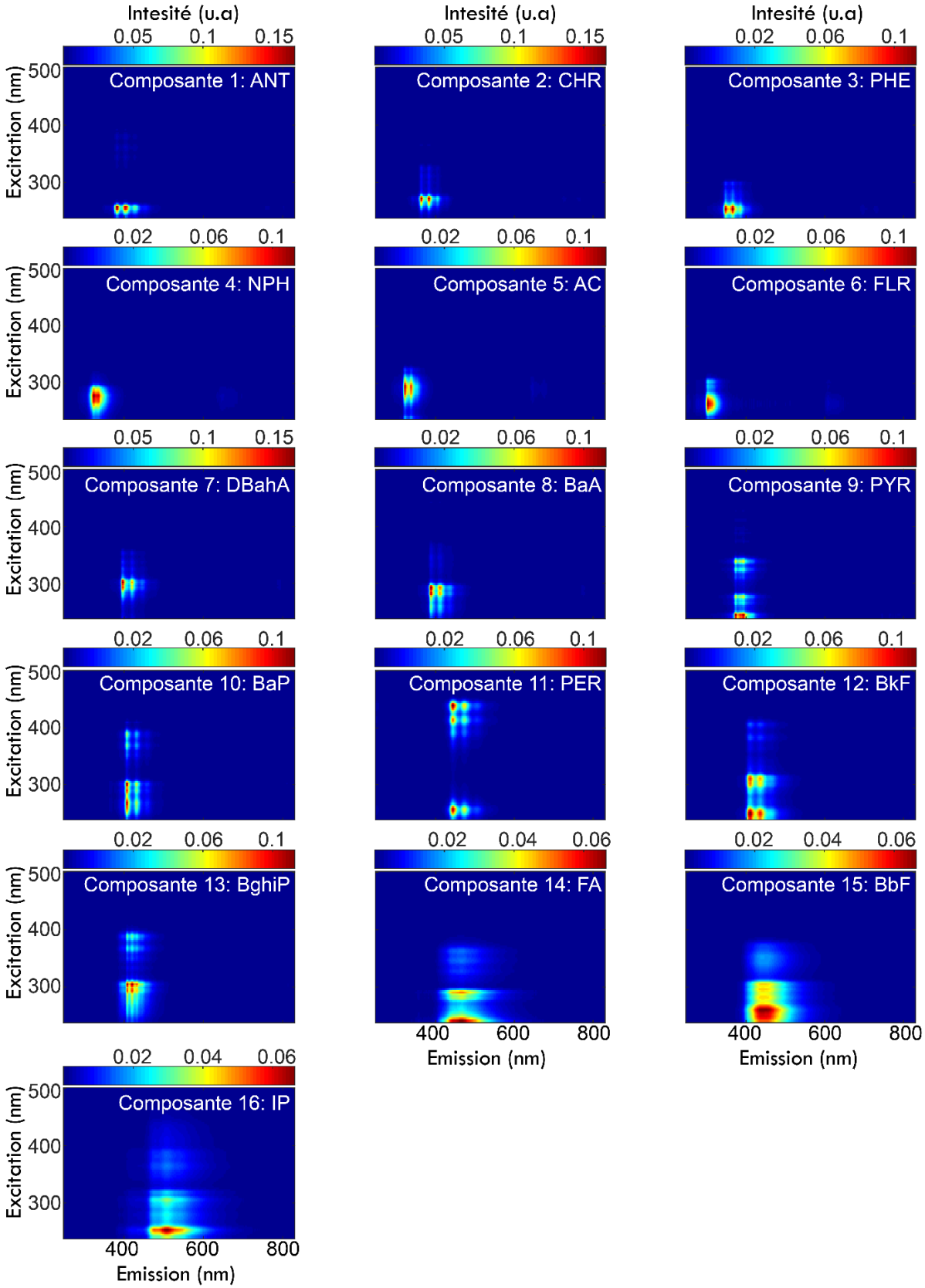


Figure IV.5 : Résultats de la décomposition PARAFAC avec 16 composantes en termes de 'loadings' et des espèces pures correspondantes.

IV.3 Validation du modèle PARAFAC

Pour s'assurer des performances et de la spécificité du modèle PARAFAC précédemment construit, une base de données de trois ensembles de validation a été construite. Pour **l'ensemble de validation #1**, 12 échantillons de deux fluorophores (i.e. NPH et BaA) de concentrations variables ont été préparés dans du dichlorométhane (Tableau IV.4a). Une MEEF est acquise pour chaque échantillon en utilisant les mêmes conditions expérimentales que précédemment. Ensuite, pour **l'ensemble de validation #2**, 11 échantillons de quatre fluorophores (i.e. NPH, BaA, PYR et ANT) de concentrations variables ont été préparés dans le même solvant (Tableau IV.4b). Une MEEF est acquise pour chaque échantillon en utilisant également les mêmes conditions expérimentales que précédemment. Enfin, pour **l'ensemble de validation #3**, des dilutions de la solution commerciale (PAH Mix 64) de chez DR EHRENSTORFER^{TM} ont été préparées dans du dichlorométhane. Cette solution commerciale est un mélange de 16 HAP, chacun à une concentration de 2 mg.L^{-1}. La solution fille ayant une concentration de 0.01 mg.L^{-1} a été retenue en raison du bon rapport signal sur bruit de sa MEEF après prétraitement par MT-SVD (Figure IV.6a). De plus, son spectre d'absorbance présente une absorbance maximale de 0.028 u.a à 239 nm (Figure IV.6b). Parmi les 16 HAP de ce mélange, 15 sont inclut dans le modèle PARAFAC construit précédemment. L'acénaphtylène est présent dans le mélange, mais il est absent du modèle. A l'inverse, le perylène est présent dans le modèle PARAFAC, mais il est absent de la solution commerciale (Tableau IV.4c).

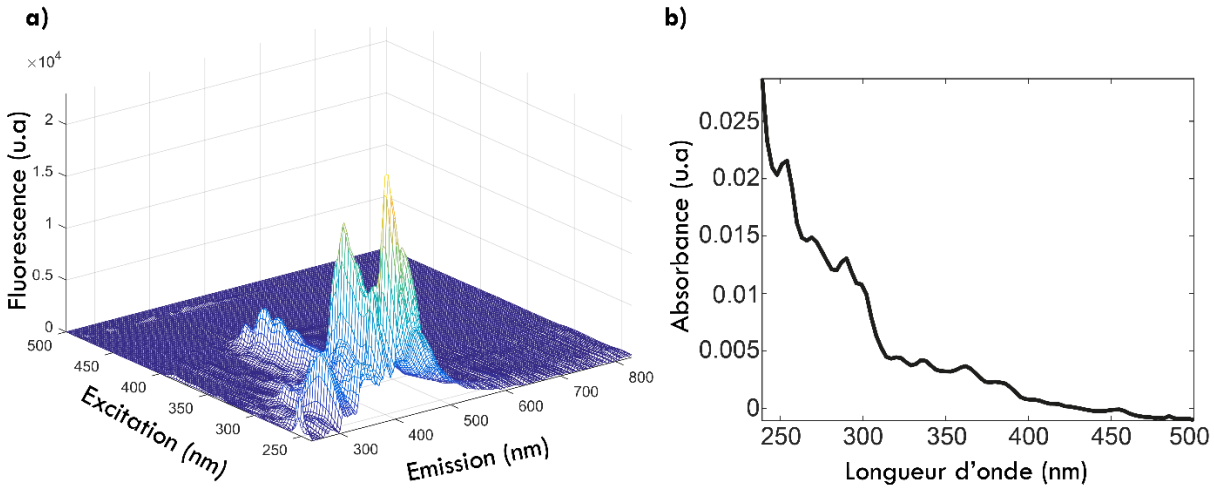


Figure IV.6 : MEEF, après prétraitement par MT-SVD, de la solution commerciale PAH MIX 64 à la concentration de 0.01 mg.L^{-1} (a) et son spectre d'absorbance (b).

Tableau IV.4 : Base de données des ensembles de validation du modèle PARAFAC. E signifie échantillon.

Concentrations des CAP en $mg.L^{-1}$																
a		**NPH**	**ANT**	**BaA**	**PYR**	**FA**	**PHE**	**FLR**	**BaP**	**BghiP**	**AC**	**BbF**	**DBahA**	**BkF**	**IP**	**PER**	**CHR**
Ensemble de validation 1	E1	**0.05**	/	**0.05**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E2	**0.05**	/	**0.1**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E3	**0.1**	/	**0.05**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E4	**0.25**	/	**0.1**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E5	**1**	/	**0.05**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E6	**1**	/	**0.01**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E7	**1**	/	**0.025**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E8	**0.125**	/	**0.05**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E9	**0.025**	/	**0.025**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E10	**0.5**	/	**0.025**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E11	**1**	/	**0.005**	/	/	/	/	/	/	/	/	/	/	/	/	/
	E12	**0.5**	/	**0.05**	/	/	/	/	/	/	/	/	/	/	/	/	/
b		**NPH**	**ANT**	**BaA**	**PYR**	**FA**	**PHE**	**FLR**	**BaP**	**BghiP**	**AC**	**BbF**	**DBahA**	**BkF**	**IP**	**PER**	**CHR**
Ensemble de validation 2	E1	**0.005**	**0.005**	**0.005**	**0.005**	/	/	/	/	/	/	/	/	/	/	/	/
	E2	**1**	**0.010**	**0.010**	**0.010**	/	/	/	/	/	/	/	/	/	/	/	/
	E3	**1**	**0.025**	**0.010**	**0.100**	/	/	/	/	/	/	/	/	/	/	/	/
	E4	**1**	**0.100**	**0.005**	**0.025**	/	/	/	/	/	/	/	/	/	/	/	/
	E5	**1**	**0.100**	**0.010**	**0.005**	/	/	/	/	/	/	/	/	/	/	/	/
	E6	**0.100**	**0.010**	**0.005**	**0.025**	/	/	/	/	/	/	/	/	/	/	/	/
	E7	**1**	**0.025**	**0.005**	**0.010**	/	/	/	/	/	/	/	/	/	/	/	/
	E8	**1**	**0.025**	**0.100**	**0.010**	/	/	/	/	/	/	/	/	/	/	/	/

	E9	**0.050**	**0.050**	**0.050**	**0.050**	/	/	/	/	/	/	/	/	/	/	/	/
	E10	**0.100**	**0.100**	**0.100**	**0.100**	/	/	/	/	/	/	/	/	/	/	/	/
	E11	**1**	**0.100**	**0.050**	**0.250**	/	/	/	/	/	/	/	/	/	/	/	/
c		**NPH**	**ANT**	**BaA**	**PYR**	**FA**	**PHE**	**FLR**	**BaP**	**BghiP**	**AC**	**BbF**	**DBahA**	**BkF**	**IP**	**PER**	**CHR**
Ensemble de validation 3	E1	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	/	**0.01**

IV.3.1 Résultats de la caractérisation par PARAFAC des 2 HAP présents dans l'ensemble de validation #1

Le modèle PARAFAC, construit sur la base des 16 espèces de références, présente de bons résultats pour identifier les deux espèces chimiques, que sont le NPH et le BaA, en concentrations variables, dans les 12 mélanges de l'ensemble de validation #1 (Figure IV.7).

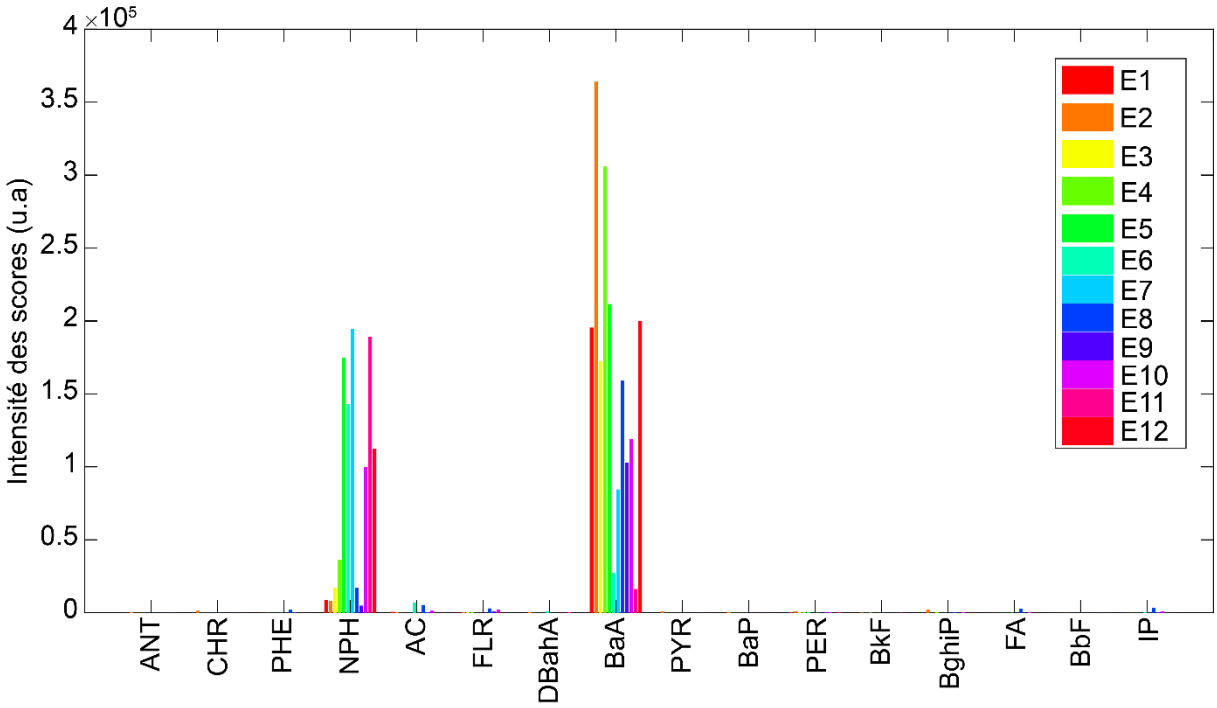


Figure IV.7 : Résultats de la décomposition PARAFAC de l'ensemble de validation #1. E signifie échantillon.

La matrice des '*scores*' indique la présence ou l'absence du NPH et du BaA au sein de ces mélanges de manière proportionnelle. Par exemple, pour l'échantillon E7, le NPH prédomine par rapport au BaA en raison des concentrations de ces deux espèces, qui sont de 1 $mg.L^{-1}$ pour le NPH et de 0.025 $mg.L^{-1}$ pour le BaA (Figure IV.8). Il convient toutefois de noter que la proportion des espèces n'est pas uniquement déterminée par la concentration, mais aussi par le rendement de fluorescence de la molécule, qui augmente généralement avec l'augmentation du degré de conjugaison de la molécule (cf. §I.7.1). Par exemple, pour l'échantillon E3, le NPH est à une concentration de 0.1 $mg.L^{-1}$ et le BaA à 0.05 $mg.L^{-1}$, mais la proportion de ce dernier en termes de '*scores*' est bien plus importante (Figure IV.8). Cela s'explique par le fait que le rendement de fluorescence du BaA est plus élevé que celui du NPH. Par conséquent, il est important de faire preuve de prudence lors de la quantification à l'aide des '*scores'* de PARAFAC. En effet, cette quantification doit être ciblée sur une seule espèce ou être alimentée par d'autres données quantitatives.

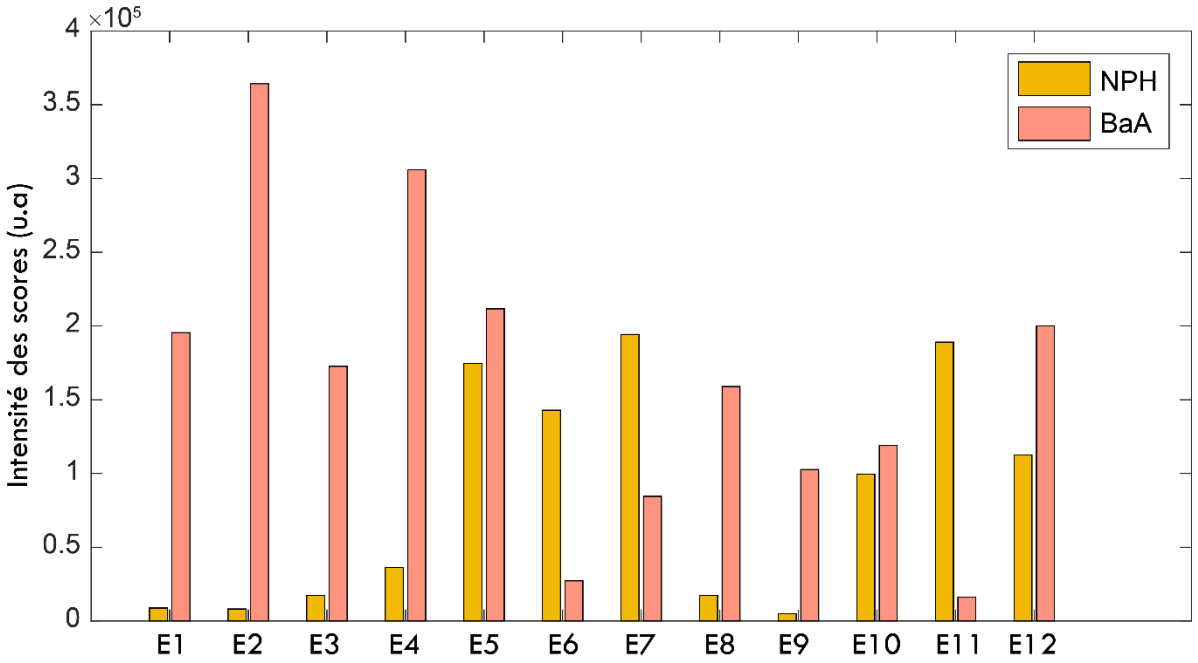


Figure IV.8 : Estimation de la présence ou de l'absence du NPH et BaA dans les échantillons de l'ensemble de validation #1. NPH : Naphtalène. BaA : Benz[α]anthracène.E signifie échantillon.

Une régression linéaire avec une ordonnée à l'origine égale à zéro a été établi entre les '*scores*' des deux composantes 4 et 8 et les concentrations réelles des HAP correspondants aux mélanges associés, à savoir le NPH et le BaA, respectivement. Les ajustements obtenus sont satisfaisants, car les valeurs des coefficients de détermination (i.e. R^2) observées se situent entre 0.95 et 0.96 (Figure IV.9). Cela indique une tendance linéaire entre les '*scores*' et les concentrations réelles en HAP.

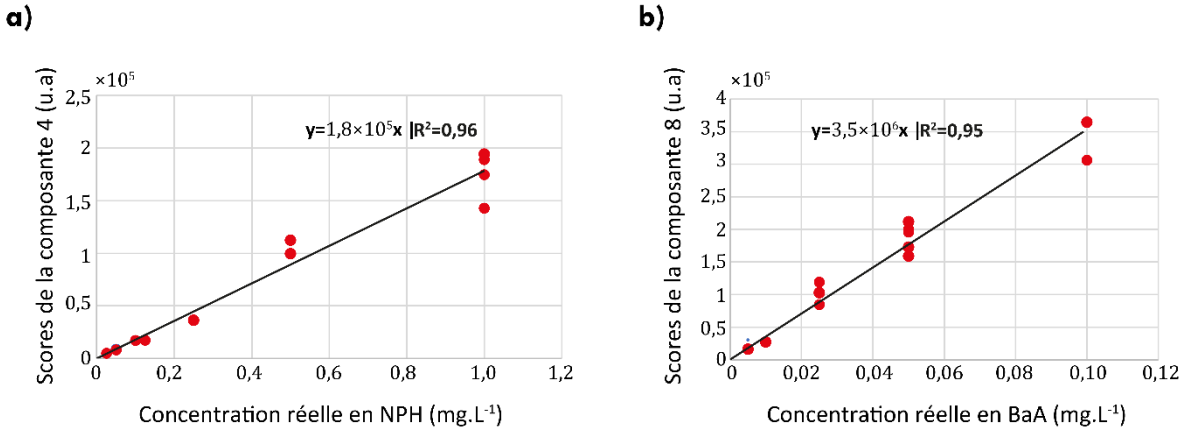


Figure IV.9 : Concentration réelle en NPH des mélanges de l'ensemble de validation#1 versus les 'scores' de la composante 4 (a). Concentration réelle en BaA des mélanges de l'ensemble de validation#1 versus les 'scores' de la composante 8 (b).

IV.3.2 Résultats de la caractérisation par PARAFAC des 4 HAP présents dans l'ensemble de validation #2

Le modèle PARAFAC présente de bons résultats dans l'identification des quatre espèces chimiques (i.e. NPH, ANT, BaA et PYR) présentes, en concentrations variables, dans les 11 mélanges de l'ensemble de validation #2 (Figure IV.10).

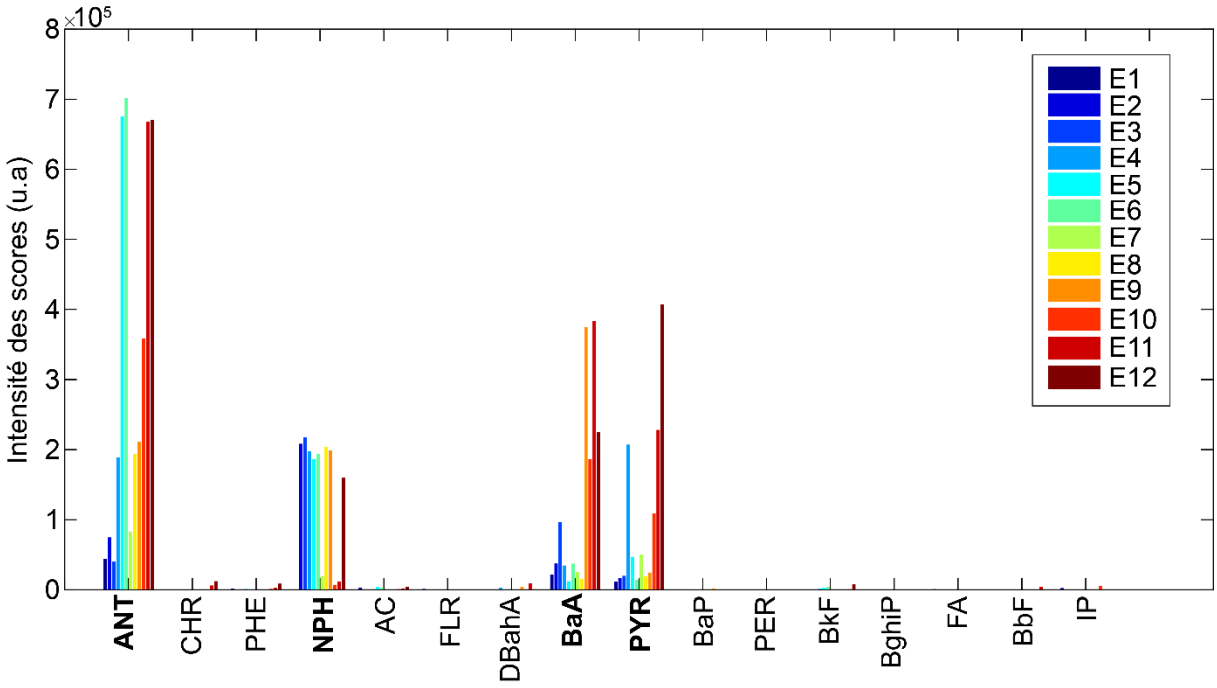


Figure IV.10 : Résultats de la décomposition PARAFAC de l'ensemble de validation #2. E signifie échantillon.

Par ailleurs, un autre type de représentation (Figure IV.11) permet de regrouper les '*scores*' des quatre HAP dans chaque échantillon, sur une seule barre. Cela permet de visualiser rapidement les proportions de chaque HAP dans chaque mélange. La hauteur de chaque barre correspond à la somme des intensités des '*scores*' des quatre HAP. Ainsi, l'intensité du '*score*' de chaque HAP est la différence entre l'intensité maximale et l'intensité minimale lues après projection sur l'axe des '*scores*' à gauche de la figure. Par exemple, dans l'échantillon 2 noté E2, le NPH est prédominant en termes de '*scores*' avec une valeur d'environ 2×10^5 u.a. Il est suivi de l'ANT, qui a un '*score*' inférieur à 1×10^5 u.a. Les '*scores*' du BaA et du PYR sont les plus faibles pour cet échantillon, avec moins de 0.5×10^5 u.a chacun. Nous remarquons que le NPH a une limite de détection de 0.05 $mg.L^{-1}$ (e.g. E9), en dessous de laquelle cette espèce, avec un rendement de fluorescence plus faible que les autres HAP de l'étude, n'est plus détectée par fluorescence 3D (e.g. E1 où le NPH est à une concentration de 0.005 $mg.L^{-1}$) (Figure IV.11).

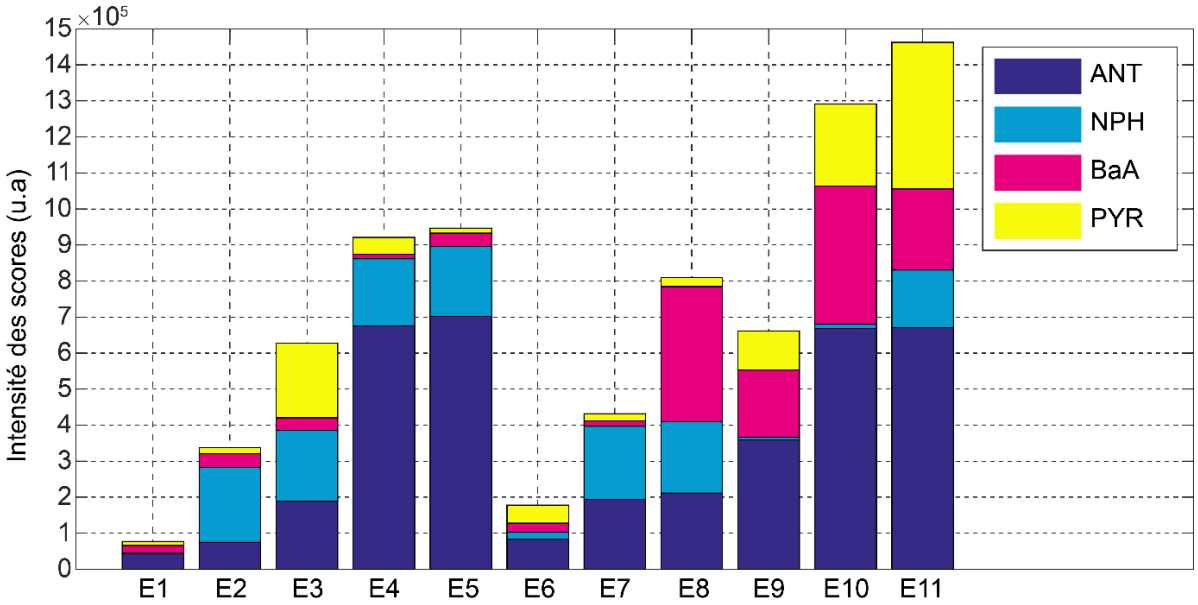


Figure IV.11 : Estimation de la présence ou de l'absence des 4 HAP dans les échantillons de l'ensemble de validation #2. NPH : naphtalène. BaA : benz[α]anthracène. ANT : anthracène. PYR : pyrène. E signifie échantillon.

Une régression linéaire avec une ordonnée à l'origine imposée à zéro a été établi entre les '*scores*' des quatre composantes 4, 8, 1 et 9 et les concentrations réelles des HAP correspondants dans les mélanges, à savoir le NPH, le BaA, l'ANT et le PYR, respectivement. Les résultats obtenus sont relativement satisfaisants, car les valeurs des coefficients de détermination (i.e. R^2) observées sont respectivement 0.98, 0.99, 0.99 et 0.97 pour le le NPH, le BaA, l'ANT et le PYR (Figure IV.12). Cela indique une tendance linéaire entre les '*scores*' et les concentrations réelles en HAP.

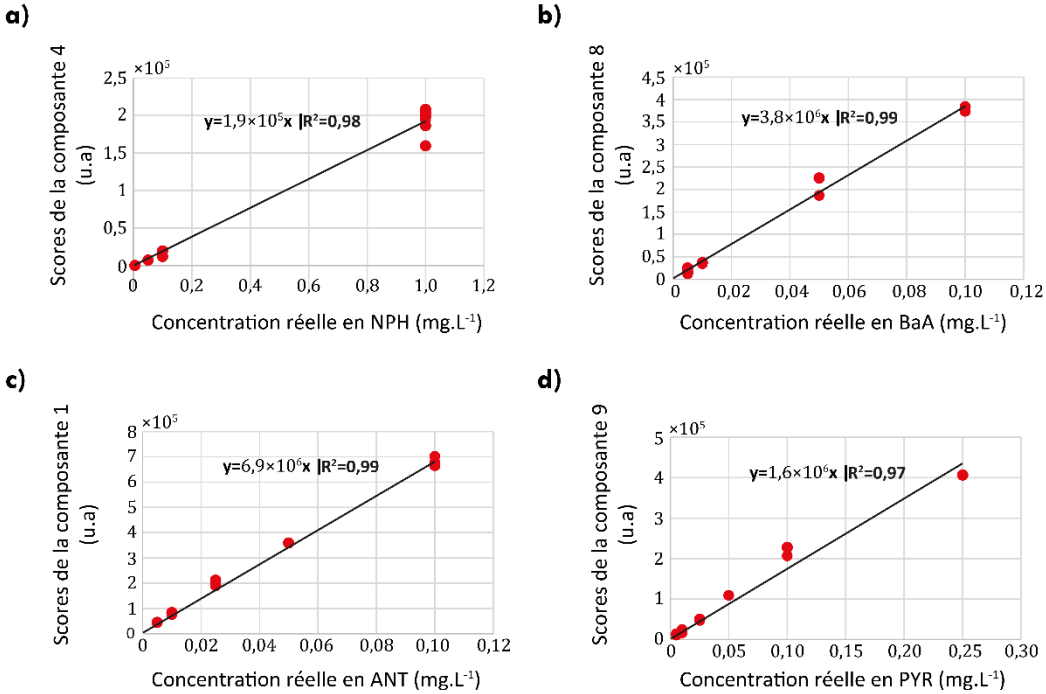


Figure IV.12 : Concentration réelle en NPH des mélanges de l'ensemble de validation#2 versus les 'scores' de la composante 4 (a). Concentration réelle en BaA des mélanges de l'ensemble de validation#2 versus les 'scores' de la composante 8 (b).

Concentration réelle en ANT des mélanges de l'ensemble de validation#2 versus les 'scores' de la composante 1 (c). Concentration réelle en PYR des mélanges de l'ensemble de validation#2 versus les 'scores' de la composante 9 (d).

IV.3.3 Résultats de la caractérisation qualitative par PARAFAC de 15 HAP présents dans l'ensemble de validation #3

Le modèle PARAFAC construit présente de bons résultats dans l'identification de l'ensemble des HAP présents dans la solution commerciale PAH Mix 64, à l'exception bien évidemment de l'acénaphtylène qui n'est pas inclus dans le modèle. De plus, ce cas de validation met en évidence la spécificité du modèle PARAFAC dans les mélanges complexes avec plusieurs fluorophores et face aux phénomènes de recouvrement spectral. En effet, le PER, qui est inclus dans le modèle, n'est pas détecté car il est absent dans la solution (Figure IV.13). En outre, ce cas de validation met également en évidence la complexité dans le calcul des '*scores'* de PARAFAC lié à la dualité entre la concentration et le rendement de fluorescence d'une molécule. En effet, si seule la concentration influençait leur calcul, les intensités des '*scores'* devraient être les mêmes pour toutes les molécules, car elles sont toutes à la même concentration or, nous observons une variabilité significative des intensités des '*scores'*. De plus, nous remarquons que les molécules avec deux cycles aromatiques (i.e. NPH, AC) ne sont pas les seules à présenter des '*scores'* faibles. Le DBahA, par exemple, présente également un '*score*' relativement faible par rapport à ce qui aurait pu être attendu en fonction de sa structure (i.e. 5 cycles aromatiques). Cela peut s'expliquer par le phénomène de recouvrement spectral et par l'influence des phénomènes physico-chimiques tels que l'inhibition de la fluorescence (cf. §I.6) qui viennent s'ajouter à la dualité entre concentration et rendement de fluorescence (Figure IV.13).

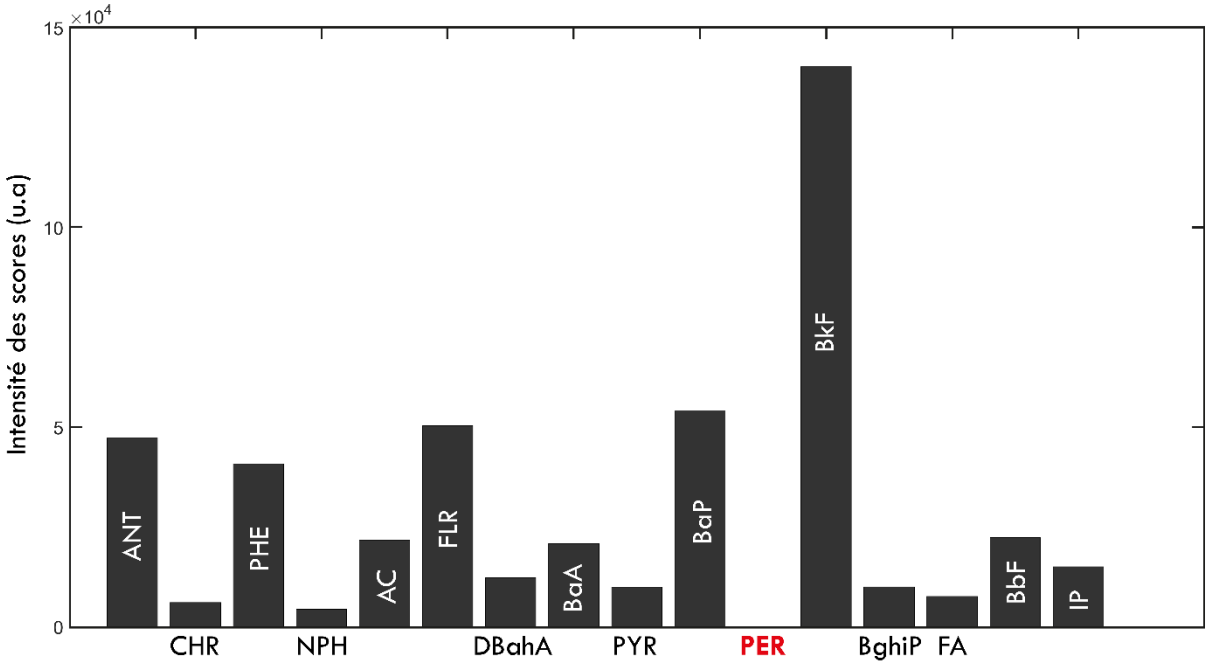


Figure IV.13 : Identification des 15 HAP présents dans la solution commerciale (PAH mix 64).

Par conséquent, Les résultats obtenus sur la série des trois ensembles de validation ont permis de valider le modèle d'identification PARAFAC en termes de spécificité. En effet, avec l'ensemble de validation #1, nous avons observé une identification précise des deux HAP présents uniquement. De même, pour l'ensemble de validation #2, seuls les quatre HAP présents dans les mélanges ont été identifiés par le modèle, malgré les concentrations variables et parfois faibles. Enfin, grâce à l'ensemble de validation #3, nous avons pu identifier 15 des 16 HAP présents dans le mélange. Cet ensemble de validation nous a également permis de confirmer l'absence connue a priori du perylène dans ce mélange. Un autre résultat important avec ces différents ensembles de validation est qu'il n'y a pas de corrélation simple entre concentration et signal de fluorescence. Ceci nous montre déjà la difficulté sur des échantillons complexes de pourvoir prédire la teneur en HAP d'échantillon complexe.

À présent, ce modèle PARAFAC va être utilisé pour l'identification de ces 16 HAP cibles dans des matrices environnementales complexes que sont les extraits organiques de sols pollués par les HAP.

IV.4 Analyse des MEEF des extraits de sols pollués par les HAP

L'étude qui sera présentée dans cette section vise à caractériser la pollution par les HAP de sols industriels contaminés par une activité humaine, présente ou passée. Pour atteindre cet objectif, des extraits organiques de sols prélevés sur différents sites répertoriés, ont été préparés pour pouvoir les mesurer en fluorescence 3D. Les MEEF obtenues ont ensuite été projetées dans le modèle PARAFAC construit précédemment et validé. Les profils des '*scores*' résultants ont été analysés afin de déterminer les caractéristiques de la pollution de chaque sol par les 16 HAP du modèle.

IV.4.1 Origine, préparation et nomenclature des sols

Les 37 échantillons de sol dans le cadre de cette thèse ont été prélevés sur sept sites industriels européens, à savoir six en France et un en Suède. Certains de ces sites sont en état de friche, tandis que d'autres sont en cours d'exploitation. Sept sols ont, en effet, été prélevés en France sur d'anciens sites de cokéfaction dans la région Lorraine, à savoir Neuves-Maisons (NM), Moyeuvre-Grande (MG), Homécourt (H, HEOM et Href) et Thionville (Thi et Thi_Ra). De plus, un sol a été également prélevé sur le site d'une ancienne usine à gaz localisée à Rennes (GR). Et enfin, les deux derniers sols ont été prélevés dans des sites de traitement du bois, l'un situé en France dans les Midi-Pyrénées (B) et l'autre en Suède (BS). Concernant le cas particulier du sol Thi, quatre échantillons de sols (Thi 87, Thi 163, Thi 166 et Thi 170) ont été prélevés dans quatre endroits différents de la même cokerie (Thi). Tous les sols ont été stockés à -20 °C pour prévenir la volatilisation des composés de faible poids

moléculaire. Ils ont ensuite été soumis à plusieurs étapes de préparation. Ils ont d'abord été séchés à l'air, homogénéisés et tamisés à 2 mm, sauf les sols Thi et HEOM qui ont été tamisés à 5 mm. Puis, ils ont été lyophilisés, broyés et tamisés à 500 µm et à 100 µm, sauf pour les sols HEOM, Href, Thi et Thi_Ra qui ont été tamisés uniquement à 500 µm[23]. Des réplicas ont été réalisées pour chaque sol, à l'exception du sol Thi (Tableau IV.5).

Tableau IV.5 : Répartition et nomenclature des 37 échantillons de sols pollués aux HAP.

Industrie	Site de traitement du bois							
Nom du sol	B				BS			
Tamisage	100 µm		500 µm		100 µm		500 µm	
Réplicas	R1	R2	R1	R2	R1	R2	R1	R2
Industrie	Usine de transformation de la houille							
Activité	Production de gaz (usine à gaz)							
Nom du sol	GR							
Tamisage	100 µm				500 µm			
Réplicas	R1		R2		R1		R2	
Industrie	Usine de transformation de la houille							
Activité	Production de la coke (Cokerie)							
Nom du sol	MG				NM			
Tamisage	100 µm		500 µm		100 µm		500 µm	
Réplicas	R1	R2	R1	R2	R1	R2	R1	R2
Industrie	Usine de transformation de la houille							
Activité	Production de la coke (Cokerie)							
Nom du sol	HEOM			Href			H	
Tamisage	500 µm			500 µm			100 µm	500 µm
Réplicas	R1	R2	R3	R1	R2	R3	R1	R2	R1	R2
Industrie	Usine de transformation de la houille							
Activité	Production de la coke (Cokerie)							
Nom du sol	Thi				Thi_Ra			
	Thi87	Thi163	Thi166	Thi170				
Tamisage	500 µm				500 µm			
Réplicas	/	/	/	/	R1		R2	R3

Dans la suite de ce document, chaque échantillon répliqué et tamisé à 500 µm et 100 µm est identifié par le nom du sol, suivi de la taille de tamisage et du numéro du réplica. Par exemple, MG500R1 fait référence au premier réplica du sol MG ayant subi un tamisage à 500 µm. Les échantillons n'ayant qu'un seul tamisage (i.e. 500 µm) sont représentés uniquement par le nom du sol suivi du numéro du réplica (ou uniquement le nom du sol pour les échantillons sans réplicas). Par exemple, Thi_RaR1 signifie le premier réplica du sol Thi_Ra. De plus, l'ensemble des échantillons d'un même sol est qualifié de groupe. Par exemple, le groupe Href comprend les échantillons HrefR1, HrefR2 et HrefR3. Dans le cas particulier des échantillons Thi87, Thi163, Thi166 et Thi170, ils sont regroupés dans un même groupe (i.e. Thi) car ils ne possèdent pas de réplica. Une fois les échantillons préparés, la matière organique porteuse de la pollution aux HAP est extraite à l'aide de dichlorométhane par une méthode d'extraction par solvant accélérée '*Accelerated solvent extraction*' (ASE).

IV.4.2 Extraction ASE de la phase organique porteuse de la pollution aux HAP[126,127]

L'extraction de la matière organique porteuse de la pollution aux HAP a été réalisée à l'aide d'un extracteur ASE 350 (Dionex). Pour cela, à la base de la cellule de l'extracteur, 2 g de sulfate de sodium anhydre et 1 g de cuivre activé ont été introduits entre deux filtres en fibre de verre (filtres GF/F Wathman). Le sulfate de sodium permet d'absorber les traces d'eau résiduelles présentes dans l'échantillon qui pourraient être mobilisées lors de l'extraction, tandis que le cuivre activé permet de piéger le soufre moléculaire présent naturellement dans le sol, le faisant précipiter sous forme de CuS. La présence de soufre peut fortement perturber les analyses réalisées par chromatographie en phase gazeuse, c'est pourquoi cet ajout de cuivre est fait systématiquement. Une fois la cellule nettoyée, 2 g d'échantillon de terre ont été introduits dans la cellule et extraits lors de 2 cycles de 5 minutes chacun, en utilisant du dichlorométhane comme solvant, à une température de 130 °C et une pression de 100 bars. L'extrait a été collecté dans un flacon et ramené à un volume connu puis des aliquotes ont été prélevées et placées dans des flacons préalablement pesés. La teneur en matière organique extractible a été déterminée en pesant les flacons après évaporation complète du solvant sous un flux léger d'azote.

IV.4.3 Base de données des MEEF des extraits de sols

Pour la construction de la base de données, les extraits de sols pollués aux HAP ont été analysés par fluorescence 3D. Les mêmes conditions expérimentales que celles fixées pour la construction de la base de données des espèces de référence ont été utilisées (i.e. chaque MEEF est de taille $(88,250)$). Plusieurs dilutions ont été effectuées afin d'établir la concentration optimale en matière organique extractible (i.e. phase organique porteuse de la

pollution) des échantillons en recherchant, le meilleur rapport signal sur bruit des MEEF. Une concentration de 0.3 mg.L^{-1} a été choisie pour tous les échantillons. Pour chaque échantillon, 5 acquisitions ont été réalisées, ce qui donne un total de 185 acquisitions (i.e. 185 MEEF) pour les 37 échantillons de l'étude. Toutes les MEEF ont été prétraitées par MT-SVD (i.e. correction des effets de diffusion de la lumière) et chaque MEEF est normalisée suivant sa norme L1. La MEEF moyenne de chaque groupe, ainsi que la MEEF moyenne de chaque classe du groupe Thi, sont illustrées dans la Figure IV.14.

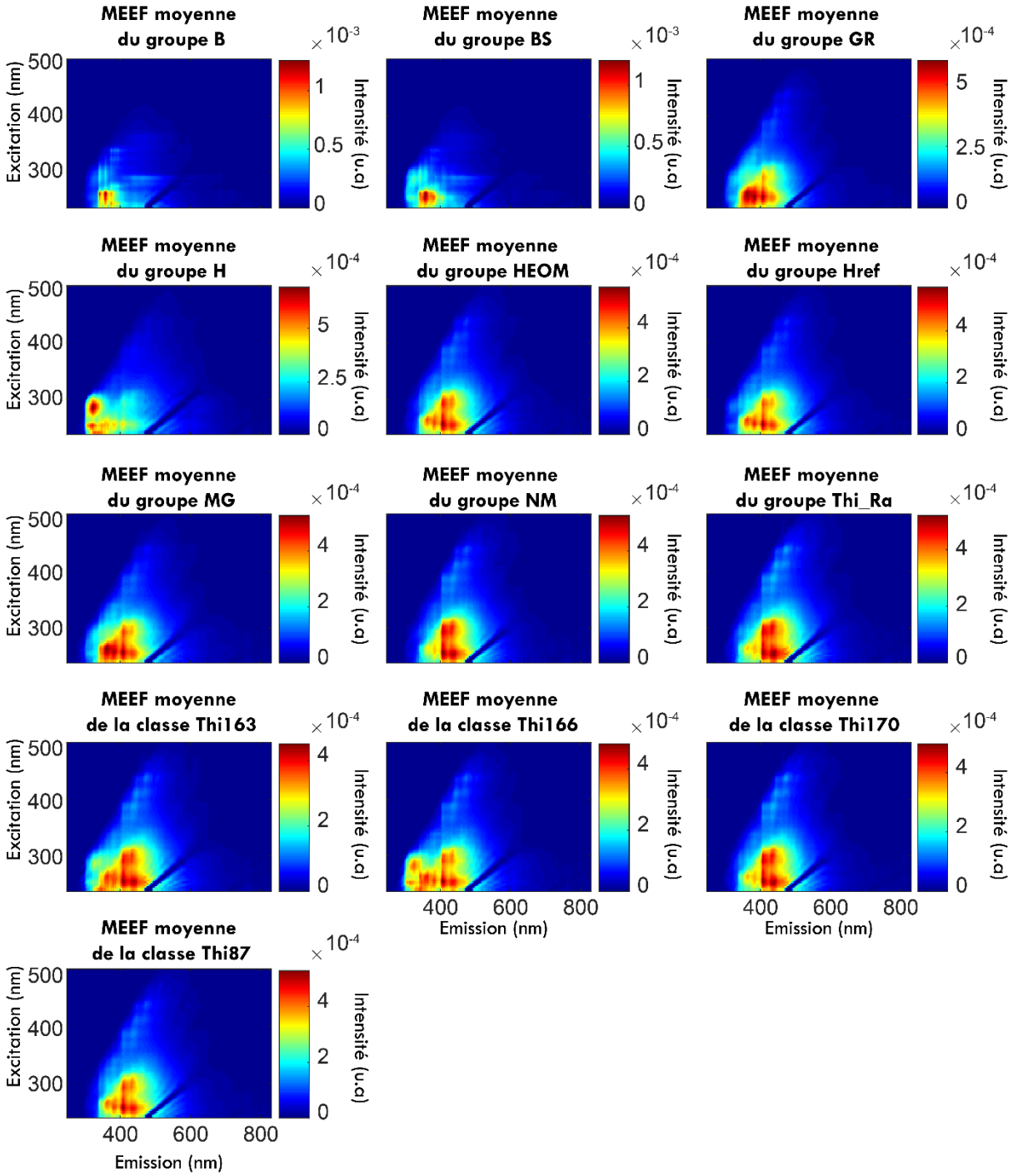


Figure IV.14 : MEEF moyenne (88,250) de chaque groupe, ainsi que la MEEF moyenne de chaque classe du groupe Thi.

De plus, un spectre d'absorbance est mesuré pour chaque acquisition en fluorescence 3D (i.e. 5 acquisitions par échantillon) sur la même plage spectrale que les spectres d'excitation afin de vérifier que l'absorbance de chaque échantillon à la concentration choisie respecte la condition de linéarité de la loi de Beer-Lambert. Les spectres obtenus montrent des absorbances maximales proches de 0.05 u.a (majoritairement inférieures). Par exemple, le spectre moyen des cinq spectres d'absorbance de l'échantillon B100R1 présente une absorbance maximale de 0.029 u.a à 239 nm, tandis que celui de B100R2 présente une absorbance maximale de 0.064 u.a à la même longueur d'onde (Figure IV.15).

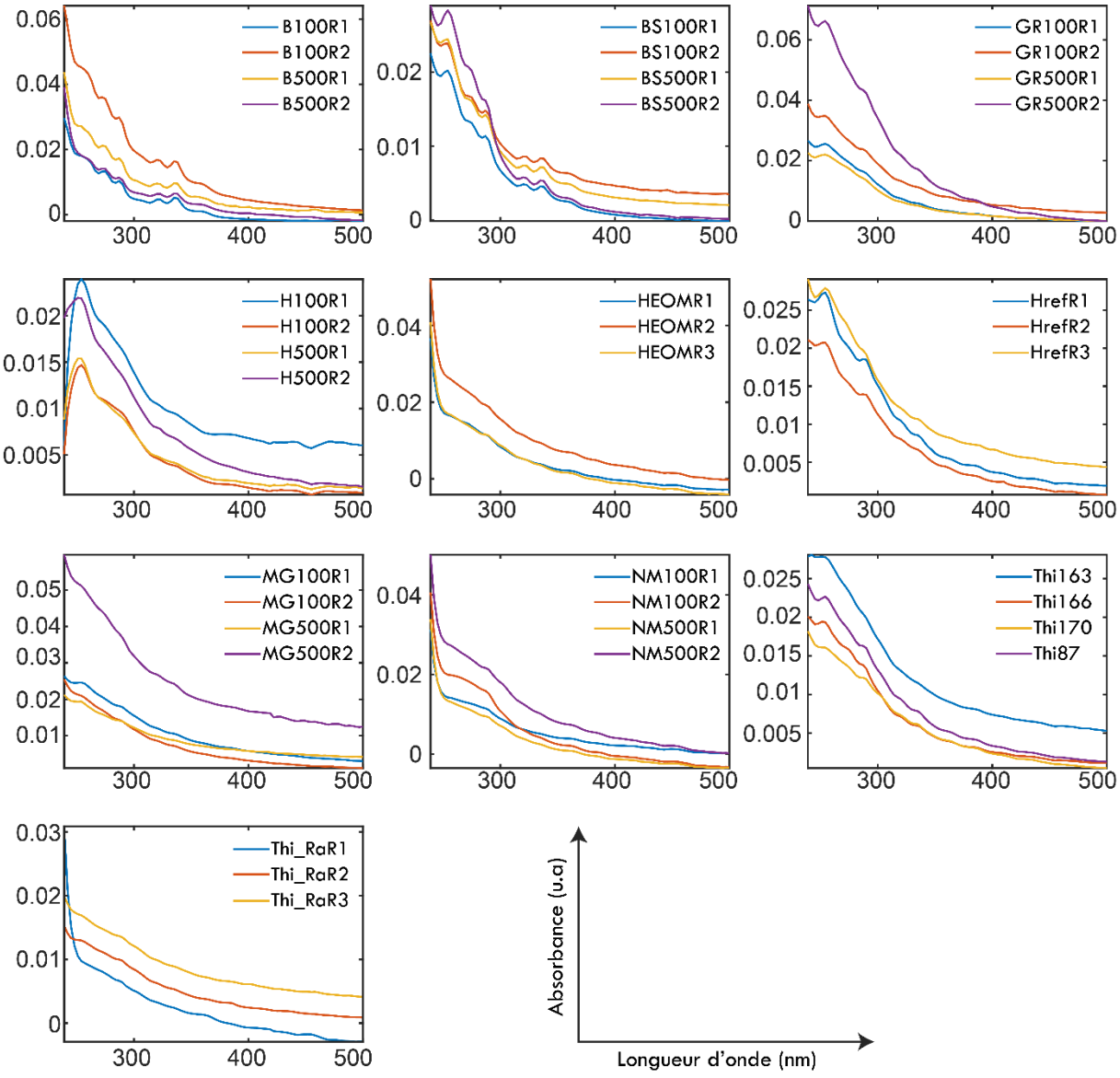


Figure IV.15 : Spectre moyen des 5 spectres d'absorption de chaque échantillon.

Compte tenu du nombre important de MEEF (i.e. 185), un modèle ACPM (Analyse en Composantes Principales Multivoie) a été construit pour analyser le cube de données de taille (185,250,88) et visualiser les MEEF dans l'espace réduit créé par les composantes principales. Ce modèle a été utilisé pour décrire l'ensemble des données, en évaluant les variations intra-

classe (i.e. distance entre les 5 acquisitions d'un même échantillon), les variations inter-classe (i.e. distance entre les échantillons d'un même groupe) ainsi que les variations inter-groupe (i.e. distance entre les différents groupes (sols)). Par ailleurs, ce modèle a aussi été utilisé pour détecter d'éventuels '*outliers*'. Un '*outlier*' est un individu (i.e. une acquisition, dans notre cas une MEEF) extrême ou atypique par rapport à la population étudiée (i.e. 185 MEEF). Il n'est pas obligatoirement un point aberrant, et donc n'est pas forcément à retirer de l'analyse de données[79]. Ainsi, cette analyse permet de repérer d'éventuels hétérogénéités et/ou erreurs liés à la réplication des extractions ou aux mesures en fluorescence.

IV.4.4 Etude exploratoire des extraits de sols par Analyse en Composantes Principales Multivoie

Le modèle ACPM a été construit avec 3 composantes principales qui permettent d'expliquer 97.86% de la variance totale. Les composantes principales CP1, CP2 et CP3 expliquent respectivement 77.77%, 16.36% et 3.73% de la variance. Deux distances multivariées ont été calculées pour détecter la présence éventuelle de '*outliers*' forts ou modérés. La première est la distance d'Hotelling (T^2) qui permet la détection de '*outliers*' forts. La deuxième est la distance '*Q residuals*' qui permet de détecter les '*outliers*' modérés. Alors que les '*Q residuals*' représentent l'amplitude de la variation restante dans chaque échantillon après projection à travers le modèle, les valeurs de Hotelling's T^2 représentent une mesure de la variation dans chaque échantillon au sein du modèle. Elles indiquent à quelle distance chaque échantillon se trouve du centre (i.e. '*scores'* = 0) du modèle[76].

La projection de l'ensemble des MEEF dans l'espace CP1-CP2 nous indique une bonne discrimination entre les groupes **H**, **BS**, **B**, **NM** et **Thi_Ra**. On note également une faible variation intra- et inter-classe au sein de ces groupes, ce qui indique une bonne répétabilité des mesures en fluorescence. Cependant, les groupes **MG**, **Href** et **HEOM** semble être très proches entre eux ce qui indique qu'ils présentent des signatures spectrales proches. Par ailleurs, le groupe des **Thi** présente une variabilité relativement importante entre les différentes classes qui le composent. Cette variation était attendue compte tenu de l'hétérogénéité de l'échantillonnage (i.e. prélèvements à différents endroits du site **Thi**). La classe **Thi 170** est proche du groupe **NM** et du groupe **Thi_Ra**, tandis que les classes **Thi 163** et **Thi 166** en sont éloignées mais restent assez proches entre elles (Figure IV.16a).

La projection des MEEF dans l'espace CP2-CP3 nous montre une discrimination selon l'axe de la CP3, du groupe **GR** et du groupe **Thi_Ra** par rapport aux autres. Cette projection met également en évidence la variation d'un individu par rapport à sa classe, qui est **MG500R2**, le long de l'axe CP3. Cependant, cette variation est très faible, car l'axe CP3 ne représente que 3.73% de l'information totale. Ainsi, cet individu qualifié de '*outlier*' modéré n'est pas

supprimé des données (Figure IV.16b). Néanmoins, il sera surveillé lors de la modélisation PARAFAC.

Enfin, la projection des MEEF dans l'espace CP1-CP3 met en évidence la variation d'un individu par rapport au reste de sa classe le long de la CP3. Il s'agit d'un individu de la classe **MG500R1**. Tout comme dans le cas précédent (espace CP2-CP3) pour l'individu **MG500R2**, cet individu n'est pas supprimé des données ('*outlier*' modéré) mais il sera surveillé lors de la modélisation PARAFAC. Par ailleurs, grâce à cette projection, on observe que la classe **Thi 87** est proche du groupe **HEOM** (Figure IV.16c).

La projection des MEEF dans l'espace distance T^2-'*Q residuals*' nous renseigne sur la présence éventuelle de '*outliers*' forts et/ou modérés. Les '*outliers*' forts se caractérisent par une distance T^2 élevée, tandis que les '*outliers*' modérés présentent un '*Q residuals*' élevé. On constate l'absence de '*outliers*' forts parmi les MEEF, ce qui indique une bonne modélisation ACPM de l'ensemble des MEEF (i.e. il n'y a pas de contribution forte de la part d'une minorité de MEEF). En revanche, les '*Q residuals*' nous indiquent la présence de deux classes (**MG500R2** et **Thi166**) avec un '*Q residuals*' élevé par rapport aux autres. Ces deux classes ne sont pas retirées des données, mais elles doivent être surveillées lors de la projection PARAFAC, d'autant plus que l'une des MEEF de chacune de ces deux classes se comporte comme un '*outlier*' modéré (Figure IV.16d).

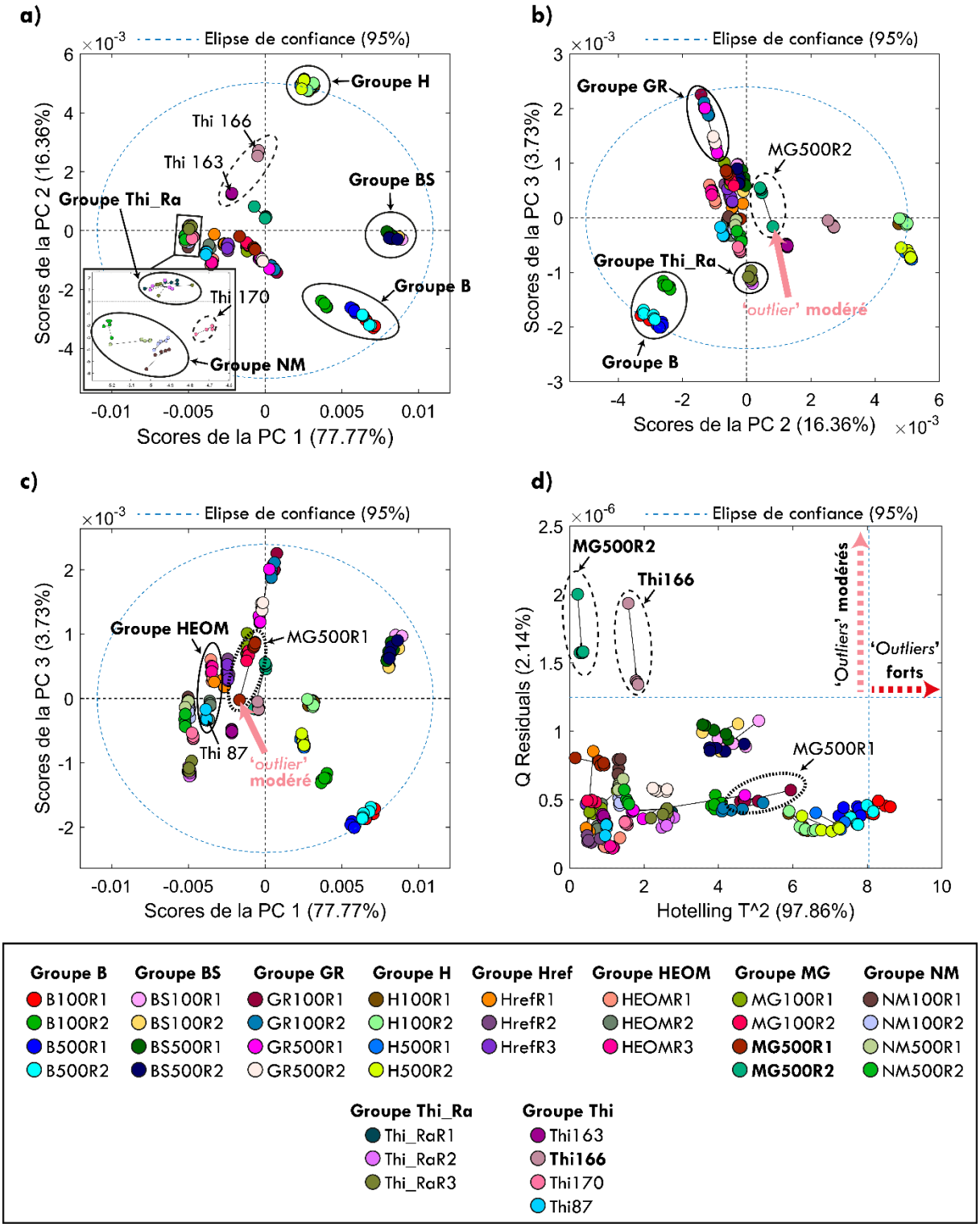


Figure IV.16 : Résultats de la modélisation ACPM des 185 MEEF des 37 échantillons d'extraits de sols pollués aux HAP. 'Scores' de la 1ère composante principale versus 'scores' de la 2ème composantes principale (a). 'Scores' de la 2ème composante principale versus 'scores' de la 3ème composantes principale (b). 'Scores' de la 1ère composante principale versus 'scores' de la 3ème composantes principale (c). Distance T^2 versus 'Q residuals'. PC signifie composante principale (d).

En termes de '*loadings*', les profils d'émission et d'excitation présentent des variations positives ou négatives, indiquant les longueurs d'onde relatives aux individus se situant soit vers des valeurs positives, soit vers des valeurs négatives des '*scores*' correspondants, respectivement. Pour plus de lisibilité, les résultats sont discutés par CP.

- CP1 : les groupes **BS** et **B** présentent des '*scores'* CP1 positifs (Figure IV.16a), ce qui indique qu'ils sont représentés par les signaux positifs de la CP1 se situant entre 300 et 400 nm pour l'émission et entre 250 et 350 nm pour l'excitation (Figure IV.17a). A l'inverse, les groupes **NM** et **Thi_Ra** sont représentés par les profils d'émission et d'excitation négatifs de la CP1, se situant entre 400 et 550 nm pour l'émission et entre 250 et 350 nm pour l'excitation (Figure IV.17a). Il est à noter que dans ces cas, la discrimination est principalement liée à la dimension d'émission.

- CP2 : le groupe **H** présente des '*scores'* CP2 positifs, tandis que le groupe **B** présente des '*scores'* négatifs (Figure IV.16a). Le groupe **H** est caractérisé par les signaux positifs dans la plage de 300 à 340 nm pour l'émission et de deux bandes principales à 250 nm et 280 nm pour l'excitation (Figure IV.17b). En revanche, le groupe **B** est représenté par une plage d'émission allant de 340 nm à 430 nm. Du côté de l'excitation, il est représenté par une plage allant de 240 nm à 340 nm avec une bande principale à 260 nm (Figure IV.17b). Enfin, le groupe **BS** présente des '*scores'* pratiquement nuls pour la CP2, ce qui indique qu'il n'est pas bien représenté par l'ensemble des signaux de cette dernière.

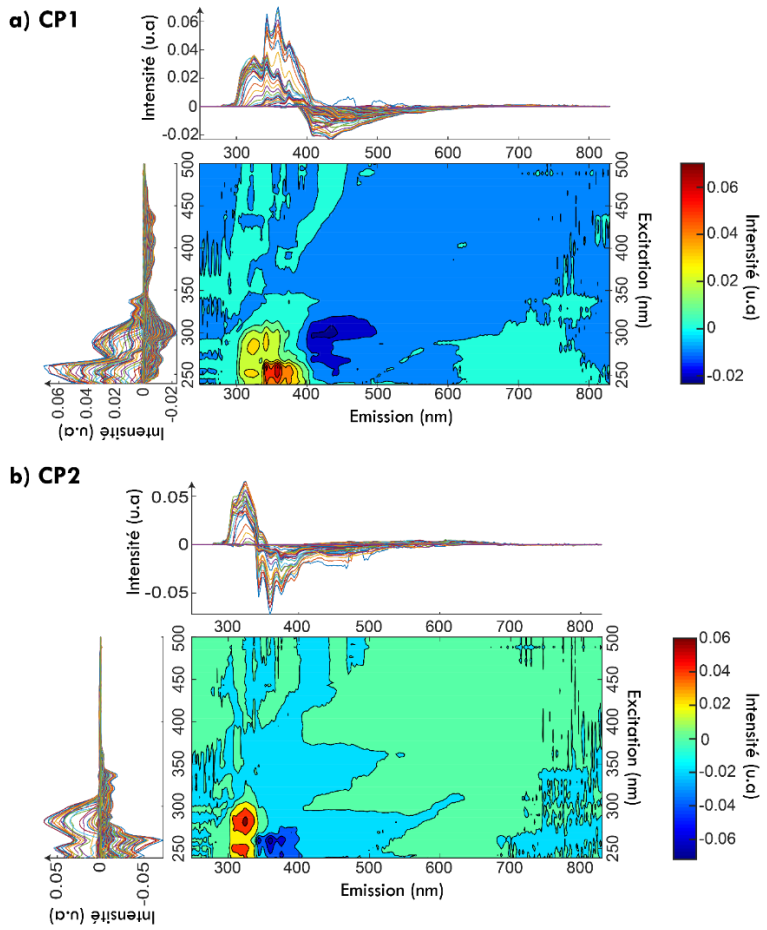


Figure IV.17 : 'Loadings' de la modélisation ACPM des 185 MEEF des 37 échantillons d'extraits de sols pollués aux HAP. 'Loading' de la 1ère composante principale (a). 'Loading' de la 2ème composante principale (b).

- CP3 : le groupe **GR** présente des '*scores'* CP3 positifs, tandis que les groupes **B** et **Thi_Ra**, présentent des '*scores'* négatifs. Les autres groupes se rapprochent plutôt des '*scores'* nuls de cette CP (Figure IV.16b). Le groupe **GR** est caractérisé par des signaux positifs en émission dans la plage de 300 à 470 nm, avec une succession de bandes principales à 310 nm, 350 nm, 368 nm, 384 nm, 407 nm et 428 nm. Du côté de l'excitation, il présente une plage allant de 240 nm à 425 nm, avec une bande majoritaire à 254 nm (Figure IV.18). Les groupes **B** et **Thi_Ra** sont représentés par des bandes d'émission situées à 343 nm, 359 nm, 375 nm et 393 nm, et des bandes d'excitation situées entre 240 nm et 380 nm, avec une bande majoritaire à 380 nm (Figure IV.18).

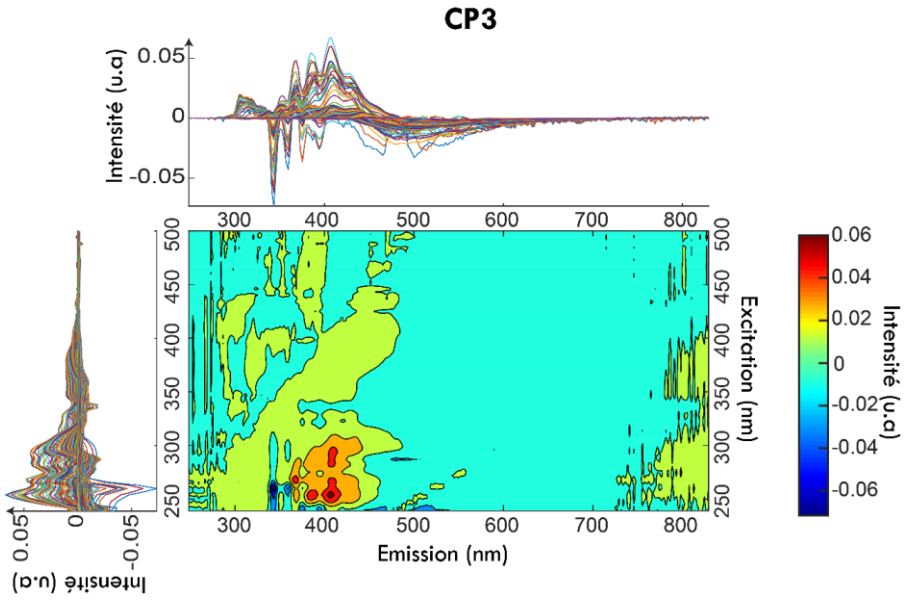


Figure IV.18 : 'Loading' de la 3ème composante principale de la modélisation ACPM des 185 MEEF des 37 échantillons d'extraits de sols pollués aux HAP.

En conclusion, la modélisation ACPM de notre cube de données de dimensions $(185,250,88)$ nous a permis de visualiser les 185 MEEF dans l'espace réduit des 3 composantes principales. Elle nous a fourni des informations sur les similarités et les dissimilarités des groupes d'une part mais d'autre part aussi sur la nature des échantillons analysés. De plus, elle nous a permis d'identifier des '*outliers*' modérés sur lesquelles il faudra rester prudent dans la suite du travail. Toutes ces informations seront utiles, voire cruciales, pour l'analyse PARAFAC.

IV.4.5 Caractérisation qualitative des HAP cibles dans les extraits de sols par le modèle PARAFAC

Les 185 MEEF, de taille $(88,250)$ chacune, des extraits de sols pollués ont été introduites dans le modèle PARAFAC afin de prédire la présence ou l'absence des 16 HAP cibles en analysant les profils des '*scores'* obtenus. Ainsi, une fois les MEEF projetées dans

le modèle PARAFAC, le profil des '*scores*' de chaque groupe a été étudié pour caractériser la nature de la pollution au HAP de chaque sol et ainsi, identifier les HAP impliqués. Des comparaisons entre les groupes permettent de mettre en évidence des tendances communes ou contradictoires, observés sur la Figure IV.19.

- **Profil des '*scores'* du groupe B et du groupe BS Figure IV.19a et b :**

Concernant le groupe **B**, le HAP qui prédomine en termes de '*scores'* est le PHE, suivi du PYR puis du FA. Cependant, les '*scores*' le plus élevé du PHE ne signifie pas nécessairement que cette espèce est prédominante quantitativement, car la concentration n'est pas le seul facteur influençant le calcul des '*scores'* (cf. §IV.3.3). Par ailleurs, ce groupe se caractérise par une faible contribution des composés : ANT, BaA, BaP, PER, BkF, BghiO et IP (Figure IV.19a). En ce qui concerne le profil des '*scores'* du groupe **BS**, il présente des similitudes avec celui du groupe **B**, à l'exception du FLR et de l'ANT qui semblent être plus présents sur ce sol (Figure IV.19b). Cette ressemblance entre les deux profils est logique, car il s'agit d'extraits de sols provenant de deux sites de la même activité industrielle toujours en fonctionnement, à savoir le traitement du bois.

- **Profil des '*scores'* du groupe GR Figure IV.19c :**

Le profil du groupe **GR** se différencie de celui des groupes **B** et **BS** avec des '*scores*' plus faibles pour le PHE, le NPH et le PYR, et même très faibles pour le FA. En revanche, les '*scores'* sont plus élevés pour le CHR, le DBahA, le BaA, le BaP, le PER, le BkF, le BghiP et le BbF (Figure IV.19c).

- **Profil des '*scores'* du groupe H Figure IV.19d :**

Pour ce sol, les '*scores'* les plus élevés sont ceux du PHE, de l'AC et du FLR, tandis que l'ANT et le BaP présentent les '*scores'* les plus faibles (Figure IV.19d).

- **Profil des '*scores'* du groupe HEOM et du groupe Href Figure IV.19e et f :**

Le groupe **HEOM** se caractérise par un '*score*' élevé pour le BbF, suivi par celui du PHE. En revanche, l'ANT, le FLR, le BaP, le FA et le NPH présentent les '*scores'* les plus faibles (Figure IV.19e). Le profil du groupe **Href** est similaire à celui du groupe **HEOM**, avec toutefois une légère augmentation du '*score*' pour le FA dans ce groupe. Par ailleurs, on note une baisse des '*scores'* du PHE et du PYR pour la classe HrefR1 (Figure IV.19f).

- **Profil des '*scores'* du groupe MG Figure IV.19g :**

Le '*score*' dominant de ce groupe est celui du PHE, suivi de près par celui du BbF. En revanche, l'ANT, le BaP et le FA présentent les '*scores'* les plus faibles (Figure IV.19g). La classe **MG500R2** montre un profil similaire aux autres classes du groupe, à l'exception du FLR qui présente un '*score*' plus élevé pour cette classe. En effet, la MEEF moyenne des 5 acquisitions de cette classe présente des intensités de fluorescence plus élevées dans la zone

relative au FLR allant de 300 à 340 nm pour l'émission et de 240 à 310 nm pour l'excitation (Figure IV.20). Compte tenu de la répétabilité des résultats pour l'ensemble des classes, cet écart, observé également en ACPM, peut être expliqué par une hétérogénéité ponctuelle de la pollution au sein de la classe **MG500R2**, ou encore par une hétérogénéité liée à l'échantillonnage.

- **Profil des '*scores'* du groupe NM et du groupe Thi_Ra Figure IV.19h et i :**

Le groupe **NM** présente un profil similaire à celui du groupe **HEOM**, à l'exception d'une légère baisse du '*score*' du PHE (Figure IV.19h). Le groupe **Thi_Ra** présente un profil similaire à celui du groupe **NM**, à l'exception d'une inversion du rapport de dominance entre le BkF et le BghiP, avec une légère dominance en faveur du BghiP pour le groupe **Thi_Ra** (Figure IV.19i).

- **Profil des '*scores'* du groupe Thi Figure IV.19j :**

Dans ce groupe, on observe une hétérogénéité entre les profils des différentes classes (Figure IV.19j). Le profil de la classe **Thi 170** présente des similitudes avec celui du groupe **NM** et du groupe **Thi_Ra**, comme cela a été observé dans les résultats de l'analyse ACPM (Figure IV.16a). Quant au profil de la classe **Thi 87**, il se rapproche davantage de celui du groupe **HEOM**, ce qui a également été observé dans l'analyse ACPM (Figure IV.16c). Pour la classe **Thi 166**, on remarque une présence plus importante du NPH, de l'AC et du FLR par rapport aux autres classes. En ce qui concerne la classe **Thi 163**, on observe une présence plus importante de l'AC. Les profils des classes **Thi 163** et **Thi 166** se ressemblent, à l'exception des '*scores'* du NPH et du FLR qui sont plus élevés pour la classe **Thi 166** (Figure IV.19j).

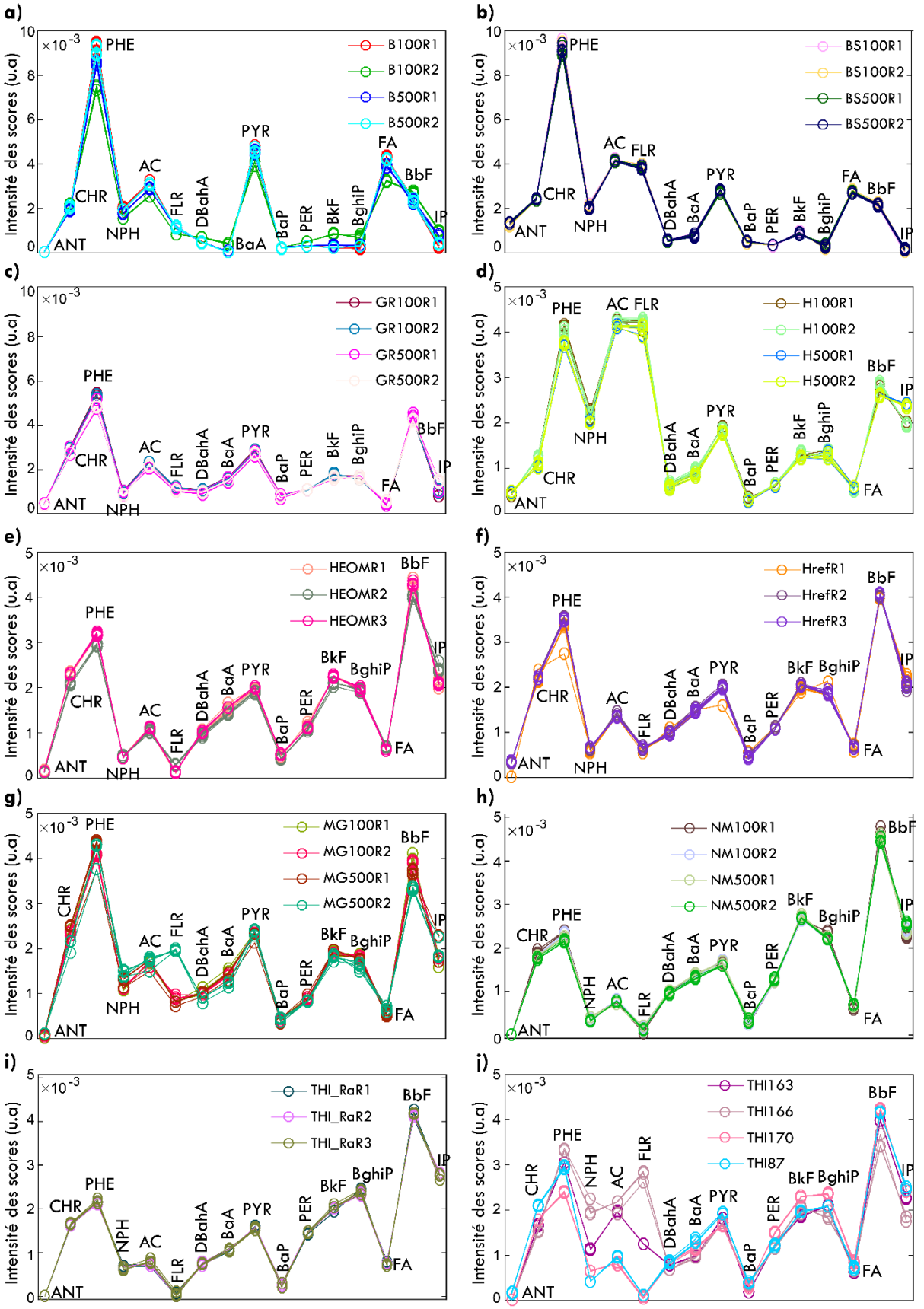


Figure IV.19 : Profil des 'scores' PARAFAC du : groupe B (a), du groupe BS (b), du groupe GR (c), du groupe H (d), du groupe HEOM (e), du groupe Href (f), du groupe MG (g), du groupe NM (h), du groupe Thi_Ra (i) et du groupe Thi (j).

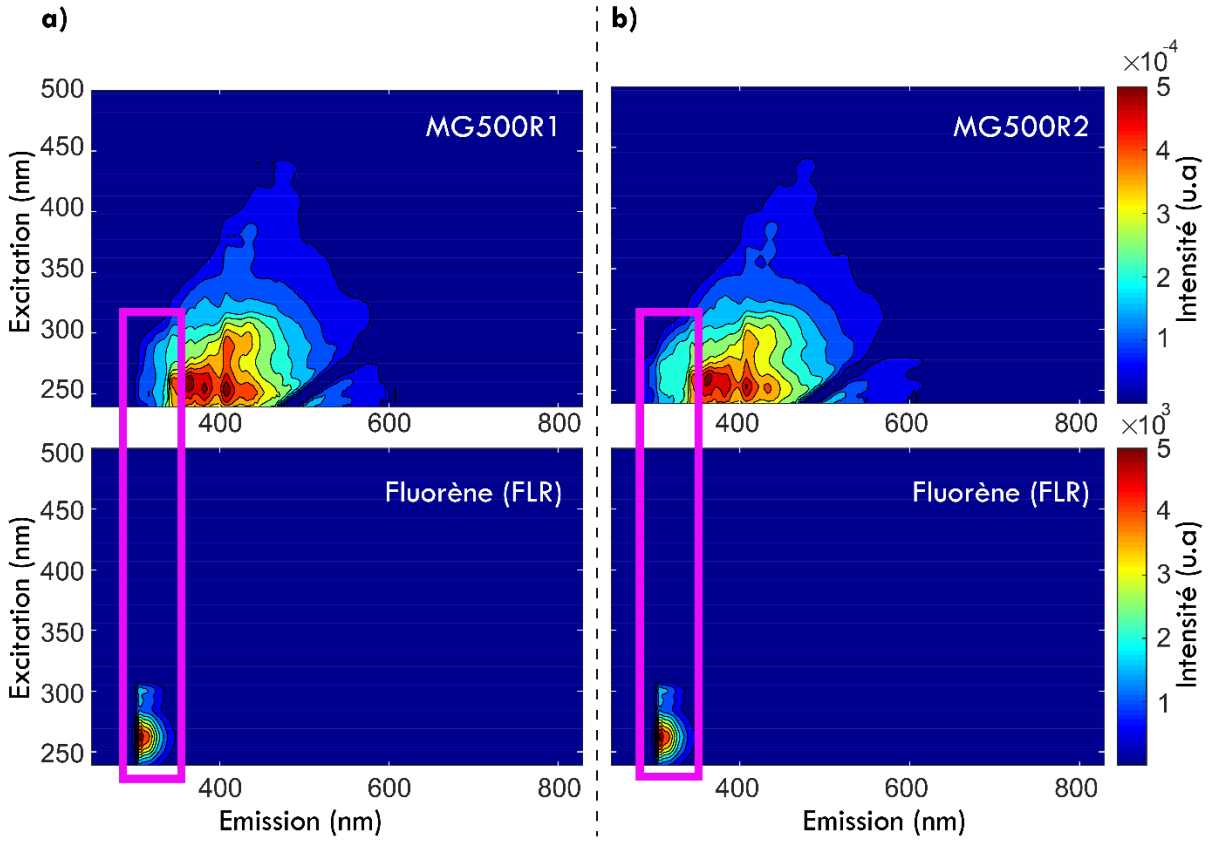
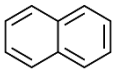
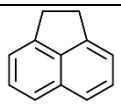
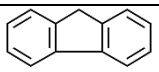
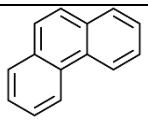
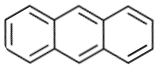
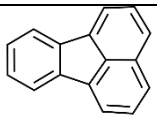
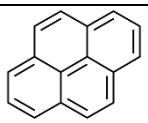
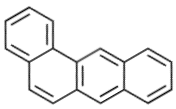
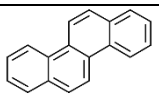
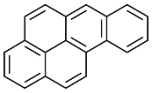
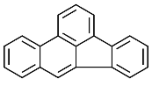
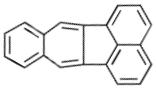


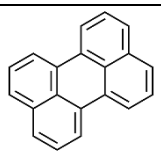
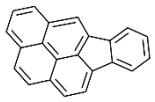
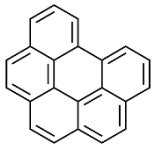
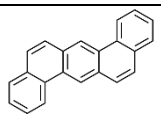
Figure IV.20 : MEEF moyenne des 5 acquisitions de la classe MG500R1 (a). MEEF moyenne des 5 acquisitions de la classe MG500R2 (b).

Grâce à cette première analyse des '*scores*' issus du modèle PARAFAC, nous constatons que chaque sol présente un profil de '*scores*' spécifique, qui est lié à l'abondance des polluants, mais aussi aux caractéristiques du signal de fluorescence. De plus, nous observons de manière générale que la taille de la maille du tamis (100 µm et 500 µm) a peu d'impact sur les profils des '*scores*'. Enfin, nous observons une hétérogénéité de réponse des classes du groupe Thi, ce qui met en avant le caractère hétérogène de la pollution du sol d'un même site. En effet, la nature d'une pollution est liée à son origine (en relation avec le type d'industrie, le type d'activités et le type de procédé utilisé), comme par exemple l'utilisation de la créosote (produit issu de la distillation du goudron de houille et riche en HAP légers) dans les sites de traitement de bois et la génération de goudron de houille comme sous-produits de cokéfaction dans les cokeries, mais aussi à son évolution et à sa préservation dans le temps (en relation avec sa migration et sa dégradation), ainsi qu'à la zone d'échantillonnage. Les HAP les plus légers sont ceux qui sont les plus susceptibles de subir en premier les mécanismes de dégradation (ex. volatilisation, biodégradation, etc.) et les mécanismes de migration (ex. lixiviation). Ainsi, un profil de pollution riche en HAP légers est un indicateur soit d'une pollution avec un produit comme la créosote, soit d'une pollution fraîche, soit d'une pollution bien préservée dans le temps, par exemple dans de petites poches de polluants. A contrario, un profil de pollution riche en HAP lourds est un indicateur d'une pollution ancienne.

Dans le cadre d'une deuxième analyse des '*scores'* issus du modèle PARAFAC, les HAP ont d'abord été triés par ordre croissant de leur poids moléculaire, et trois catégories ont été définies : HAP légers, HAP intermédiaires et HAP lourds (Tableau IV.6).

Tableau IV.6 : Catégorisation des 16 HAP de l'étude en fonction de leurs poids moléculaires. Les structures moléculaires sont issues de la base de données Chemspider (https://chemspider.com).

Catégorie	HAP	Masse molaire (g/mol)	Structure moléculaire
HAP légers	NPH	128.2	
	AC	154.2	
	FLR	166.2	
	PHE	178.2	
	ANT		
HAP intermédiaires	FA	202.3	
	PYR		
	BaA	228.3	
	CHR		
HAP lourds	BaP	252.3	
	BbF		
	BkF		

	PER		
	IP	276.3	
	BghiP		
	DBahA	278.3	

Ensuite, les '*scores'* de chaque HAP dans toutes les classes et tous les groupes sont tracés sur un seul graphique en vue de comparer l'abondance de ce dernier dans l'ensemble des échantillons (Figure IV.21 pour les **HAP légers**, Figure IV.22 pour les **HAP intermédiaires**, Figure IV.23 et Figure IV.24 pour les **HAP lourds**).

Lors de l'analyse des '*scores'* de la catégorie des **HAP légers** (Figure IV.21), nous observons que les '*scores'* les plus élevés pour le **NPH** se trouvent dans les groupes B, BS et H et dans la classe Thi 166. Le groupe MG et la classe Thi 163 présentent des '*scores'* relativement élevés pour cette HAP léger comparés aux autres sols issus de cokeries. Le groupe MG est suivi de près du groupe GR. Pour l'**AC**, les '*scores'* les plus élevés se trouvent dans les groupes B, BS et H, suivi de ceux du groupe GR et de ceux des lasses Thi 163 et Thi166. Le **FLR** présente les '*scores'* les plus élevés dans les groupes BS et H, suivis des groupes B et GR et des classes Thi 163 et Thi166, ainsi que du groupe MG et particulièrement de la classe MG500R2. Pour le **PHE**, les groupes B et BS présentent les '*scores'* les plus élevés. Enfin, pour l'**ANT**, les '*scores'* les plus élevés se trouvent dans le groupe BS, suivi des groupes H et GR.

De manière générale, des '*scores*' élevés de **HAP légers** sont observés dans les groupes B et BS comparé aux autres sites, avec une distinction notable pour l'**ANT**, qui est fortement abondant dans le groupe BS et très faible dans le groupe B. Cette pollution importante aux **HAP légers** de ces deux sites de traitement du bois toujours en exploitation peut s'expliquer par plusieurs facteurs. Tout d'abord, il s'agit de sites toujours en activité, ce qui entraîne une pollution récente et donc une abondance élevée de **HAP légers**. De plus, le type de produit utilisé pour le traitement du bois, à savoir la créosote, est principalement composé de **HAP légers**, étant issu de la distillation du goudron de houille[127]. Les différences

observées entre les deux sols peuvent être expliquées par les variations dans les procédés et les types de créosote utilisés, étant donné que l'un des sites est situé en France (B) tandis que l'autre est situé en Suède (BS).

Concernant les '*scores*' élevés des **HAP légers** du groupe H comparé aux autres sites de cokéfaction, cela peut s'expliquer par un échantillonnage qui a été réalisé dans une zone où la pollution aux **HAP légers** est bien préservée, car ce site de cokéfaction est en arrêt d'activité depuis de nombreuses décennies. Ce même fait peut expliquer l'hétérogénéité des '*scores'* du groupe Thi. En effet, les classes Thi 163 et Thi 166 ont une pollution importante aux **HAP légers** (en particulier THI166) par rapport aux deux autres classes Thi 87 et Thi 170. Pour rappel, ces 4 échantillons de sols ont été prélevés dans quatre zones différentes d'un même site (Thi) ayant abrité une cokerie.

Nous constatons aussi, lors de l'analyse des '*scores'* de la catégorie des **HAP intermédiaires** (Figure IV.22), que les '*scores'* les plus élevés pour le **FA** sont ceux des groupes B et BS. Pour le **PYR**, les '*scores'* les plus élevés se trouvent dans les groupes B, BS et GR. En ce qui concerne le **BaA**, les '*scores'* les plus élevés se trouvent dans les groupes GR, HEOM, Href, MG, NM, Thi_Ra et Thi. Les '*scores'* les plus faibles de cet HAP sont observés dans les groupes B, BS et H. À l'exception du groupe H, tous les autres groupes présentent des '*scores'* relativement élevés de **CHR**.

De manière générale, il semble exister une corrélation entre la pollution des sols de cokeries et le poids moléculaire croissant des HAP, à l'exception du sol H qui présente des '*scores*' très faibles de **CHR**. En effet, à mesure que le poids moléculaire des HAP augmente, on observe davantage de convergences, voire des dépassements, des intensités des '*scores*' des sols de cokeries par rapport aux sols B et BS. Pour ces deux groupes, la prédominance de la pollution semble d'ailleurs diminuer par rapport aux autres groupes à mesure que le poids moléculaire des HAP augmente. Cependant, il est essentiel de rappeler que les variations d'intensité des '*scores*' sont influencées non seulement par l'aspect quantitatif de la pollution, mais aussi par le rendement de fluorescence et les phénomènes physico-chimiques qui influencent la réponse en fluorescence (cf. §IV.3.3). Par conséquent, les variations quantitatives de la pollution ne peuvent être déduites que de manière comparative entre les groupes et non entre les HAP. Enfin, on observe que les '*scores'* du groupe GR semblent correspondre aux '*scores'* des cokeries. Cette similitude peut s'expliquer par la nature de la pollution de ce sol, qui provient, tout comme pour les cokeries, de la transformation de la houille conduisant à la production de goudron. Cependant, bien que le type de transformation soit de même nature que celui des cokeries, l'objectif est différent, car il s'agit de produire principalement du gaz de ville et non du coke. Ainsi, les procédés et la qualité de la houille

utilisés peuvent différer. Ces faits combinés, peuvent expliquer à la fois les similitudes et les différences entre le profil du groupe GR et les profils des cokeries (à l'exception du groupe H).

En analysant les '*scores'* de la catégorie des **HAP lourds** (Figure IV.23 et Figure IV.24), nous constatons que les '*scores'* les plus élevés pour le **BaP** se situent dans le groupe GR, suivi des groupes BS, HEOM, Href et MG. Les '*scores'* les plus faibles sont ceux du groupe B. Concernant le **BbF** et le **BkF**, ils présentent des profils similaires, avec les '*scores'* les plus élevés se trouvant dans le groupe NM, suivi de près par tous les groupes, à l'exception des groupes B et BS. Le **PER**, l'**IP** et le **BghiP** présentent également des profils similaires, avec les '*scores'* les plus élevés dans le groupe Thi_Ra, suivi de près par tous les groupes, à l'exception des groupes B et BS. Les '*scores'* les plus élevés du **DBahA** se trouvent dans les groupes GR, HEOM, Href, MG et NM, suivis de près par ceux de Thi_Ra. Les '*scores'* les plus faibles pour ce HAP sont dans les groupes B, BS et H.

En conséquence, la tendance qui a été observée avec les **HAP intermédiaires** semble se confirmer davantage avec les **HAP lourds**. Une augmentation des intensités des '*scores*' des sols de cokeries avec le poids moléculaire croissant des HAP est, en effet, observée, alors qu'une diminution intensités des '*scores*' des sols B et BS est constatée. En outre, cette analyse confirme les similitudes observées entre les sols des cokeries et le sol GR. Enfin, on remarque une variation moins prononcée au sein des classes du groupe Thi pour ces **HAP lourds**.

Pour conclure, cette deuxième analyse des '*scores'* issus du modèle PARAFAC nous a permis une caractérisation qualitative des 16 HAP cibles dans les extraits de sols. Il convient toutefois de souligner que ces observations sont spécifiques aux échantillons utilisés dans cette étude, car toute généralisation pourrait être dangereuse compte tenu de la complexité et de l'hétérogénéité de la pollution des sols, même au sein d'un même site industriel (e.g. résultats du groupe Thi).

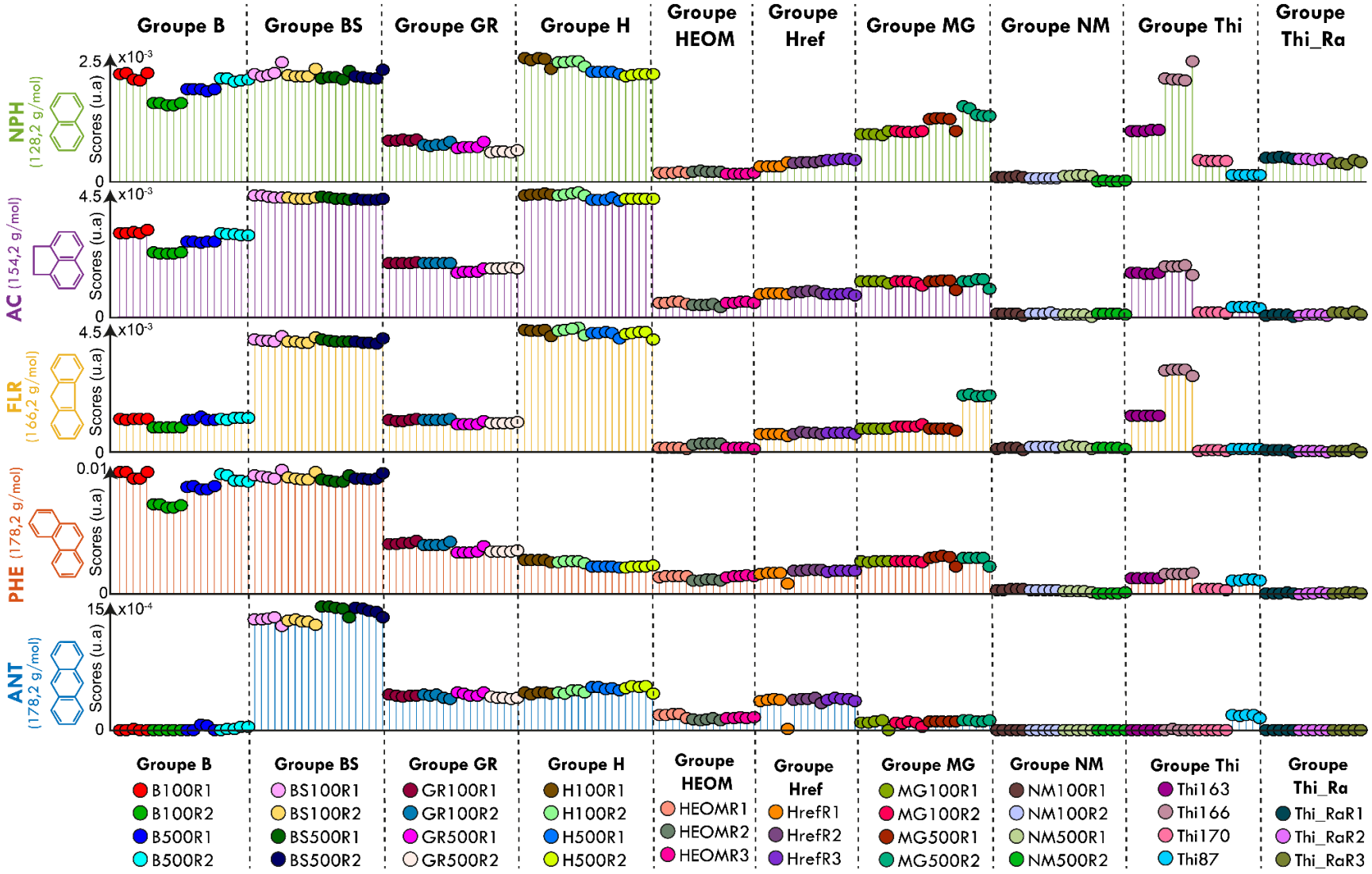


Figure IV.21 : Profils des 'scores' du NPH, AC, NPH, FLR, PHE et ANT au sein de l'ensemble des groupes de l'étude. Les structures moléculaires sont issues de la base de données Chemspider (https://chemspider.com).

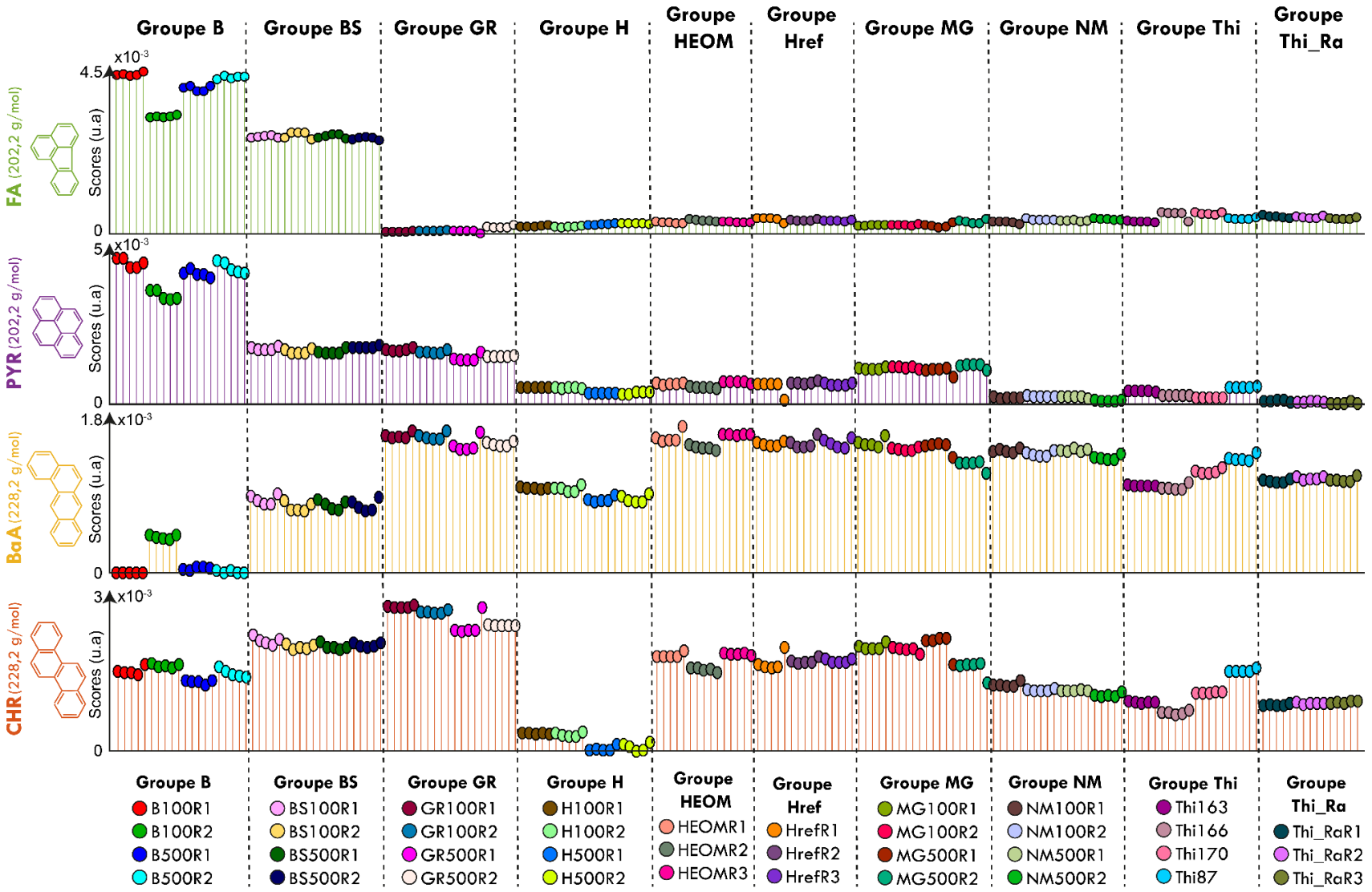


Figure IV.22 : Profils des 'scores' du FA, PYR, BaA et CHR au sein de l'ensemble des groupes de l'étude. Les structures moléculaires sont issues de la base de données Chemspider (https://chemspider.com).

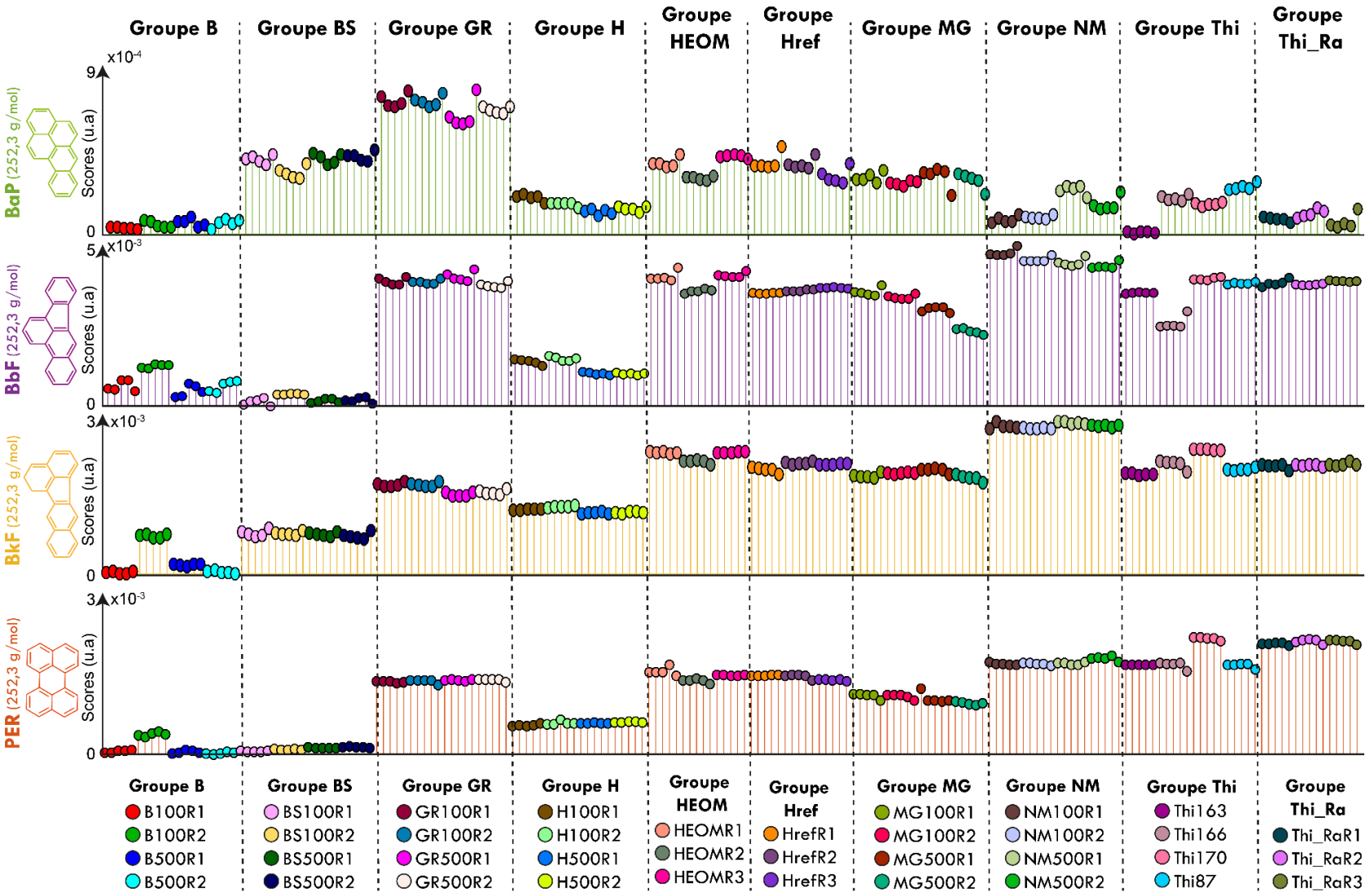


Figure IV.23 : Profils des 'scores' du BaP, BbF, BkF et PER au sein de l'ensemble des groupes de l'étude. Les structures moléculaires sont issues de la base de données Chemspider (https://chemspider.com).

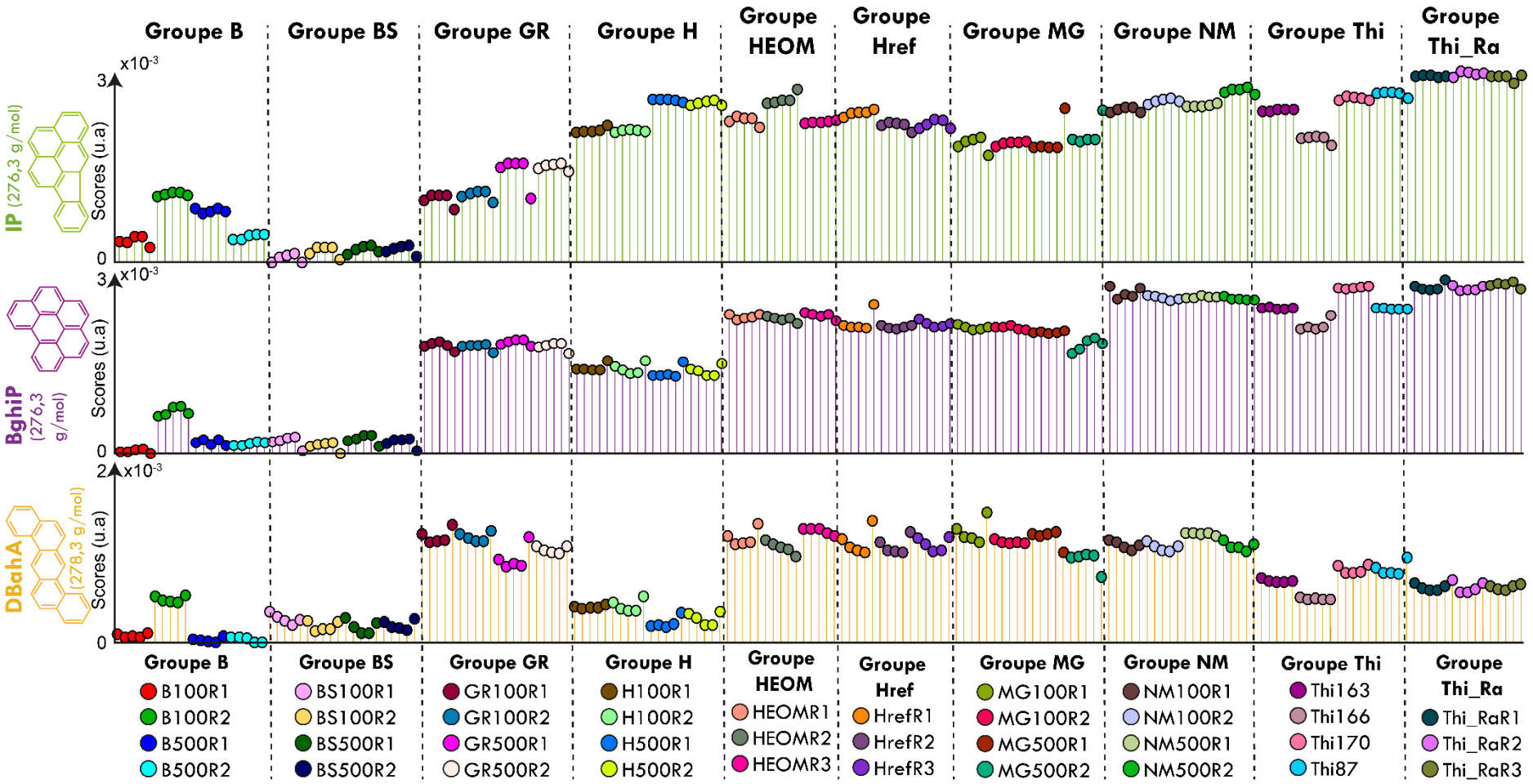


Figure IV.24 : Profils des 'scores' du IP, BghiP et DBahA au sein de l'ensemble des groupes de l'étude. Les structures moléculaires sont issues de la base de données Chemspider (https://chemspider.com).

IV.5 Conclusion

Dans ce chapitre IV, l'algorithme de décomposition trilinéaire PARAFAC a été utilisé pour identifier les 16 HAP ciblés présents dans des sols industriels contaminés. La construction et la validation d'un modèle PARAFAC à partir de MEEF de référence est, par conséquent, la clé pour identifier sans ambiguïtés des HAP dans des matrices environnementales complexes. L'analyse conjointe avec une ACPM permet quant à elle, de donner des informations sur la nature même des échantillons, et ainsi, pouvoir discuter de la nature même des sols anthropisés. En outre, l'identification des HAP ciblés dans ces sols associés à leurs poids moléculaires, à savoir HAP lourds, intermédiaires et légers, peut donner des informations sur le degré de conservation de la pollution et si l'activité est toujours en cours ou non sur le site industriel. Les résultats prometteurs issus de ce premier travail d'identification restreint à 16 HAP démontre qu'il est parfaitement envisageable de l'appliquer à de nombreux autres CAP. Le modèle PARAFAC construit précédemment, couplé à une méthodologie expérimentale adaptée, peut parfaitement être complété ou construit sur d'autres pollutions organiques en fluorescence 3D.

Cependant, nous avons constaté que l'influence du rendement de fluorescence et d'autres phénomènes physico-chimiques sur la réponse en fluorescence ne permet pas une quantification directe de cette pollution. Pour pallier à cette difficulté, le prochain chapitre, proposera de combiner nos données issues de MEEF avec des données de quantification déterminées par chromatographie en phase gazeuse couplée à la spectrométrie de masse. Des modélisations quantitatives seront effectuées pour développer des modèles prédictifs permettant de quantifier les 16 HAP à partir des données de fluorescence.

Dans un dernier temps, nous évaluerons les limitations de la spectroscopie de fluorescence dans la détection des CAP, qui va bien au-delà des HAP. Nous proposerons ainsi une ouverture vers la spectroscopie Raman, dont les spectres seront analysés par transformée en Ondelette, pour à terme, coupler les différentes informations chimiques de différentes méthodes analytiques.

CHAPITRE V

MODELISATION QUANTITATIVE DE LA POLLUTION EN HAP DE SOLS INDUSTRIELS ET INVESTIGATION AUTOUR DE LA CARACTERISATION QUALITATIVE DES CAP PAR SPECTROSCOPIE RAMAN : PERSPECTIVE D'UNE ANALYSE MULTIBLOCS

V.1 Introduction

Pour la modélisation quantitative de la pollution des sols par les HAP, trois approches sont évalués dans le cadre de ce chapitre. Tout d'abord, une régression linéaire simple est calculée entre les '*scores*' de PARAFAC obtenus dans le cadre de l'étude du chapitre IV et les concentrations réelles en HAP mesurées par GC-MS, méthode de référence pour le dosage de ces composés. Cependant, les résultats ont montré des tendances non linéaires indiquant à ce stade, un éventuel non-respect de la loi de Beer-Lambert. Cette observation était toutefois prévisible compte tenu de la complexité de la pollution des sols et de l'influence de divers phénomènes physico-chimiques, tels que le phénomène d'extinction de fluorescence, sur la réponse en fluorescence et le calcul des '*scores*' PARAFAC.

Par conséquent, des modèles de régression pour la prédiction sont développés en utilisant des méthodes d'apprentissage supervisé, notamment la N-PLS et la SVR non linéaire avec un noyau RBF (cf. §II.5), afin de modéliser les liens complexes entre les données spectrales et les concentrations de chaque HAP déterminées par GC-MS. Pour ce faire, les MEEF initiales contenues dans la base de données des MEEF des extraits de sols du chapitre IV (cf. §IV.4.3), ne sont pas utilisées en tant que données spectrales directement. Elles sont, en effet, remplacées par les MEEF modélisées après la MT-SVD et la décomposition trilinéaire PARAFAC (Figure V.1). Cette approche permet ainsi de réduire la complexité des MEEF d'origine et de se focaliser uniquement sur les réponses relatives aux 16 HAP étudiés. Les MEEF modélisées sont calculées par le produit matriciel suivant :

$$\widehat{\mathbf{X}} = \mathbf{A}(\mathbf{B} \odot \mathbf{C})^{\mathrm{T}} \qquad \text{(V.1)}$$

Avec $\widehat{\mathbf{X}}(185{,}22000)$ la matrice réorganisée du cube des MEEF modélisées par le modèle PARAFAC $\underline{\widehat{\mathbf{X}}}$ qui a pour dimensions $(185{,}250{,}88)$. $\mathbf{A}(185{,}16)$ est la matrice des '*scores*'. $\mathbf{B}(250{,}16)$ et $\mathbf{C}(88{,}16)$ sont les deux matrices de '*loadings*'. Puis, un travail d'exploration des données est réalisé afin de repérer d'éventuels '*outliers*', pour ainsi maximiser les performances des modèles subséquents. Pour ce faire, le groupe Thi est exclu de l'étude en raison de son hétérogénéité et du manque de réplicas. La classe GR100R2 est également exclue de l'étude en raison de valeurs de référence (i.e. valeurs GC-MS) anormalement plus basses que les deux autres classes du même groupe et ce, malgré une bonne reproductibilité des '*scores*' PARAFAC dans ce groupe. La classe HEOMR1 est aussi exclue en raison de valeurs de référence anormalement plus élevées que les deux autres tripliquas et ce, malgré une bonne reproductibilité des '*scores*' PARAFAC dans ce groupe. Quelques autres MEEF sont retirées de l'étude en raison d'une plus grande variation par rapport aux autres MEEF de la même classe affectant les performances des modèles. Ces

variations ont été observés après une analyse minutieuse, groupe par groupe et classe par classe, des résultats de la modélisation ACPM, exposés dans le chapitre IV (cf. §IV.4.4). Enfin, les gammes spectrales d'émission et d'excitation sont réduites uniquement aux zones de fluorescence, dans le but d'accélérer les temps de calcul des modèles. La dimension émission est réduite entre 248.276 nm et 619.957 nm et la dimension excitation entre 239 nm et 458 nm. En fin de compte, le cube des données spectrales est devenu : $\underline{\hat{\mathbf{X}}}(140,162,74)$.

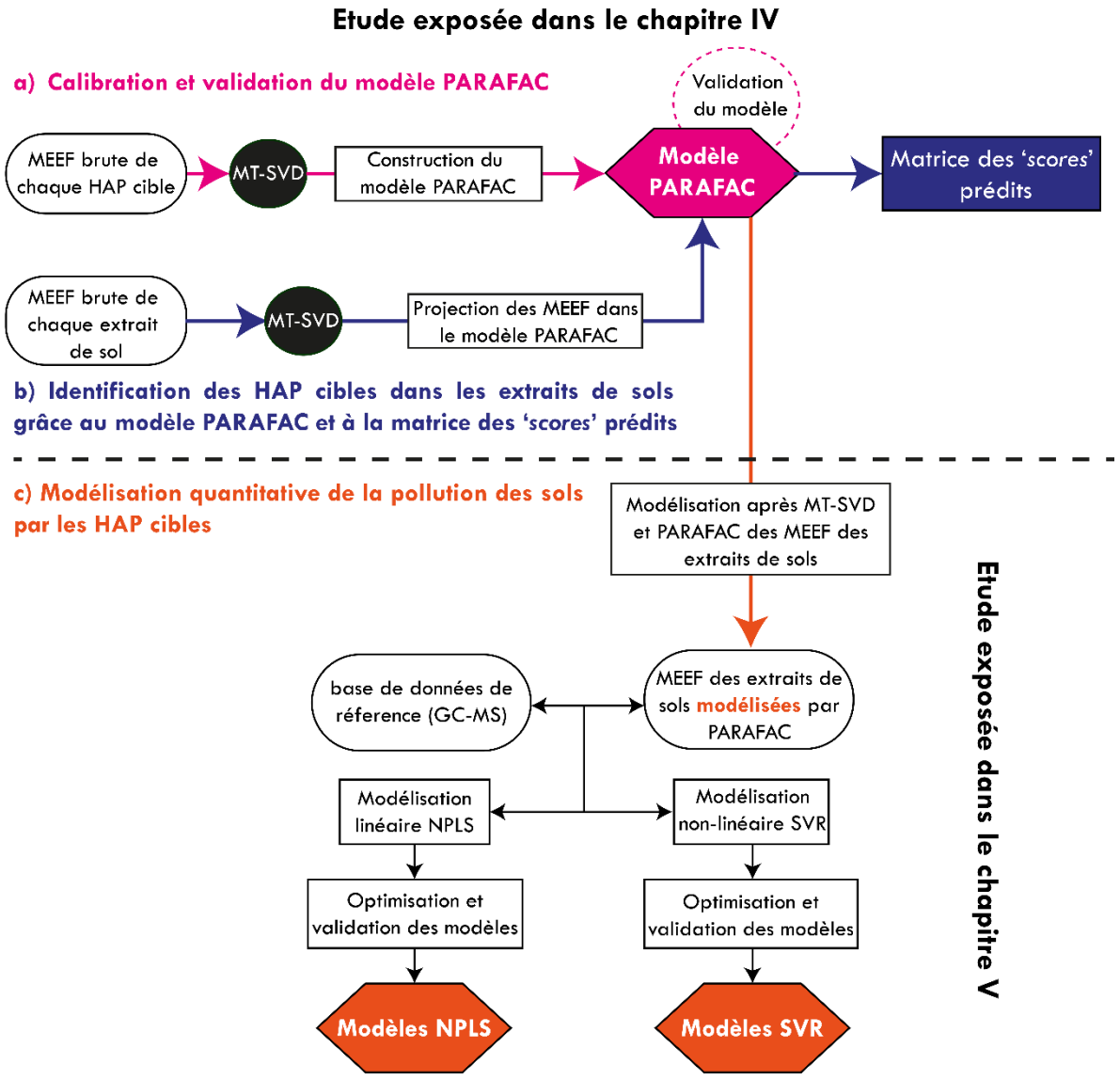


Figure V.1 : Stratégie globale de la modélisation quantitative de la pollution des sols par les HAP. de la calibration et la validation du modèle PARAFAC (a) jusqu'à la modélisation quantitative (c).

Pour calibrer et valider les modèles NPLS et SVR, la stratégie de validation croisée '*venetian blinds*'[128] est utilisée en gardant les réplicas et les tripliquas ensemble. Elle est ajustée pour construire les modèles avec une division des données en 10 ensembles. Un échantillon de chaque ensemble est utilisé dans la validation (i.e. exclu de la calibration). Pour évaluer les performances des modèles en calibration et en validation croisée, deux critères

sont utilisés : le coefficient de corrélation (R^2) et la racine de l'écart quadratique moyen '*Root Mean Square Error*' ($RMSE$)[103] (Tableau V.1).

Tableau V.1 : Critères d'évaluation des performances des modèles en calibration et en validation (RMSE et R^2).

Calibration	Validation
Coefficient de corrélation de la calibration : R^2_{cal} $R^2_{cal} = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}}$	Coefficient de corrélation de la validation croisée : R^2_{cv} $R^2_{cv} = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_{cv,i} - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_{cv,i} - \overline{y})^2}}$
$RMSE$ de calibration : $RMSEC$ $RMSEC = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$	$RMSE$ de validation croisée : $RMSECV$ $RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{cv,i} - y_i)^2}{n}}$

Avec comme notations : (i) y_i est la valeur de référence pour un échantillon i, (ii) n est le nombre d'échantillons, (iii) $\hat{y}_i$ est la valeur prédite par le modèle pour un échantillon i, (iv) $\overline{y}$ est la valeur moyenne de l'ensemble de valeurs de référence dans l'ensemble de calibration (v) et enfin $\hat{y}_{cv,i}$ est la valeur prédite par validation croisée pour un échantillon i.

L'optimisation (i.e. minimisation des erreurs) des modèles N-PLS est réalisée en ajustant le nombre de variables latentes en fonction de la $RMSEC$ et de la $RMSECV$. L'optimisation des modèles SVR non linéaires est, quant à elle, réalisée en ajustant les paramètres des modèles (i.e. ε, C et γ, cf. §II.5.3.2) en fonction de la $RMSEC$ et de la $RMSECV$.

V.2 Protocole de la quantification des HAP par GC-MS[23,127]

Tous les extraits organiques issus de sols contaminés et utilisés dans le chapitre IV ont été également analysés par GC-MS afin de quantifier la concentration de chacun des 16 HAP cibles. L'appareillage et la méthodologie de quantification des HAP est présentée dans ce qui suit :

V.2.1 Appareillage de GC-MS :

Le couplage de chromatographie en phase gazeuse-spectrométrie de masse (GC-MS) utilisé, un Agilent Technologies 7820A, est équipé d'un injecteur Split/Splitless utilisé en mode Splitless et maintenu à 320°C. Le four de chromatographie est doté d'une colonne capillaire en silice DB5-MS (phase apolaire de (5% phényl)-methylpolysiloxane) de 60 m de long, de 0.25 mm de diamètre intérieur et de 0.25 µm d'épaisseur de film. La séparation chromatographique est réalisée avec l'hélium comme gaz vecteur à un débit de 1.5 mL.min^{-1} et selon le programme de température suivant : de 70°C (maintenu 2 minutes) à 130°C à

raison de 15°C.min^{-1}, puis de 130°C à 315°C à raison de 4°C.min^{-1}, et un maintien à 315°C pendant 25 minutes. Les analyses sont effectuées par un spectromètre de masse (Agilent Technologies 5975) utilisé en mode SIM (Single Ion Monitoring) et en mode fullscan alternés. Le mode fullscan permet une détection de l'ensemble des ions et fragments formés lors de l'impact électronique (70 eV) dont le rapport masse/charge est compris entre 50 et 600 alors que le mode SIM permet de cibler les ratios m/z spécifiques aux molécules ciblées (les 16 HAP suivis dans l'étude) en augmentant la sensibilité de leur détection et donc les limites de quantification.

V.2.2 Calibration

Afin de pouvoir quantifier les CAP par GC-MS, une calibration interne de l'appareil a été préalablement réalisée. Avant injection dans le GC-MS, 20µL d'un mélange d'étalons internes (EI) de HAP deutérés comprenant le [$^2H_8$]naphtalène, [$^2H_{10}$]acénaphtène, [$^2H_{10}$]phénanthrène, [$^2H_{12}$]chrysène et le [$^2H_{12}$]périylène à une concentration de 16 µg.mL^{-1} pour chaque composé (fournisseur : Dr Ehrentorfer) est ajouté à 80µL d'un mélange de calibration contenant les 16 HAP ciblés dans cette étude et préparé à différentes concentrations . Pour chaque composé à quantifier, une courbe de calibration a été construite avec six solutions de concentration croissante (0.3, 0.9, 1.5, 3.0, 6.0 et 9.0 µg.mL^{-1}).

V.2.3 Méthode de quantification

Pour l'analyse de chaque extrait organique, 80 µL de solution sont mélangés avec 20 µL d'étalon interne (EI) de quantification (concentration finale des EI dans l'échantillon de 2.4 µg.mL^{-1}). Pour chaque série d'injection, des blancs entre chaque échantillon ainsi que des points de contrôle pour les gammes haute (3 µg.mL^{-1}) et basse (0.06 µg.mL^{-1}) sont analysés toutes les 10 injections afin de vérifier la stabilité de l'appareil. La série d'injection est validée si leur déviation est inférieure à 20%.

A noter que pour certains extraits organiques, la quantification des HAP a été réalisée en utilisant le même protocole de calibration et de quantification décrit précédemment mais sur d'autres couplage GC-MS équivalent. Le protocole d'injection peut également être légèrement différent d'un groupe d'échantillon à l'autre (concentration d'extrait injectée) pour éviter, dans le cas de concentrations élevées, une éventuelle saturation de la colonne chromatographique. Ces différences d'appareillage et de protocole n'induisent aucune différence en termes de qualité de quantification des molécules.

V.3 Modélisation NPLS des MEEF et des données de GC-MS

Pour chaque HAP cibles de l'étude, plusieurs modèles N-PLS sont construits en augmentant progressivement le nombre de variables latentes. Ensuite, le modèle qui présente les plus basses valeurs de $RMSEC$ et de $RMSECV$ est sélectionné. La plupart des modèles montrent les valeurs les plus faibles de $RMSEC$ et de $RMSECV$ pour 16 variables latentes. Le IP et le BghiP montrent les valeurs les plus faibles de $RMSEC$ et de $RMSECV$ avec 11 variables latentes. Le PYR et le FA sont des cas particuliers. En effet, pour ces HAP, deux gammes de concentrations différentes ont été considérées en raison de l'influence très forte du groupe B, qui présente des concentrations voisines de 5000 µg.g^{-1} pour le PYR et voisines de 8000 µg.g^{-1} pour le FA. Le nombre de variables latentes est fixé à 16 pour le modèle de PYR à la gamme réduite entre 0-500 µg.g^{-1} (PYR*) et à 13 pour le modèle de FA à la gamme réduite entre 0-1000 µg.g^{-1} (FA*) (Tableau V.2). Les valeurs de R^2_{cal} et R^2_{cv} obtenues sont de manière générale acceptables, car la valeur la plus basse de R^2_{cv} est observée pour le PYR dans la gamme entre 0-500 µg.g^{-1} avec $R^2_{cv} = 0.73$. La valeur la plus basse de R^2_{cal} est observée pour l'ANT avec $R^2_{cal} = 0.88$ (Tableau V.2).

Tableau V.2 : Résultats de la modélisation NPLS en termes de RMSEC, RMSECV, R^2_{cal} et R^2_{cv}. VL : variables latentes. PYR désigne le modèle du PYR à la gamme des concentration 0-500 µg.g^{-1}. FA* désigne le modèle du FA à la gamme des concentration 0-1000 µg.g^{-1}. RMSEC : racine de l'écart quadratique moyen de la calibration. RMSECV : racine de l'écart quadratique moyen de la validation croisée.*

HAP	$RMSEC$ µg.g^{-1}	$RMSECV$ µg.g^{-1}	R^2_{cal}	R^2_{cv}	Nombre VL	Gamme de concentration (µg.g^{-1})	Remarques
ANT	23.61	32.95	0.88	0.77	16	0-300	
CHR	16.88	24.26	0.92	0.84	16	0-250	
PHE	46.04	63.40	0.92	0.85	16	0-450	
NPH	07.87	11.37	0.95	0.78	16	0-70	
AC	19.79	31.43	0.97	0.89	16	0-300	
FLR	10.71	15.81	0.97	0.93	16	0-200	
DBahA	03.54	04.17	0.93	0.89	16	0-40	
BaA	23.96	31.54	0.91	0.84	16	0-350	
PYR	225.00	435.50	0.97	0.93	12	0-5000	
PYR*	36.48	75.90	0.94	0.73	16	0-500	Groupe B exclu
BaP	19.64	23.19	0.92	0.88	16	0-220	
PER	05.18	06.14	0.92	0.88	16	0-60	
BkF	11.71	14.40	0.92	0.86	16	0-140	
BghiP	09.36	10.82	0.94	0.91	11	0-110	
FA	392.50	710.60	0.97	0.90	10	0-8000	

FA*	75.07	120.77	0.97	0.87	13	0-1000	Groupe B exclu
BbF	17.76	22.94	0.93	0.88	16	0-250	
IP	11.94	13.88	0.94	0.91	11	0-140	

Cependant, les résultats obtenus en termes de $RMSEC$ et de $RMSECV$ montrent de faibles performances des modèles N-PLS. Par exemple, le modèle du DBahA, qui présentait la plus faible $RMSEC$ et $RMSECV$ avec des valeurs de 3.54 µg.g^{-1} et 4.17 µg.g^{-1}, respectivement, couvrait également la plus petite gamme de concentration allant de 0 à 40 µg.g^{-1}. Cela indique donc des erreurs relativement élevées (Tableau V.2). De plus, le nombre important de variables latentes, égale à 16 pour la majorité des modèles, suggère des modèles instables (i.e. modèles entachés d'erreurs à cause des variables latentes à haut rang et donc à fort niveau de bruit) et potentiellement non linéaires. Ainsi, une modélisation SVR non linéaire à l'aide d'un noyau RBF a été réalisée.

V.3.1 Modélisation SVR des MEEF et des données GC-MS

Pour chaque HAP de l'étude, un modèle SVR a été construit et optimisé à partir de la matrice concaténée du cube $\underline{\widehat{\mathbf{X}}}(140,162,74)$, à savoir $\widehat{\mathbf{X}}(140,11988)$ (Figure V.2). Pour les cas particuliers du PYR et du FA, un modèle SVR est construit pour chaque gamme de concentration et chaque HAP (Figure V.3). Les résultats obtenus en termes de $RMSEC$, de $RMSECV$, de R^2_{cal} et de R^2_{cv} ont démontré d'excellentes performances sur l'ensemble des modèles construits, à l'exception des modèles PYR et FA incluant le groupe B avec une gamme de concentration entre 0 et 5000 µg.g^{-1} pour le PYR et une gamme de concentration entre 0 et 8000 µg.g^{-1} pour le FA (Tableau V.3). Malgré de faibles $RMSEC$, ces modèles présentaient des $RMSECV$ relativement élevées en raison d'un biais introduit par les MEEF du groupe B, ce qui a affecté leurs performances comparativement aux autres modèles (Figure V.3).

Tableau V.3 : Résultats de la modélisation SVR en termes de RMSEC, RMSECV, R^2_{cal} et R^2_{cv}. VS : vecteurs de support. PYR désigne le modèle du PYR à la gamme des concentration 0-500 µg.g^{-1}. FA* désigne le modèle du FA à la gamme des concentration 0-1000 µg.g^{-1}. RMSEC : racine de l'écart quadratique moyen de la calibration. RMSECV : racine de l'écart quadratique moyen de la validation croisée.*

HAP	$RMSEC$ µg.g^{-1}	$RMSECV$ µg.g^{-1}	R^2_{cal}	R^2_{cv}	Nombre VS	Gamme de concentration (µg.g^{-1})	Remarques
ANT	0.6019	01.8277	1	0.99	88	0-300	
CHR	0.5534	01.4542	1	0.99	84	0-250	
PHE	1.3928	03.8121	1	0.99	85	0-450	
NPH	0.3019	01.1981	1	0.99	132	0-70	

AC	0.3778	01.4243	1	1.00	121	0-300	
FLR	0.5902	01.1911	1	0.99	127	0-200	
DBahA	0.0653	00.1932	1	1.00	126	0-40	
BaA	0.6293	01.7189	1	1.00	127	0-350	
PYR	1.3700	13.8834	1	1.00	108	0-5000	
PYR*	1.3240	03.3039	1	1.00	81	0-500	Groupe B exclu
BaP	0.2365	01.0325	1	1.00	125	0-220	
PER	0.0693	00.3234	1	1.00	122	0-60	
BkF	0.1454	00.6801	1	1.00	126	0-140	
BghiP	0.2005	00.7088	1	1.00	124	0-110	
FA	2.1698	26.5206	1	1.00	104	0-8000	
FA*	1.1899	05.9847	1	1.00	111	0-1000	Groupe B exclu
BbF	0.5875	01.2329	1	1.00	124	0-250	
IP	0.2210	00.7869	1	1.00	121	0-140	

Les valeurs élevées de R^2_{cal} et de R^2_{cv} (i.e. entre 0.99 et 1) indiquent que les modèles ont réussi à capturer la majorité de la corrélation entre les données spectrales et les données quantitatives. Les valeurs basses de la $RMSEC$ et de la $RMSECV$, proportionnellement à leurs plages de concentration correspondantes, indiquent que les modèles présentent de bonnes capacités de prédiction. En effet, les valeurs basses de la $RMSEC$ indiquent une absence de sous-apprentissage, c'est-à-dire que les modèles sont capables de bien représenter les données d'entraînement. Les valeurs basses de la $RMSECV$ indiquent une absence de sur-apprentissage, c'est-à-dire que les modèles généralisent bien aux données non utilisées lors de l'apprentissage. Le nombre de vecteurs support indique quant à lui le nombre d'échantillons ayant contribué à la construction du modèle SVR et à la définition du paramètre de la marge.

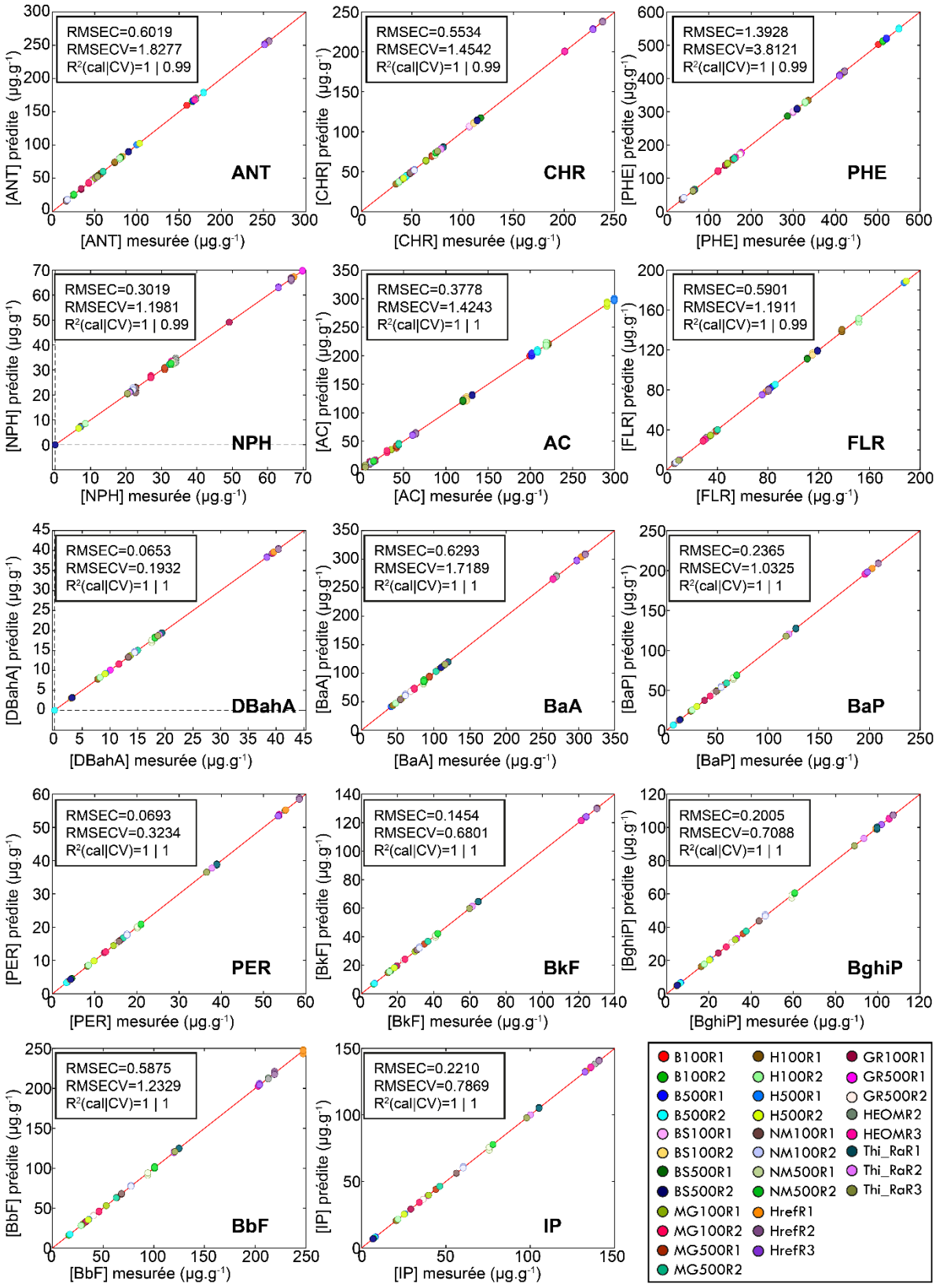


Figure V.2 : Modèles SVR des 16 HAP de l'étude à l'exception du PYR et du FA. RMSEC : racine de l'écart quadratique moyen de la calibration. RMSECV : racine de l'écart quadratique moyen de la validation croisée. Cal : calibration. CV : validation croisée.

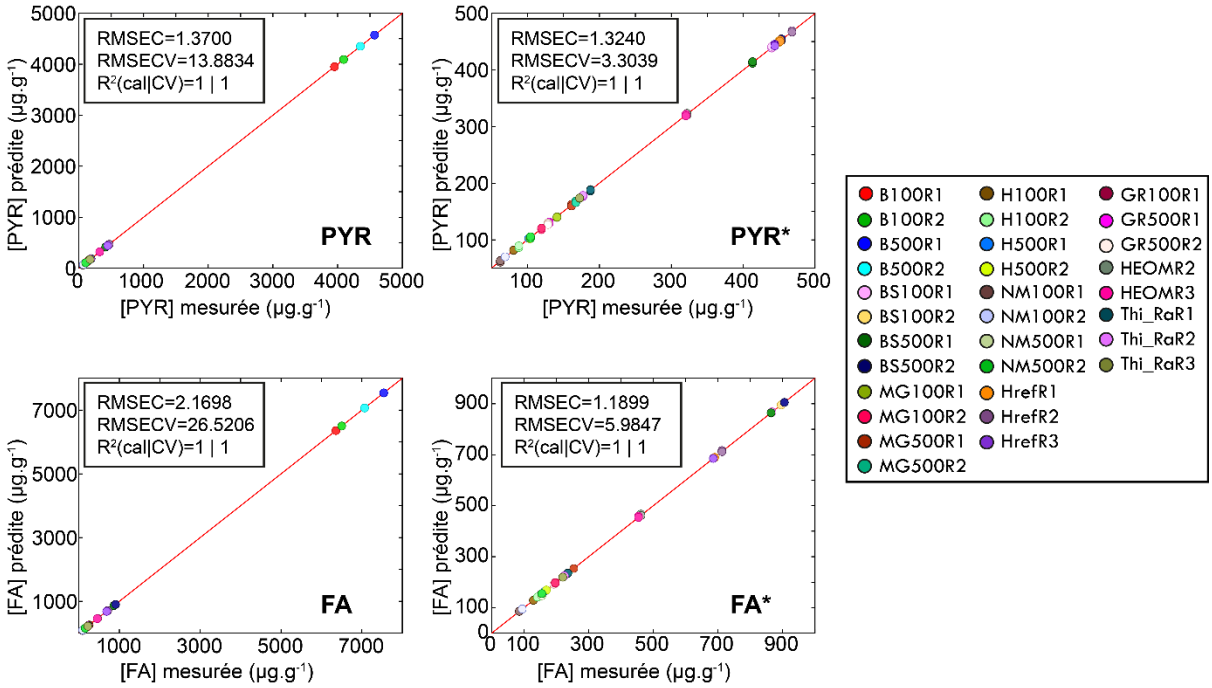


Figure V.3 : Modèles SVR du PYR et du FA. PYR désigne le modèle du PYR à la gamme des concentration 0-500 µg.g^{-1}. FA* désigne le modèle du FA à la gamme des concentration 0-1000 µg.g^{-1}. RMSEC : racine de l'écart quadratique moyen de la calibration. RMSECV : racine de l'écart quadratique moyen de la validation croisée. Cal : calibration. CV : validation croisée.*

L'optimisation de chaque modèle SVR est réalisée en ajustant les trois paramètres de la SVR, à savoir ε (i.e. taille de la marge), γ (i.e. largeur du noyau RBF) et C (i.e. paramètre de régularisation), afin d'obtenir une faible $RMSEC$ indiquant de bonnes performances du modèle et une faible $RMSECV$ indiquant une bonne généralisation du modèle et l'absence de surapprentissage. Les valeurs ε obtenues sont soit 10^{-3} soit 10^{-2}. Dans le cas où $\varepsilon = 10^{-3}$, nous avons plus de vecteurs de support que dans le cas où $\varepsilon = 10^{-2}$, car la marge est plus fine (Tableau V.4). En effet, il existe une relation inversement proportionnelle entre la taille de la marge et le nombre de vecteurs de support. Cela signifie qu'avec une marge fine, la plupart des échantillons se retrouvent à l'extérieur de cette marge et deviennent donc des vecteurs de support définissant la structure du modèle. Les valeurs γ obtenues sont soit 1×10^{-3} ou 3.17×10^{-4}. Ces valeurs reflètent le degré de non-linéarité des données, car plus γ est élevée, plus les données sont non-linéaires. Dans notre cas, γ n'est pas excessivement élevée, ce qui indique un degré de non-linéarité raisonnable et un risque moindre de surapprentissage (Tableau V.4). En effet, plus la modélisation est non-linéaire, plus le risque de surapprentissage est élevé. Les valeurs C obtenues se situent entre 3.16 et 1000. Ces valeurs varient significativement d'un modèle à un autre (Tableau V.4), car le paramètre C est un paramètre de régularisation qui établit un compromis entre la généralisation du modèle et l'erreur d'apprentissage (cf. §II.5.3). Plus la valeur de C est élevée, plus les erreurs peuvent avoir un impact sur le modèle, rendant ainsi le modèle sensible aux '*outliers*'. Cependant, les

valeurs de C obtenues ne sont pas excessivement élevées, et il convient de noter que la détection et l'élimination des '*outliers*' ont été effectuées en amont par ACPM.

Tableau V.4 : Valeurs des trois paramètres d'optimisation des modèles SVR et du nombre de vecteurs support (VS).

HAP	ε	Nombre VS	γ	C
ANT	10^{-2}	88	3.17×10^{-4}	1000.00
CHR	10^{-2}	84	3.17×10^{-4}	100.00
PHE	10^{-2}	85	3.17×10^{-4}	1000.00
NPH	10^{-3}	132	1×10^{-3}	3.16
AC	10^{-3}	121	3.17×10^{-4}	31.62
FLR	10^{-3}	127	3.17×10^{-4}	10.00
DBahA	10^{-3}	126	3.17×10^{-4}	31.62
BaA	10^{-3}	127	3.17×10^{-4}	31.62
PYR	10^{-3}	108	1×10^{-4}	1000.00
PYR*	10^{-2}	81	3.17×10^{-4}	1000.00
BaP	10^{-3}	125	3.17×10^{-4}	100.00
PER	10^{-3}	122	3.17×10^{-4}	100.00
BkF	10^{-3}	126	3.17×10^{-4}	100.00
BghiP	10^{-3}	124	3.17×10^{-4}	100.00
FA	10^{-3}	104	1×10^{-4}	1000.00
FA*	10^{-3}	111	3.17×10^{-4}	31.62
BbF	10^{-3}	124	3.17×10^{-4}	31.62
IP	10^{-3}	121	3.17×10^{-4}	100.00

V.3.2 Prédiction SVR des concentrations des 16 HAP dans les classes GR100R2 et HEOMR1

Les classes GR100R2 et HEOMR1 ont été exclues de l'ensemble de données destiné à la calibration des modèles SVR en raison de l'incohérence de leurs valeurs de référence. En effet, la classe HEOMR1 présentait des valeurs de référence anormalement élevées par rapport aux deux autres réplicas du groupe HEOM. Les valeurs de référence de la classe GR100R2 étaient quant à elles anormalement basses par rapport aux valeurs des autres classes du groupe GR. A l'inverse, les '*scores*' de PARAFAC indiquaient une bonne répétabilité entre les classes du groupe GR et entre les classes du groupe HEOM (cf. §IV.4.5). Par ailleurs, les '*scores*' de l'ACPM ne mettaient pas en évidence une variabilité inter-classe significative (cf. §IV.4.4). Par conséquent, des investigations ont été menées pour déterminer l'origine de ces incohérences et pour décider de l'utilisation ou non de ces deux classes comme ensemble de prédiction. Cela a été fait en examinant les MEEF de chacune des classes concernées en lien avec leurs '*scores*' en ACPM.

- **GR100R2**

Après un agrandissement de la zone du groupe GR dans le plan CP1-CP3, nous observons une similitude très forte entre les classes GR100R1 et GR100R2 en raison des distances réduites entre les '*scores*' des MEEF qui les composent. De plus, les classes les plus proches par la suite sont la classe GR500R1 et GR500R2. Cela indique donc une faible variabilité inter-classe au sein du groupe GR (Figure V.4a). Il est donc probable que les signatures spectrales des différentes classes soient similaires, hypothèse confirmée par l'analyse des images moyennes des 5 MEEF de chaque classe. Ces images présentent, en effet, une grande similarité en termes de longueurs d'onde d'émission et d'excitation, ainsi qu'en termes d'intensité relative (Figure V.4b). Toutes ces observations suggèrent que la variabilité significative rencontrée en GC-MS provient probablement d'un problème pré-analytique (e.g. erreur de quantité de EI ajouté), ou analytique (e.g. mauvaise injection dans le GC-MS) de la classe GR100R2, et non de l'échantillon lui-même.

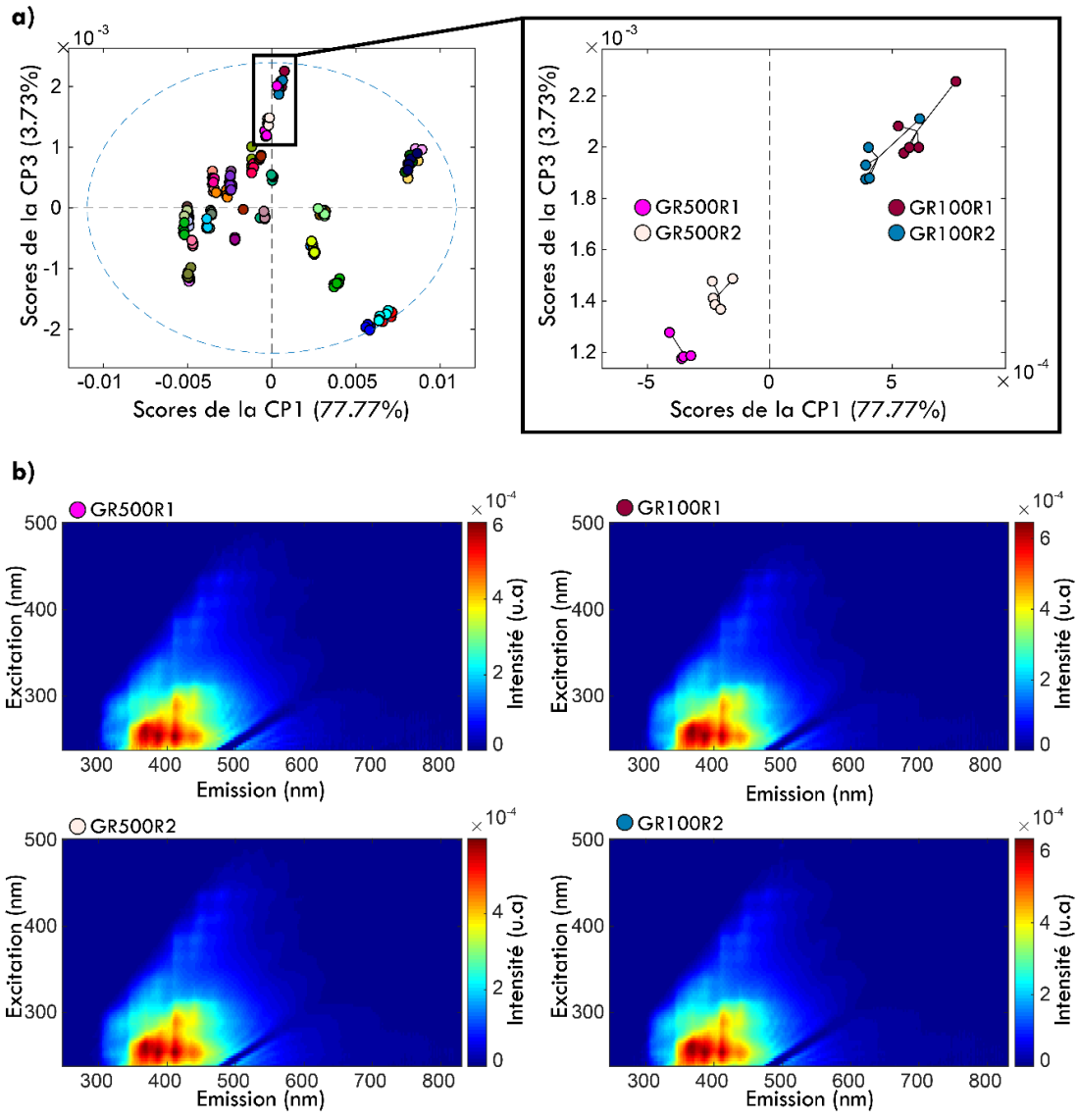


Figure V.4 : 'Scores' de 1ère composante principale versus 'scores' de la 3ème composante principale. b) image moyenne des 5 MEEF de chaque classe du groupe GR.

- **HEOMR1**

Après un agrandissement de la zone du groupe HEOM dans le plan CP1-CP3, nous observons une similitude très forte entre les classes HEOMR1 et HEOMR3 en raison des distances réduites entre les '*scores*' des MEEF qui les composent. De plus, la classe restante HEOMR2 est également proche. Cela indique donc une faible variabilité inter-classe au sein de ce groupe HEOM (Figure V.5a). Il est fort probable que les signatures spectrales de ces différentes classes soient très similaires. L'observation des images moyennes des 5 MEEF des groupes HEOMR1, HEOMR2 et HEOMR3 confirme cette hypothèse. Ces images présentent, en effet, une grande similarité en termes de signature de fluorescence mais également en termes d'intensités relatives (Figure V.5b). Toutes ces observations prouvent que la variabilité rencontrée en GC-MS est significative et est dépendante d'une erreur provenant des mêmes problèmes décrits précédemment pour la classe GR100R2, et non du signal de fluorescence ou de l'extraction associé.

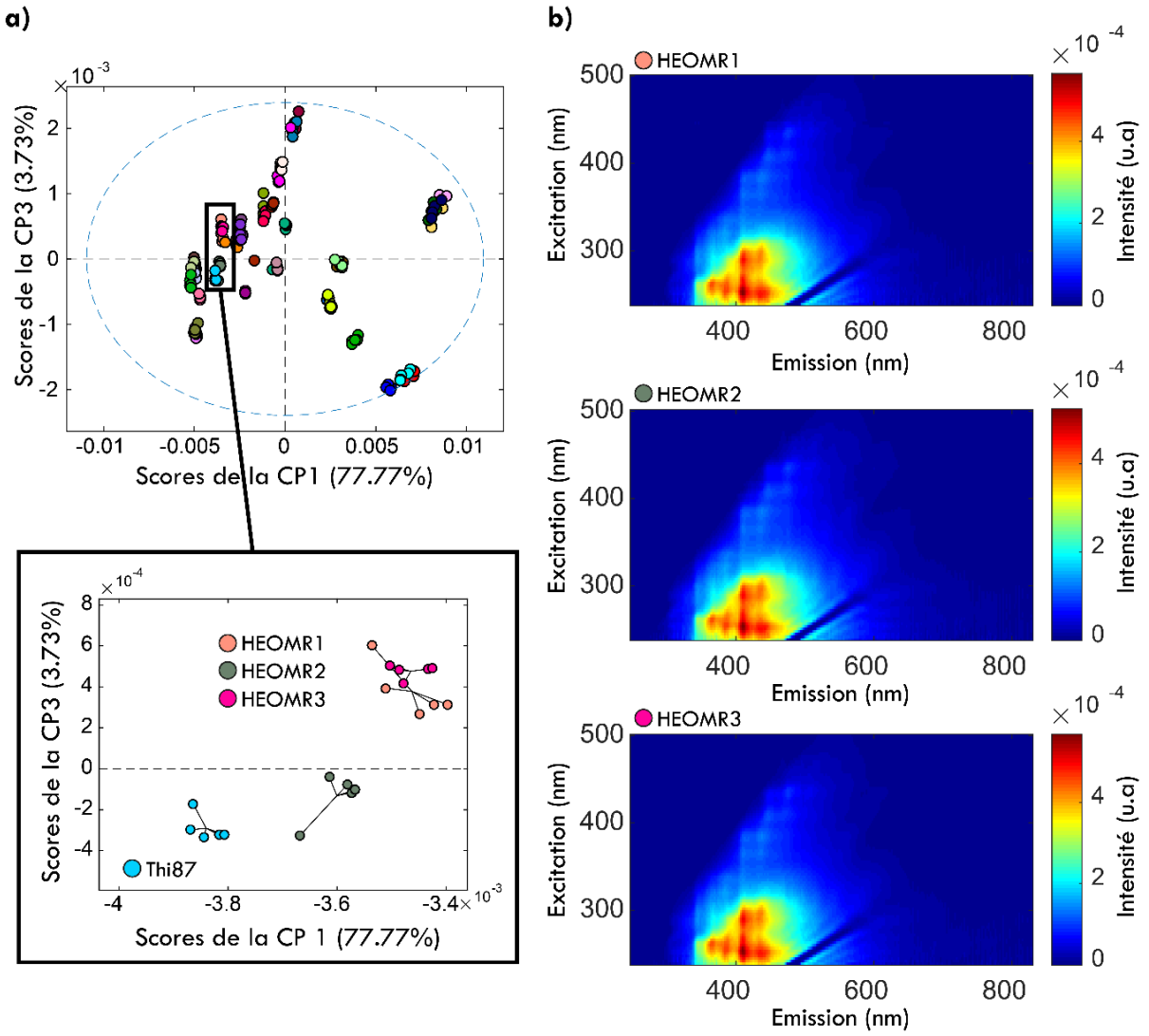


Figure V.5 : 'Scores' de 1ère composante principale versus 'scores' de la 3ème composante principale. b) image moyenne des 5 MEEF de chaque classe du groupe HEOM.

Suite à la confirmation que les incohérences rencontrées proviennent des données de référence et non des échantillons ou du rendement de l'extraction, l'ensemble des données composé des 5 MEEF de la classe GR100R2 et des 5 MEEF de la classe HEOMR1 a été considéré comme un ensemble de prédiction. Cet ensemble de données a été fourni aux 16 modèles SVR pour la prédiction de la concentration des 16 HAP dans ces extraits de sols. Les valeurs prédites ont été comparées à la valeur de référence de la GR100R1 pour la classe GR100R2, et à la valeur moyenne des classes HEOMR2 et HEOMR3 pour la classe HEOMR1.

Les valeurs prédites en concentration obtenues par SVR, ainsi que les écarts entre ces valeurs et celles de références, sont satisfaisantes. L'écart le plus important est, en effet, de 6.42 $\mu g.g^{-1}$ et concerne la prédiction de la concentration du PHE dans la classe HEOMR1. Si nous regardons de plus près, la valeur prédite est de 168.93 $\mu g.g^{-1}$, alors que la valeur de référence est de 162.51 $\mu g.g^{-1}$. La meilleure performance est observée pour la prédiction de la concentration en PER dans la classe GR100R2, avec un écart de seulement 0.03 $\mu g.g^{-1}$. La valeur prédite corresponds à 12.34 $\mu g.g^{-1}$, tandis que la valeur de référence est de 12.31 $\mu g.g^{-1}$ (Tableau V.5).

*Tableau V.5 : Résultats de la prédiction des concentrations des 16 HAP dans les classes MG100R1 et HEOMR1. *Moyenne des valeurs prédites pour les 5 MEEF de chaque classe. ** Pour la classe GR100R2 : valeur de référence de la GR100R1 obtenue par GC-MS. Pour la classe HEOMR1 : moyenne des valeurs de références des classes HEOMR2 et HEOMR3 obtenues par GC-MS. *** valeur absolue de l'écart entre la moyenne des valeurs prédites et la valeur de référence.*

HAP	Classe	Moyenne des valeurs prédites* $\mu g.g^{-1}$	Valeur de référence** $\mu g.g^{-1}$	Valeur absolue de l'écart*** $\mu g.g^{-1}$
ANT	GR100R2	35.00	33.99	1.01
	HEOMR1	168.75	169.14	0.38
CHR	GR100R2	40.75	39.80	0.95
	HEOMR1	201.26	200.64	0.62
PHE	GR100R2	143.08	138.84	4.24
	HEOMR1	168.93	162.51	6.42
NPH	GR100R2	49.24	49.13	0.10
	HEOMR1	35.74	33.42	2.32
AC	GR100R2	10.47	9.87	0.60
	HEOMR1	16.89	16.09	0.79
FLR	GR100R2	31.30	30.47	0.83
	HEOMR1	31.51	30.81	0.69
DBahA	GR100R2	7.90	7.81	0.09
	HEOMR1	39.70	39.35	0.34
BaA	GR100R2	55.65	54.57	1.08
	HEOMR1	265.79	267.01	1.21
PYR*	GR100R2	106.61	103.35	3.25
	HEOMR1	324.26	321.10	3.16
BaP	GR100R2	37.68	37.35	0.32
	HEOMR1	196.01	198.96	2.95
PER	GR100R2	12.34	12.31	0.03
	HEOMR1	55.17	54.38	0.79
BkF	GR100R2	19.92	19.57	0.35
	HEOMR1	121.18	122.57	1.38

BghiP	GR100R2	24.75	24.45	0.29
	HEOMR1	105.76	105.41	0.35
FA*	GR100R2	155.79	152.39	3.39
	HEOMR1	458.05	457.83	0.21
BbF	GR100R2	33.43	32.60	0.82
	HEOMR1	202.85	207.99	5.13
IP	GR100R2	29.56	29.24	0.31
	HEOMR1	136.92	137.06	0.14

Cette première partie du chapitre démontre la pertinence des outils de modélisation quantitative pour la construction de modèles mathématiques de régression capables d'exploiter la complémentarité entre des données de fluorescence 3D, qui constituent des signatures spectrales de la pollution aux 16 HAP des échantillons de sols, et des données GC-MS, qui fournissent une quantification de celle-ci. L'étude présentée jusqu'à présent se concentre uniquement sur les 16 HAP cibles et est complémentaire aux travaux réalisés sur leur identification (cf. §IV). Cependant, la pollution des sols ne se limite pas seulement à ces 16 HAP, car il existe un large éventail de molécules de type CAP, associées à leur contamination, telles que les dérivées des HAP, les CAP hétérocycliques et leurs dérivées qui sont tout aussi, voire plus toxique[7]. Malheureusement, la présence d'un hétéroatome au sein d'un des cycles aromatiques d'un CAP ou dans son groupement fonctionnel de substitution entraîne souvent une diminution du rendement de fluorescence, ce qui rend leur détection difficile en fluorescence 3D. Dans la section suivante, différents CAP pour lesquels une réponse en fluorescence est limitée, seront présentés.

V.4 Limitations de la spectroscopie de fluorescence dans l'analyse des CAP

Pour illustrer les limites de la spectroscopie de fluorescence dans la caractérisation des CAP, la réponse en fluorescence de plusieurs CAP a été évaluée. Nous avons ainsi sélectionné 4 dérivés de HAP, dont 3 céto-HAP (Fluorènone, Benzofluorènone et Benzanthrone) et 1 nitro-HAP (Nitropyrène), ainsi que 2 S-CAP (Benzothiophène et Dibenzothiophène) avec 2 de leurs dérivés alkylés (Methyl-Benzothiophène et Methyl-Dibenzothiophène), et enfin 2 N-CAP (Quinoléine et Benzocarbazole). La MEEF optimale pour chaque molécule a été recherchée et identifiée. Elle dépend à la fois de la concentration et du rendement de fluorescence de chaque molécule. Il s'agit de la MEEF qui présente le meilleur rapport signal sur bruit, au-delà duquel le signal sature (Figure V.6). La même instrumentation

et les mêmes conditions expérimentales que celles décrites dans le chapitre III et IV sont utilisées (cf. §III.3.1).

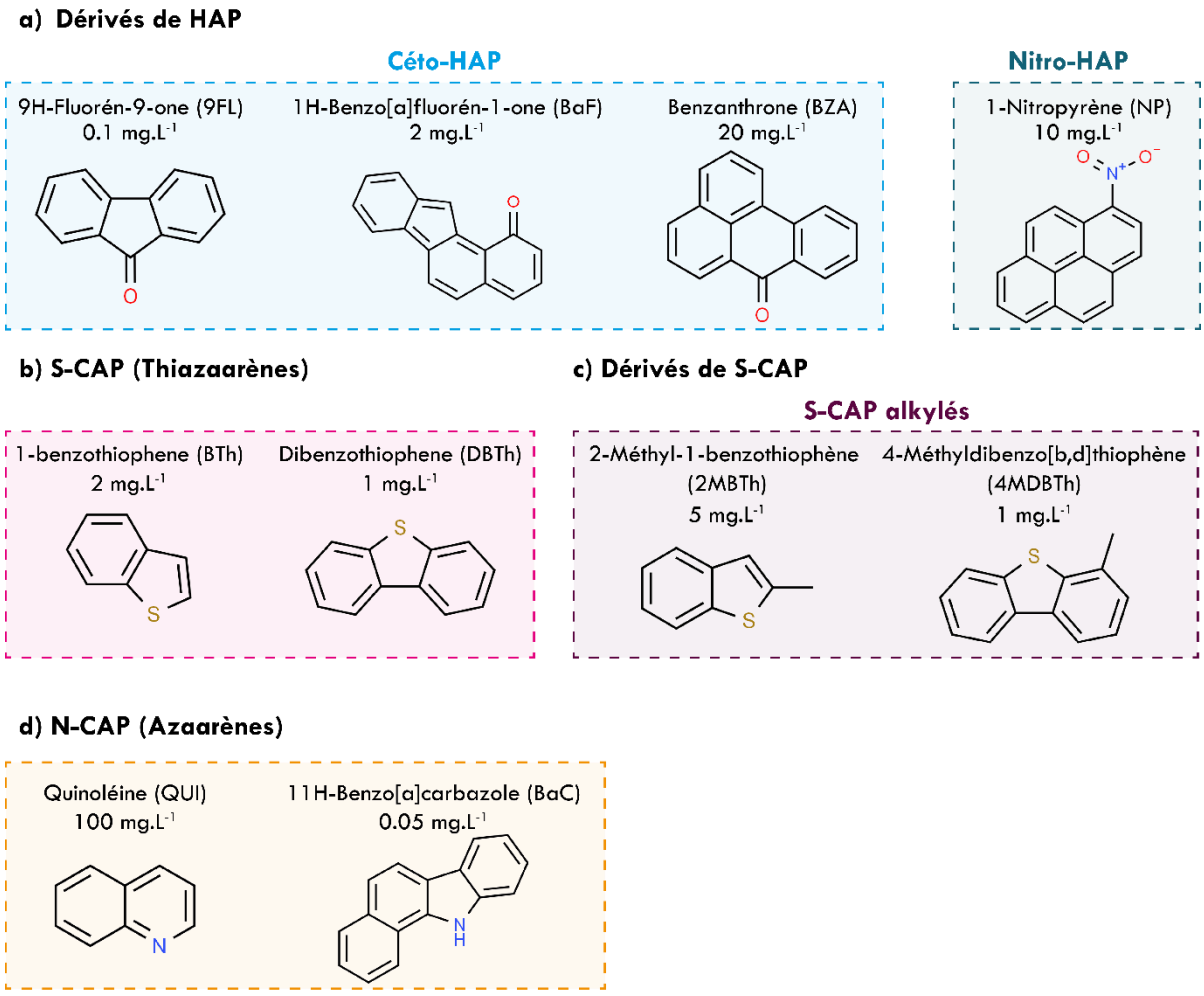


Figure V.6 : Structures chimiques des 10 CAP utilisés pour l'étude des limités de la spectroscopie de fluorescence dans la caractérisation des CAP.

La réponse en fluorescence des CAP sélectionnés est très variable. Ci-dessous sont discutés les résultats obtenus en fonction de la catégorie ou de la sous-catégorie du CAP.

(a) Dérivés de HAP

Concernant les **Céto-HAP** (Figure V.7a), le **9FL** présente une réponse en fluorescence acceptable (i.e. intensité maximale supérieure à 5000 u.a) à une faible concentration de 0.1 mg.L^{-1}. En revanche, le **BaF** nécessite une concentration relativement élevée de 2 mg.L^{-1} pour obtenir une réponse en fluorescence acceptable (i.e. intensité maximale supérieure à 12000 u.a). Ces deux molécules se caractérisent par une émission à des longueurs d'onde relativement élevées. Le **9FL** émet intensément entre 450 et 650 nm, tandis que le **BaF** émet intensément entre 500 et 700 nm. La réponse en fluorescence du **BZA** est médiocre, car il a fallu utiliser une concentration élevée (i.e. 20 mg.L^{-1}) pour obtenir une réponse acceptable (i.e. intensité maximale supérieure à 12000 u.a). De plus, deux réponses distinctes correspondant

à la présence d'un monomère (émission voisine de 365 nm) et d'un excimère (émission voisine de 453 nm) sont observées (Figure V.7a). Ce fait est confirmée par la MT-SVD, qui détecte un rang global de 2. Cette variabilité dans la réponse de fluorescence est caractéristique des CAP substitués par un groupement carbonyle (cf. §I.7.3.1). Le **NP**, qui est un **nitro-HAP** (Figure V.7b), présente une réponse en fluorescence médiocre, car il a été nécessaire d'utiliser une concentration élevée de 10 $mg.L^{-1}$ pour obtenir une émission acceptable, à savoir, une intensité maximale supérieure à 5000 u.a. De plus, deux réponses distinctes correspondant à la présence d'un monomère avec une émission aux alentours de 400 nm, et d'un excimère avec émission aux alentours de 500 nm sont observées (Figure V.7b). Ce fait est confirmée par la MT-SVD, qui détecte un rang global de 2.

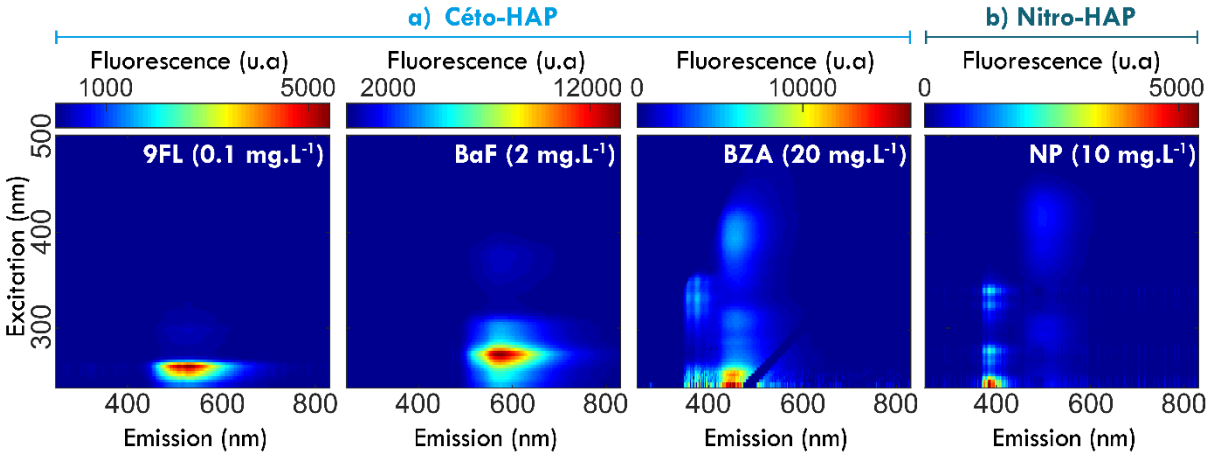


Figure V.7 : MEEF, prétraitées par MT-SVD, des quatre dérivés de HAP : les céto-HAP (a) et le nitro-HAP (b).

(b) Les Thiazaarènes (S-CAP) et leurs dérivés alkylés

Le **BTh** et le **DBTh**, qui sont des **S-CAP** dotés de cycles thiophènes (i.e. liaison σ de l'hétéroatome S aux atomes de carbone, cf. §I.7.2 et Figure V.6), présentent des émissions de fluorescence acceptables, bien que celle du **DBTh** soit meilleure. En effet, on observe pour ce dernier, une intensité de fluorescence maximale supérieure à 100 000 u.a à une concentration de 1 $mg.L^{-1}$, tandis que l'intensité du **BTh** est d'environ 50 000 u.a à une concentration deux fois plus élevée, soit 2 $mg.L^{-1}$. Ces deux composés émettent de la fluorescence à des longueurs d'ondes relativement basses. Le **BTh** émet entre 300 et 340 nm, tandis que le **DBTh** émet entre 330 et 360 nm (Figure V.8a). Les signatures spectrales du **2MBTh** et du **4MBTh**, qui sont des dérivés alkylés des deux S-CAP cités ci-dessus, ressemblent beaucoup à celles de leurs molécules parentes, c'est-à-dire le **BTh** et le **DBTh** respectivement. Cette similarité se manifeste également dans le rendement de fluorescence, où le **4MDBTh** présente une réponse en fluorescence nettement meilleure que celle du **2MBTh**. En effet, il est nécessaire d'utiliser une concentration élevée (i.e. 5 $mg.L^{-1}$) de **2MBTh** pour obtenir une réponse de fluorescence satisfaisante (i.e. intensité maximale supérieure à

2×10^4 u.a), tandis que le **4MBTh** montre une réponse de fluorescence très satisfaisante (i.e. intensité maximale supérieure à 12×10^4 u.a) à 1 mg.L^{-1} (Figure V.8b).

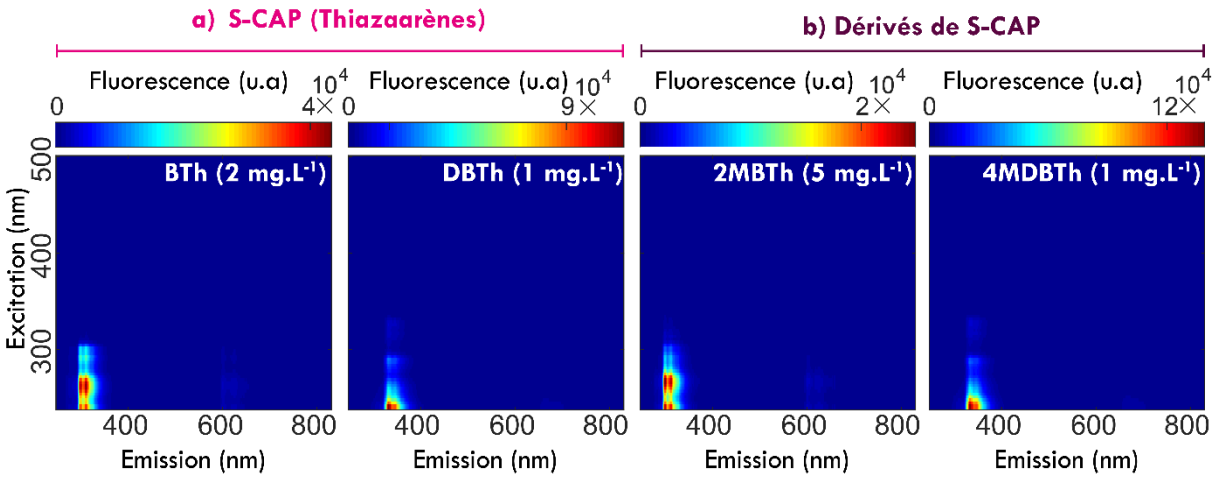


Figure V.8 : MEEF, prétraitées par MT-SVD, de deux S-CAP (a) et de deux de leurs dérivés alkylés (b).

(c) Les Azaarènes (N-CAP)

Les deux **N-CAP** de l'étude, à savoir la **QUI** et le **BaC,** présentent des réponses en fluorescence complétement différentes. Le **BaC** se caractérise par une excellente émission de fluorescence (i.e. intensité maximale supérieure à 6×10^4 u.a) à une faible concentration de 0,05 mg.L^{-1}, tandis que pour obtenir une réponse de fluorescence de la **QUI**, il a été nécessaire d'utiliser une concentration extrêmement élevée de 100 mg.L^{-1} (Figure V.9). Ces observations étaient attendues compte tenu de la nature de la liaison chimique reliant l'hétéroatome d'azote au carbone d'un des cycles de la molécule (cf. §I.7.2). Dans le cas du **BaC**, l'azote est lié avec une liaison σ à l'atome de carbone, ce qui induit une excellente réponse en fluorescence, contrairement à l'azote de la **QUI** qui est liée avec une liaison π (Figure V.6).

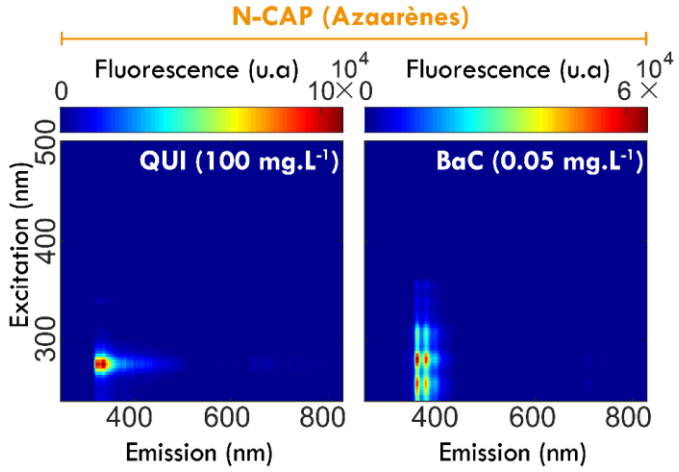


Figure V.9 : MEEF, prétraitées par MT-SVD, des deux N-CAP de l'étude.

Ces résultats démontrent la difficulté d'une caractérisation exhaustive de la contamination des sols par les CAP à l'aide uniquement de la spectroscopie de fluorescence.

Pour tenter de repousser cet inconvénient, il est nécessaire d'utiliser d'autres techniques analytiques complémentaires afin d'obtenir une caractérisation plus approfondie de la pollution aux CAP des sols contaminés. Nous avons ainsi retenu la spectroscopie Raman qui présente de nombreux avantages, en particulier sa capacité à fournir des informations sur la structuration moléculaire du système chimique étudié. De plus, cette spectroscopie est moins spécifique que la spectroscopie de fluorescence, car elle ne se limite pas aux composés fluorescents. C'est pourquoi, 27 CAP, incluant les 16 HAP de la liste de l'US-EPA[6] mais également des CAP polaires, ont été étudiés par spectroscopie Raman. Dans un premier temps nous allons démontrer la faisabilité et l'adaptabilité de cette technique pour la détection de ces 27 CAP, et dans un deuxième temps, étudier la réponse Raman des extraits organiques issus de sols pollués aux CAP.

Cette étude servira de point de départ à d'autres recherches plus approfondies qui pourront être menées à la suite de ce travail de thèse. En effet, ce travail ouvre la perspective d'une caractérisation plus approfondie de la pollution des sols par les CAP mais également la perspective de l'utilisation de stratégies de chimiométrie de type multi-blocs (i.e. analyse simultanée, pour les mêmes échantillons, de plusieurs blocs de données issus de différentes techniques analytiques).

V.5 Description par spectroscopie Raman de la pollution des sols par les CAP

Pour rappel, la spectroscopie Raman est une spectroscopie électromagnétique basée sur le principe de la diffusion inélastique de la lumière par un échantillon irradié avec une onde électromagnétique. Cette diffusion, appelé diffusion Raman, est la conséquence d'une modification de la polarisation de la molécule par le champ électrique du rayonnement électromagnétique incident[27]. La spectroscopie Raman permet d'obtenir des informations sur la structure moléculaire et la composition chimique d'un échantillon en étudiant ces vibrations moléculaires. Lors de la diffusion Raman, les photons sont diffusés avec des changements d'énergie correspondant aux transitions vibrationnelles des molécules. Il existe deux types de diffusion Raman. D'une part, il y a la diffusion Raman Stokes, où l'énergie des photons diffusés est plus faible que celle des photons incidents (i.e. perte d'énergie). D'autre part, il y a la diffusion Raman anti-Stokes, où l'énergie des photons diffusés est plus élevée que celle des photons incidents (i.e. gain d'énergie). En pratique, le signal Raman Stokes est plus intense que le signal Raman anti-Stokes (Figure V.10).

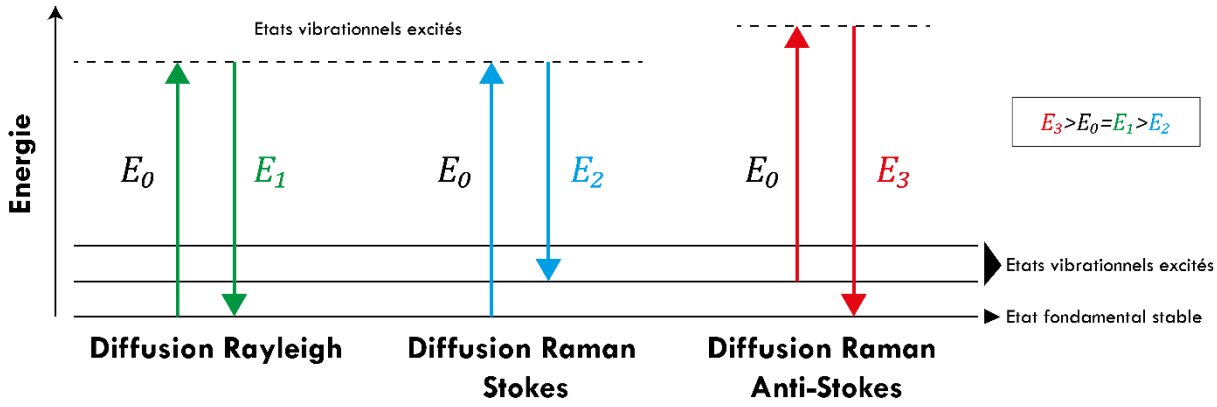


Figure V.10 : Diagramme de Perrin-Jablonski illustrant les trois proccessus de diffusion de la lumière par une molécule[48]. E_0 est l'énergie incidente.

V.5.1 Construction de la base de données des CAP et des extraits de sols

Pour la première partie de l'étude Raman, 27 solutions de différentes concentrations ont été préparées dans du dichlorométhane et un total de 27 spectres Raman ont été acquis. Cela comprend un spectre pour chaque HAP parmi les 16 HAP de l'étude précédente, le spectre de l'acénaphtylène (ACY), ainsi qu'un spectre pour chaque CAP parmi les 10 CAP étudiés lors du travail sur les limitations de la spectroscopie de fluorescence (Tableau V.6). Pour la deuxième partie de cette étude, un spectre Raman a été acquis pour chacun des 37 extraits de sols détaillés dans le chapitre IV (cf. §IV.4.1).

Les expérimentations Raman ont été menées au Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement (LASIRE, CNRS UMR 8516) à Lille, en France. Les spectres ont été enregistrés à l'aide d'un micro-spectromètre Raman HR 800 mm de la société Horiba Scientific. Un laser UV d'une longueur d'onde de 266 nm avec une intensité de 20 mW a été utilisé. L'objectif choisi était un X80, adapté pour les longueurs d'onde UV. Chaque mesure a été réalisée avec un temps d'exposition de 20 secondes et a été accumulée 3 fois. Le réseau utilisé avait une densité de 2400 lignes par millimètre. Le détecteur utilisé était un détecteur Symphony CCD de 2048 × 512 pixels. La plage spectrale couvrait de 200 à 4000 cm^{-1}, avec une résolution spectrale de 1.597 cm^{-1}.

Tableau V.6 : Base de données des 27 CAP utilisés dans le cadre de la première partie de l'étude Raman.

Catégorie	Sous-catégorie	Abréviation du nom	Concentration $mg.mL^{-1}$
HAP et leurs dérivés	HAP	ANT	1.61
		CHR	0.29
		PHE	0.75
		NPH	1.92
		AC	0.61
		ACY	0.17

		FLR	0.92
		DBahA	0.12
		BaA	1.38
		PYR	1.49
		BaP	0.46
		PER	1.11
		BkF	0.10
		BghiP	0.24
		FA	0.32
		BbF	0.06
		IP	0.24
	Céto-HAP	9FL	1.85
		BaF	0.18
		BZA	0.47
	Nitro-HAP	NP	0.05
S-CAP (Thiazaarènes) et leurs dérivés	S-CAP	BTh	0.55
		DBTh	0.28
	S-CAP alkylés	2MBTh	0.33
		4MDBTh	0.54
N-CAP (Azaarènes)	/	QUI	1.85
		BaC	0.32

V.5.2 Etude de la réponse Raman des 27 CAP purs

V.5.2.1 Correction des spectres Raman

Durant les mesures Raman, l'efficacité du flux de photons entre l'échantillon et l'instrument peut varier, en lien notamment avec les variations de la puissance du laser, l'accumulation de matière sur les dispositifs de mesure, ou encore les variations d'opacité de l'échantillon. Ainsi, les variations du flux de photons entraînent des effets multiplicatifs[129] sur les spectres acquis. En pratique, ces effets se traduisent sur un spectre Raman par une déviation non constante, souvent croissante de la ligne de base à mesure que les décalages Raman augmentent. De plus, l'émission de fluorescence, un processus concurrent souvent observé sur les spectres Raman, est plus intense que la diffusion Raman qui est intrinsèquement faible[129]. En pratique, les signaux de fluorescence induisent dans un spectre Raman au mieux une déviation de la ligne de base et au pire un recouvrement complet des signaux d'intérêt. Dans notre cas, les 27 CAP de l'étude présentent une réponse correcte en Raman aux concentrations initialement choisies, avec un bon rapport signal sur bruit et des intensités de l'ordre de 10^4 u.a. Compte tenu de la longueur d'onde de la source excitatrice (i.e. 266 nm), le phénomène de fluorescence a pu être évité et ces effets ont été très peu observés. Néanmoins, des phénomènes d'échauffement thermique relativement faibles ont été observés sur certains spectres Raman d'échantillons, en raison de l'énergie élevée de la

source excitatrice. Ainsi, ce phénomène combiné à d'autres interférences se traduit par une déviation de la ligne de base observable sur les spectres Raman bruts (i.e. effet multiplicatif) (Figure V.11).

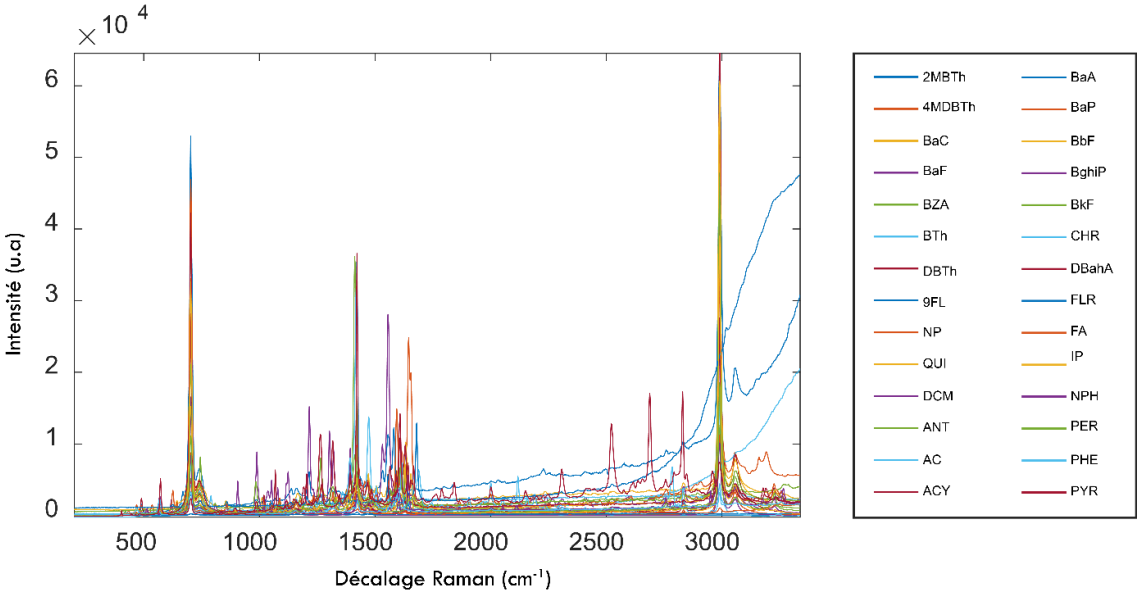


Figure V.11 : Spectres Raman bruts des 27 CAP de l'étude.

Par conséquent, des prétraitements de chimiométrie ont été utilisés pour corriger ces effets et permettre l'accès aux informations chimiques pertinentes. Tout d'abord, la correction de la ligne de base a été réalisée au moyen de l'algorithme itératif '*Asymmetric Least Squares*' (WLS), basé sur le filtre de Whittaker[103] (une explication de cette méthode est fournie en annexe C). Puis, chaque spectre est normalisé suivant sa norme L1. Et enfin, une zone d'intérêt allant de 834 à 2008 cm^{-1} a été sélectionnée sur les spectres Raman corrigés (Figure V.12).

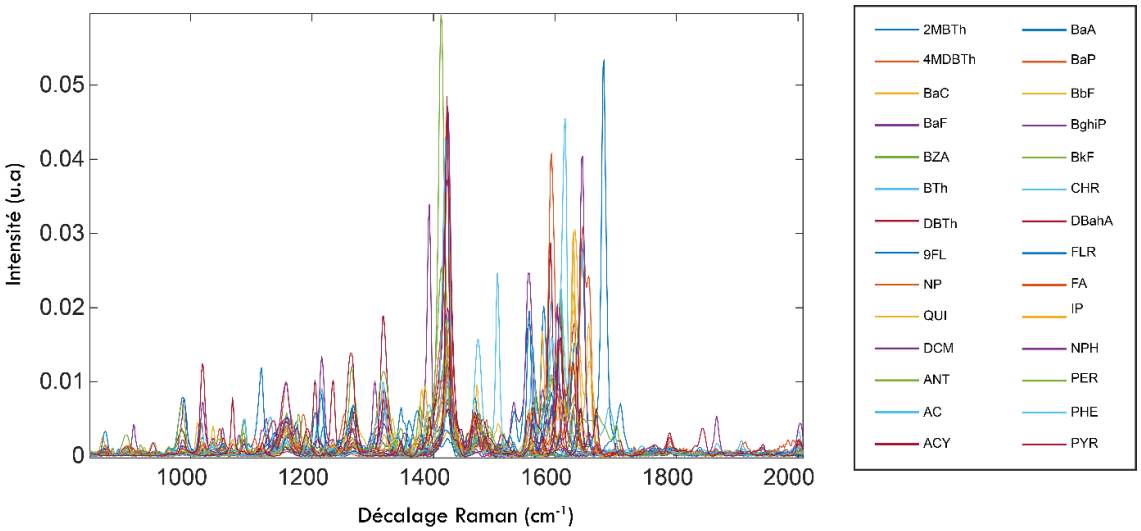


Figure V.12 : Spectres Raman corrigés des 27 CAP de l'étude.

V.5.2.2 Analyse des réponses Raman par catégories et sous-catégories des CAP

Les spectres Raman des 27 CAP sélectionnés présentent une grande variabilité. Les résultats obtenus sont discutés de manière non exhaustive en fonction de la catégorie ou de la sous-catégorie du CAP, et de leurs bandes caractéristiques ou communes :

(a) Les HAP et leurs dérivés carbonylés et nitrés

Les **17 HAP** présentent une bande commune à 1269 cm^{-1}. Le **PYR** et l'**IP** se caractérisent par une bande commune à 1069 cm^{-1}. Le **FLR** se distingue par une bande très intense à 1681 cm^{-1}, tandis que le **PHE** se distingue par une bande très intense à 1617 cm^{-1}. Le **FA**, quant à lui, présente une bande peu intense mais caractéristique à 1186 cm^{-1}. Le **BghiP** se caractérise par une bande très intense à 1393 cm^{-1}. Cette bande est également présente, mais de manière beaucoup moins intense, pour le **BkF**, le **CHR** et le **PYR**. Le **NPH** présente une bande caractéristique à 1866 cm^{-1}. Le **NPH** et l'**IP** présentent une bande commune à 1019 cm^{-1}. Le **PYR**, le **PER**, l'**IP**, le **BghiP**, le **BaP**, le **BaA** et l'**ANT** présentent une bande commune à 985 cm^{-1} (Figure V.13).

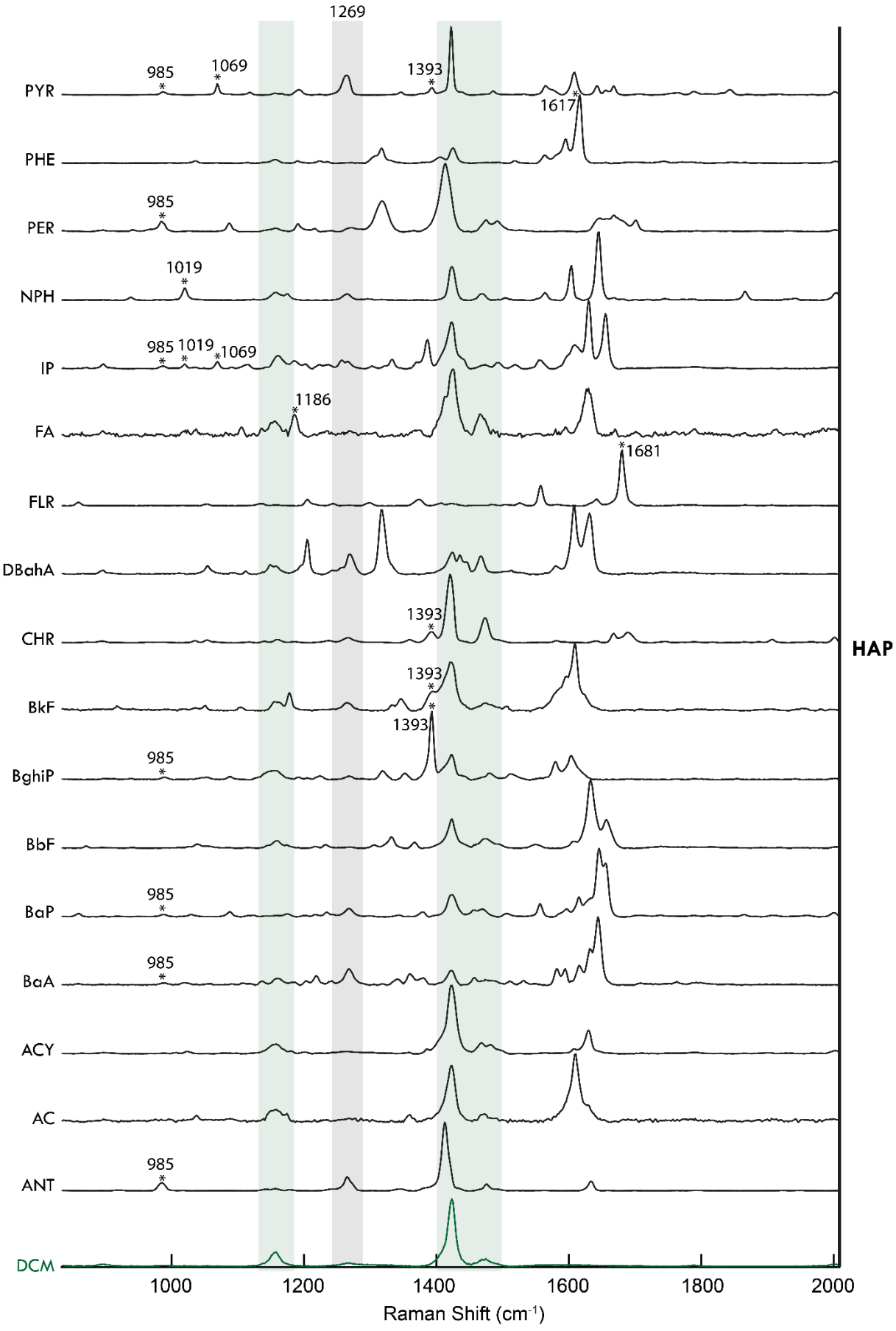


Figure V.13 : Spectres Raman corrigés des 17 HAP de l'étude. DCM : dichlorométhane.

Le **9FL**, le **BaF** et le **BZA** sont des **céto-HAP**, qui se caractérisent par des signatures Raman différentes. Le **9FL** et le **BaF** se caractérisent par une bande Raman commune à 987 cm^{-1}. Le **9FL** et le **BZA** se caractérisent par une bande commune à 1594 cm^{-1}. Cette dernière est comprise dans le cas du **BZA** dans un massif de trois bandes à 1569 cm^{-1}, 1594 cm^{-1} et 1609 cm^{-1}. Le **9FL** se caractérise par deux bandes à 1347 et à 1708 cm^{-1}. Le **BaF** se caractérise par une bande à 1216 cm^{-1}. Le **BZA** se caractérise par la bande à 1609 cm^{-1} (Figure V.14). Le **nitro-HAP**, **NP**, se caractérise, dans le cadre de cette étude, par une faible réponse Raman en raison de sa faible concentration (0.05 $mg.mL^{-1}$). Néanmoins, sa réponse est exploitable car on observe une intensité minimale de 314 u.a à la bande de 1564 cm^{-1} et une intensité maximale de 709 u.a à la bande de 1630 cm^{-1}. Cette dernière bande se caractérise par un léger épaulement vers 1644 cm^{-1} (Figure V.14).

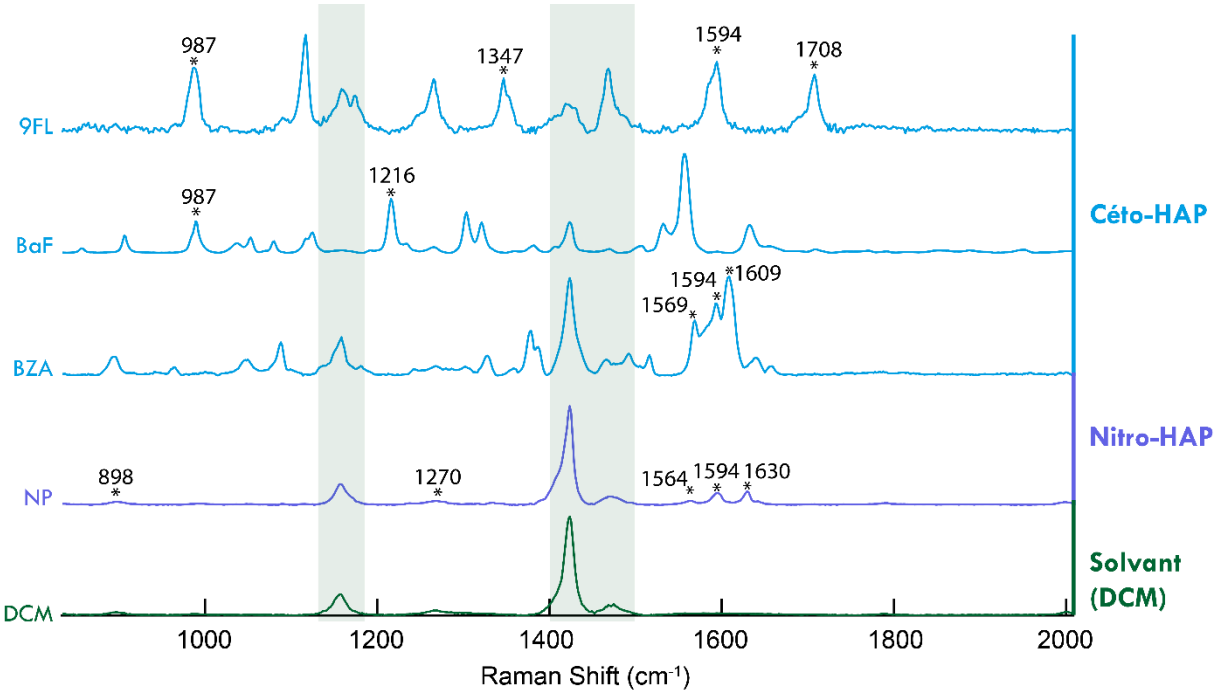


Figure V.14 : Spectres Raman corrigés des 3 céto-HAP et du nito-HAP de l'étude. DCM : dichlorométhane.

(b) Les S-CAP (Thiazaarènes) et leurs dérivés alkylés

Le **BTh**, le **2MBTh** et le **DBTh** ainsi que le **4MDBTh**, partagent les bandes Raman à 1019 cm^{-1} et 1593 cm^{-1}, à l'exception du cas du **2MBTh** où l'on observe un décalage de cette dernière vers 1582 cm^{-1}. Le **BTh** et le **2MBTh** se caractérisent par la bande Raman commune à 1216 cm^{-1}, tandis que le **DBTh** se distingue par la bande à 1235 cm^{-1}, et enfin le **4MDBTh** par la bande à 1254 cm^{-1}. Le **DBTh** et le **2MBTh** se partagent la bande à 1556 cm^{-1}, tandis que le **BTh** et le **4MDBTh** se partagent la bande à 1564 cm^{-1}. Enfin, le **BTh** se caractérise par une bande intense à 1505 cm^{-1} (Figure V.15).

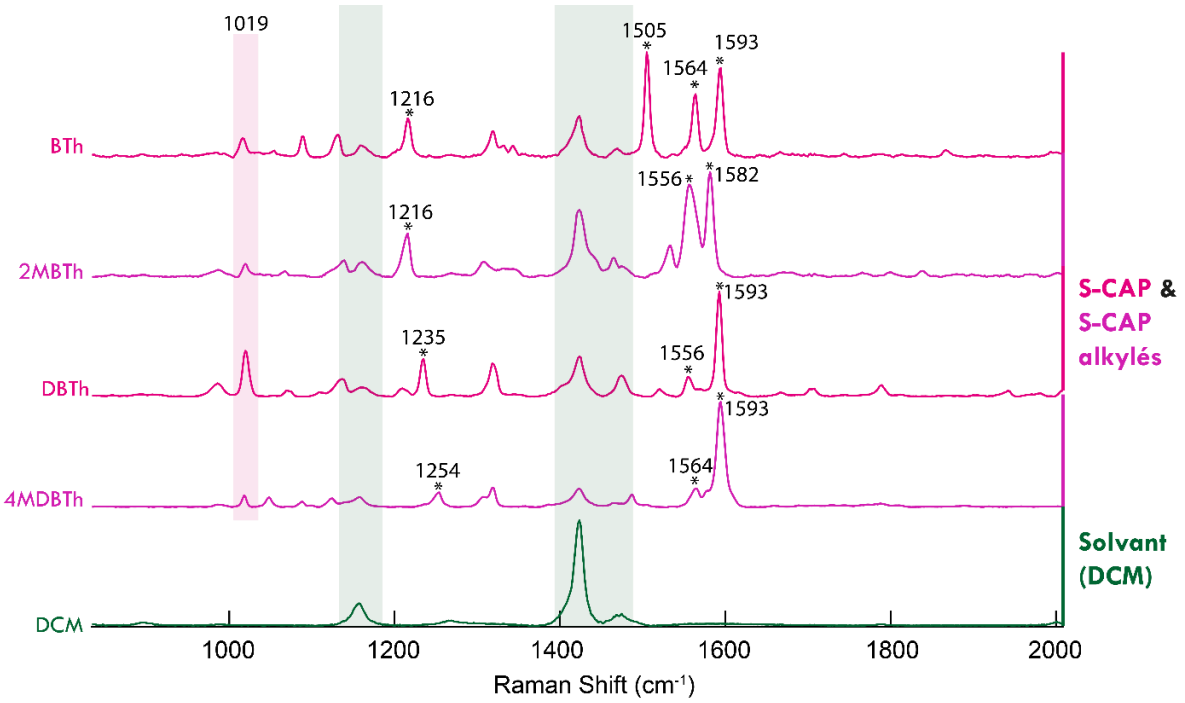


Figure V.15 : Spectres Raman corrigés des 2 S-CAP et des 2 dérivés de S-CAP de l'étude. DCM : dichlorométhane.

(c) Les N-CAP (Azaarènes)

Les deux **N-CAP** de l'étude, à savoir la **QUI** et le **BaC**, se caractérisent par une bande commune, la plus intense dans leurs spectres respectifs, qui se situe à 1630 cm^{-1}. Cependant, en dehors de cette bande, la **QUI** et le **BaC** présentent des signatures Raman totalement différentes, cohérentes avec leurs différences structurelles. Par exemple, la **QUI** montre un ensemble de bandes vibrationnelles entre 1018 cm^{-1} et 1055 cm^{-1}, tandis que le **BaC** ne présente qu'une seule bande proche de cette région, qui est située à 1011 cm^{-1}. Ce dernier se caractérise également par une bande relativement intense à 1269 cm^{-1} (Figure V.16).

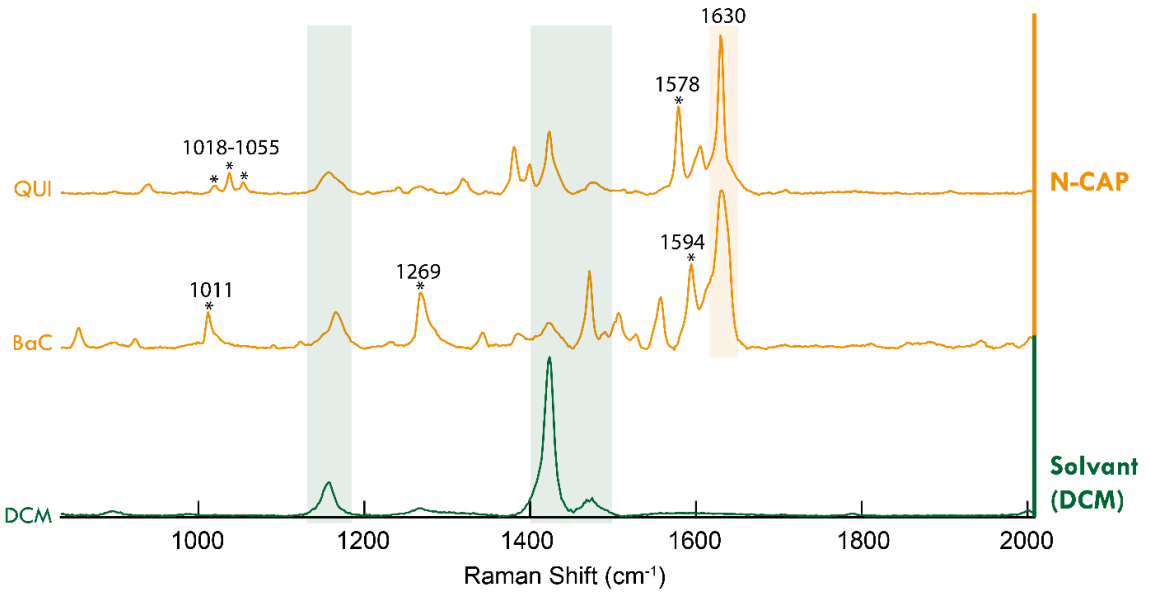


Figure V.16 : Spectres Raman corrigés des 2 N-CAP de l'étude. DCM : dichlorométhane.

Les résultats expérimentaux obtenus dans cette étude préliminaire montrent la richesse des spectres Raman et la spécificité des bandes vibrationnelles en relation avec les molécules individuelles de CAP catégorisées ou sous-catégorisées. De plus, cette première phase de l'étude met également en évidence le caractère complémentaire de la spectroscopie Raman avec la spectroscopie de fluorescence pour la caractérisation des CAP. Par exemple, l'**ACY**, le **BZA**, le **NP** et la **QUI** présentent de très mauvaises réponses en fluorescence, mais montrent de bonnes réponses en Raman. Mais quand est-il de la réponse Raman des extraits organiques de sols pollués aux CAP ?

V.5.3 Description par spectroscopie Raman des extraits de sols pollués aux CAP

L'objectif ici est d'évaluer la possibilité de discriminer les différents extraits organiques de sols en identifiant des bandes caractéristiques propres ou des tendances liées à la réponse spectrale Raman. Comme pour l'étude préliminaire, un effet multiplicatif est observé dans les spectres Raman en raison des phénomènes physiques tels l'échauffement thermique et/ou des différences d'opacité des échantillons (Figure V.17a). Il est donc important de les corriger. Une zone d'intérêt, compatible avec la zone étudiée en phase préliminaire et où la majorité des bandes vibrationnelles est observée, est sélectionnée de 200 à 1796 cm^{-1} (Figure V.17b). Ensuite, chaque spectre a été normalisé suivant sa norme L1 (Figure V.17c). Puis, l'effet additif résiduel observé (i.e. déviation quasiment constante de la ligne de base sur tous les décalages Raman) a été corrigé grâce à l'algorithme WLS, basé sur le filtre de Whittaker avec comme paramètre $p = 0.001$ et $\lambda = 1000$ (Figure V.17d).

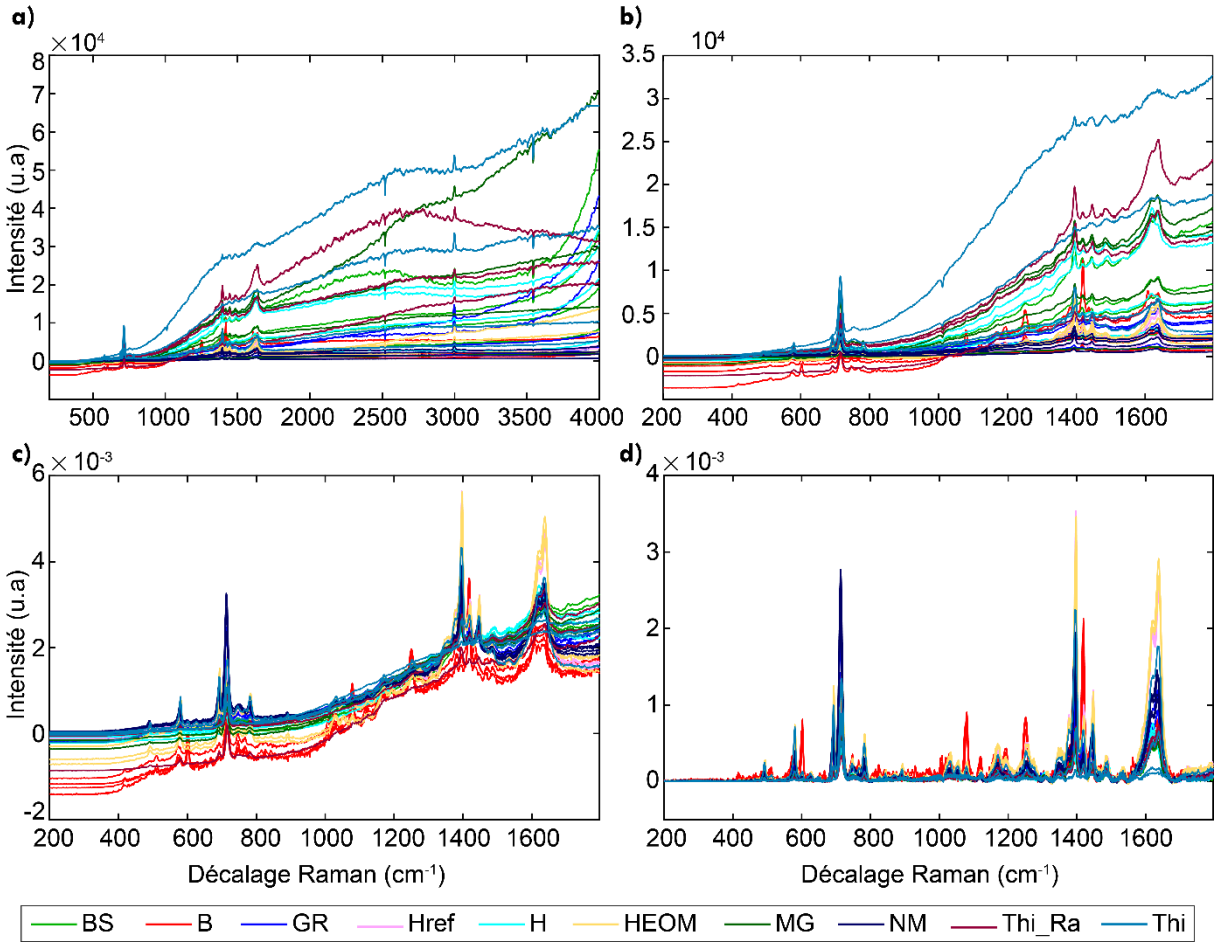


Figure V.17 : Spectres Raman bruts des 37 extraits de sols de l'étude (a). Spectres Raman bruts découpés à la zone d'intérêt allant de 200 à 1796 cm^{-1} (b). Spectres Raman découpés et normalisés suivant leur norme L1 (c). Spectres Raman découpés, normalisés et corrigés par WLS (d).

V.5.3.1 Analyse en composantes principales des spectres Raman

Une fois les spectres corrigés, une analyse ACP a été effectuée. Le modèle a été construit avec 8 composantes principales, ce qui permet d'expliquer 99,02% de la variance totale des données Raman. Les composantes principales, CP1, CP2, CP3, CP4, CP5, CP6, CP7 et CP8 expliquent 76.40%, 10.43%, 9.02%,1.49%, 0.58%, 0.51%, 0.38% et 0.21% de la variance, respectivement.

L'analyse des *'scores'* dans l'espace CP1-CP2 révèle une bonne discrimination des sols **NM**, **GR**, **B**, **Href** et **HEOM**. On constate également une faible variation inter-classe au sein de tous les groupes ayant des réplicas (i.e. tous les groupes sauf le groupe **Thi**). Cela témoigne d'une bonne reproductibilité de l'extraction des sols au départ et des mesures Raman associées. Comme dans le cas de la modélisation ACPM des MEEF du chapitre IV, le groupe **Thi** présente une variabilité relativement importante entre les différentes classes (i.e. échantillons) qui le composent. En effet, la classe **Thi 170** est très proche des groupes **H**, **MG**, **Thi_Ra** et **BS**, tandis que les classes **Thi 163** et **Thi 166** en sont légèrement éloignées mais restent relativement proches entre elles. De plus, on observe que la classe **Thi 87** est proche

des groupes **Href** et **HEOM** (Figure V.18a). Les '*loadings*' de la CP1 et de la CP2 indiquent les bandes vibrationnelles impliquées dans la discrimination des groupes observée sur la carte des '*scores*'. La CP1 présente principalement des bandes positives, les plus majoritaires sont aux positions 1396, 1618 et 1638 cm^{-1}. Cela indique que les groupes **Href** et **HEOM**, qui ont des '*scores*' CP1 positifs élevés, sont très bien représentés par ces bandes. Les autres groupes ayant des '*scores*' CP1 nuls ou négatifs sont peu représentés par ces bandes. La CP2 présente deux bandes positives majoritaires aux positions de 712 cm^{-1} et 1393 cm^{-1}, ainsi qu'une bande majoritaire négative à 1419 cm^{-1} (Figure V.18b). Le groupe **NM** est le groupe ayant les '*scores*' positifs les plus élevés en CP2, ce qui indique qu'il est bien représenté par les bandes positives (i.e. 712 cm^{-1} et 1393 cm^{-1}). En revanche, le groupe **B** est le groupe ayant les '*scores*' CP2 les plus négatifs, ce qui indique qu'il est bien représenté par la bande négative à 1419 cm-1 (Figure V.18).

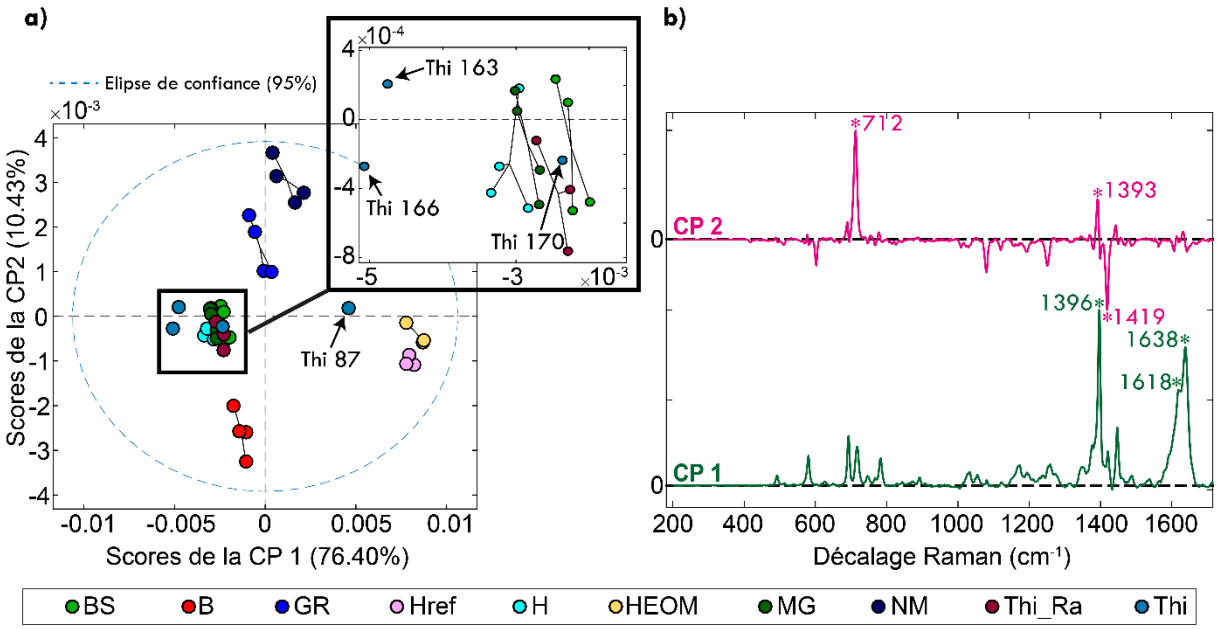


Figure V.18 : 'Scores' de la 1ère composante principale versus 'scores' de la 2ème composantes principale de la modélisation ACP des spectres Raman des 37 extraits de sols pollués aux CAP (a). 'Loadings' de la 1ère composante principale et de la 2ème composante principale (b). CP signifie composante principale.

Ensuite, en examinant les '*scores*' obtenus dans le plan CP1-CP5, il est possible d'optimiser la distinction entre les groupes **BS**, **MG**, **Thi_Ra** et **H**, qui n'étaient pas bien discriminés par les 4 premières CP. En effet, le groupe **BS** présente des '*scores*' CP5 positifs, tandis que les groupes **H** et **Thi_Ra** présentent des '*scores*' négatifs. Le groupe **MG** quant à lui présente des '*scores*' presque nuls (Figure V.19a). La CP5 présente deux bandes positives majoritaires aux positions de 688 cm^{-1} et 1392 cm^{-1}, ainsi que trois bandes négatives majoritaires aux positions de 715 cm^{-1}, 1620 cm^{-1} et 1642 cm^{-1} (Figure V.19b). Le groupe **BS** est donc bien représenté par les deux bandes positives, tandis que les groupes **H** et **Thi_Ra** sont plutôt bien représentés par les trois bandes négatives. Le groupe **MG**, quant à lui, est très

peu représenté par l'ensemble des bandes de cette CP car ses '*scores*' sont presque nuls (Figure V.19).

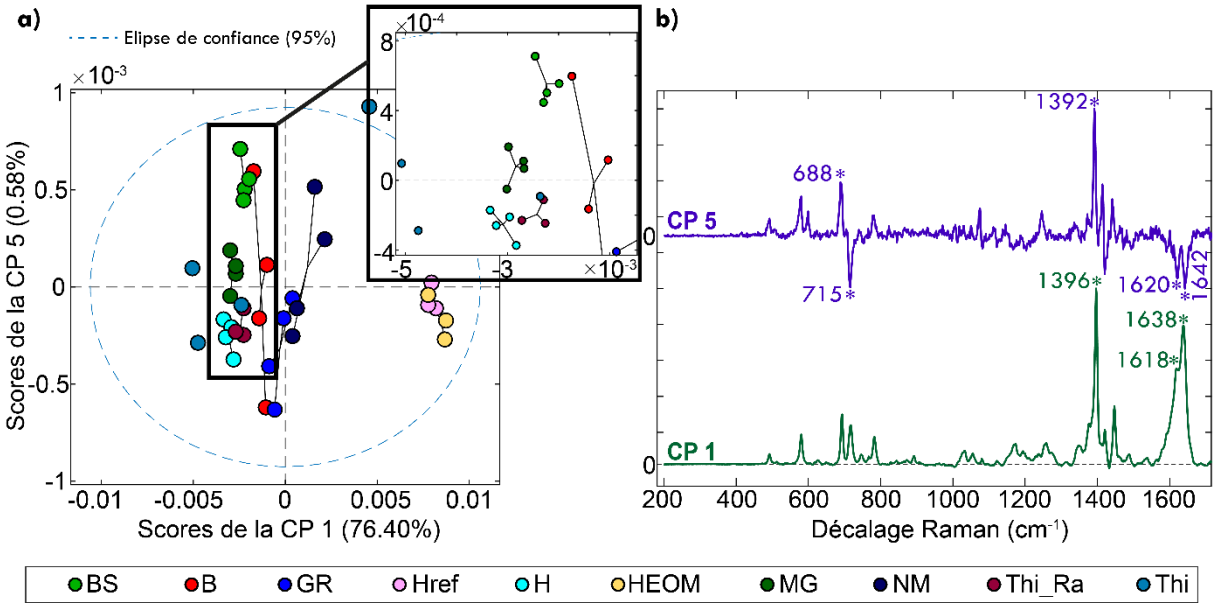


Figure V.19 : 'Scores' de la 1ère composante principale versus 'scores' de la 5ème composantes principale de la modélisation ACP des spectres Raman des 37 extraits de sols pollués aux CAP (a). 'Loadings' de la 1ère composante principale et de la 5ème composante principale (b). CP signifie composante principale.

De plus, la projection de l'ensemble des spectres dans les espaces CP1-CP6 et CP1-CP7 permet d'améliorer la discrimination entre les groupes **Href** et **HEOM** qui étaient jusque-là, très proches en termes de '*scores'*. En effet, que ce soit pour la CP6 ou la CP7, les '*scores*' de **Href** sont positifs tandis que ceux de **HEOM** sont négatifs (Figure V.20a et 20b). Ainsi, le groupe **Href** est bien représenté par les bandes positives de la CP6, qui sont situées pour les plus majoritaires aux positions de 1393 cm^{-1}, 1420 cm^{-1} et 1443 cm^{-1}, ainsi que par les bandes positives de la CP7, qui sont situées pour les plus majoritaires aux positions de 709 cm^{-1} et 1398 cm^{-1} (Figure V.20d).

Et enfin, l'analyse des '*scores*' dans l'espace CP1-CP8 permet d'améliorer la discrimination du groupe **H** par rapport aux autres groupes. En effet, ses '*scores*' pour cette CP sont négatifs (Figure V.20c). Ainsi, les bandes négatives de la CP8 situées pour les plus majoritaires aux positions de 580 cm^{-1}, 693 cm^{-1} et 1620 cm^{-1} sont caractéristiques de ce sol (Figure V.20d).

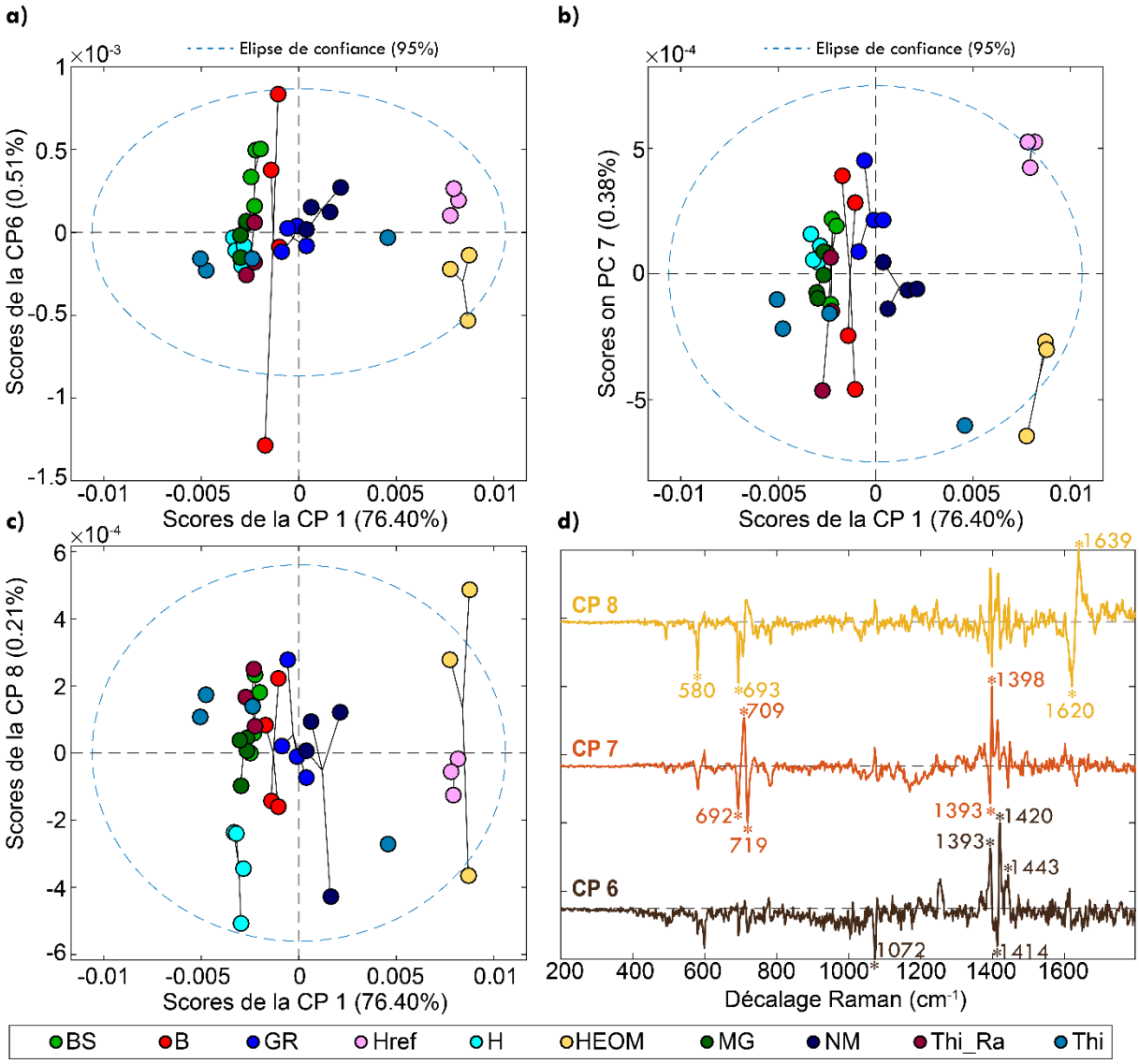


Figure V.20 : 'Scores' de la 1ère composante principale versus 'scores' de la 6ème composantes principale de la modélisation ACP des spectres Raman des 37 extraits de sols pollués aux CAP (a). 'Scores' de la 1ère composante principale versus 'scores' de la 7ème composantes principale de la modélisation ACP des spectres Raman des 37 extraits de sols pollués aux CAP (b). 'Scores' de la 1ère composante principale versus 'scores' de la 8ème composantes principale de la modélisation ACP des spectres Raman des 37 extraits de sols pollués aux CAP (c). 'Loadings' de la 6ème composante principale, de la 7ème composante principale et de la 8ème composante principale (d). CP signifie composante principale.

L'analyse ACP précédente permet de discriminer la plupart des extraits organiques de sols de l'étude et aussi d'établir, en lien avec les '*loadings*', les bandes vibrationnelles responsables de cette discrimination. En perspective à ce travail, une étude de la nature des bandes vibrationnelles (i.e. type de vibration moléculaire) pourra être menée pour approfondir la caractérisation qualitative des CAP présents dans ces sols industriels contaminés. Dans le but d'affiner davantage la capacité de discriminer les différents extraits organiques des sols contaminés grâce à la spectroscopie Raman, et dans le cadre du déploiement d'une approche multi-blocs permettant une analyse simultanée des données Raman et des images de fluorescence 3D, un autre prétraitement de chimiométrie a été mis en place. Son objectif est d'améliorer cette discrimination en mettant en évidence des caractéristiques dans le domaine

des fréquences, ainsi que de transformer un spectre Raman en une matrice (i.e. une image) caractéristique de l'échantillon étudié et riche en informations spectrales et fréquentielles. Il s'agit de la transformée en ondelettes continue qui permet, contrairement à la transformée de Fourier (fréquemment utilisé), une transition vers le domaine fréquentiel tout en préservant les informations spectrales. Étant donné l'existence également de la transformée en ondelettes discrète, le reste du texte se concentrera uniquement sur la transformée en ondelettes continue. Par ailleurs, une explication de la transformée en ondelettes continue, de la transformée de Fourrier et de la différence entre les deux méthodes est fournie en annexe D.

V.5.3.2 Analyse en composantes principales multivoie des transformées en ondelettes continues des spectres Raman

D'un point de vue mathématique, la transformée en ondelettes continue repose sur une fonction de base, appelée ondelette mère et notée $\psi(t)$, à partir de laquelle on génère une famille d'ondelettes filles $\psi_{a,b}(t)$ dilatées et translatées en ajustant les paramètres a (paramètre de dilatation) et b (paramètre de translation) pour couvrir différentes échelles et positions dans le signal d'entrée $f(t)$[117] :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \qquad (V.2)$$

$$si\ a = 1\ et\ b = 0, alors\ \psi_{a,b}(t) = \psi(t)$$

Où t est la variable en fonction de laquelle le signal évolue (ex. temps, longueur d'onde, décalage Raman, etc.). Le terme $\frac{1}{\sqrt{a}}$ garantit à toutes les ondelettes filles une norme identique à celle de l'ondelette mère (i.e. même aire et même énergie). Ces fonctions filles $\psi_{a,b}(t)$ sont convoluées par intégration avec le signal $f(t)$ pour obtenir les coefficients $W(a,b)$ qui constitue alors une mesure de la similarité entre le signal et une ondelette fille avec les paramètres a et b donnés[130] :

$$W(a,b) = \int_{-\infty}^{+\infty} \psi_{a,b}(t)^* f(t)dt \qquad (V.3)$$

Où $\psi_{a,b}(t)^*$ est le conjugué complexe de $\psi_{a,b}(t)$.

Il existe plusieurs ondelettes mères[131] plus ou moins adaptées aux signaux spectroscopiques. L'ondelette de Morlet[132] a été choisie dans le cadre de cette étude en raison de sa résolution précise en temps et en fréquence, ainsi que de sa sensibilité aux transitions à hautes fréquences, telles que les bandes vibrationnelles fines :

$$\psi(t) = e^{-t^2/2}\cos(5t) \qquad (V.4)$$

Où t est la variable en fonction de laquelle le signal évolue (ex. temps, longueur d'onde, décalage Raman, etc.). L'ondelette mère de Morlet $\psi_{1,0}(t)$, ainsi que l'ondelette de Morlet dilatée $\psi_{2,0}(t)$ et une autre translatée $\psi_{1,8}(t)$, sont représentées dans la Figure V.21.

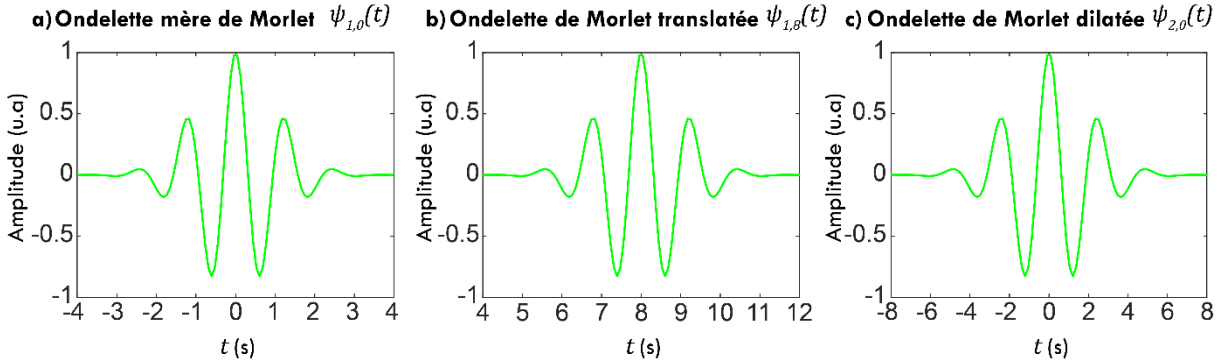


Figure V.21 : Ondelettes mère de Morlet $\psi_{1,0}(t)$ (a). Ondelette de Morlet translatée $\psi_{1,8}(t)$ (b) et ondelette de Morlet dilatée $\psi_{2,0}(t)$ (c).

Une fois le calcul de la transformée en ondelettes de chaque spectre Raman est réalisé, un scalogramme est obtenu. Il s'agit d'une image qui porte sur l'axe de ses abscisses l'échelle des décalages Raman en cm^{-1} et sur l'axe de ses ordonnées l'échelle des fréquences en Hz. L'intensité de chaque pixel est relative à la magnitude en u.a, qui est reliée aux intensités Raman de départ. Les scalogrammes des échantillons BS500R1 et MG500R1 sont représentés dans la Figure V.22 à titre d'exemple. Tous les autres scalogrammes sont disponibles en annexe E. Ces scalogrammes, qui sont donc des matrices de données, ont ensuite été analysés par ACPM pour essayer de discriminer les différents groupes de l'étude et d'établir, en lien avec les '*loadings*', les bandes vibrationnelles responsables de cette discrimination.

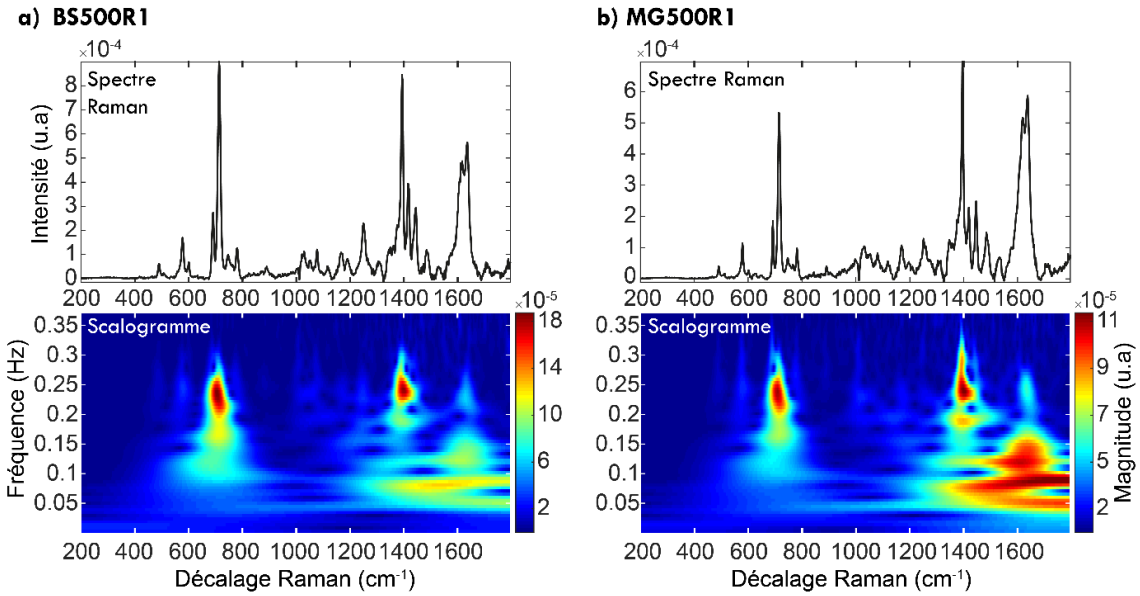


Figure V.22 : Spectre Raman en haut et son scalogramme associé en bas de l'échantillon BS500R1 (a). Spectre Raman en haut et son scalogramme associé en bas de l'échantillon MG500R1 (b).

Le modèle ACPM a été construit avec 9 composantes principales, ce qui permet d'expliquer 99.60% de la variance totale dans les données. La première CP explique 80.75% de la variance, la deuxième CP explique 11.39%, et la troisième CP explique 5.73%. Les CP4, CP5, CP6, CP7, CP8 et CP9 expliquent respectivement 0.88%, 0.27%, 0.22%, 0.17%, 0.13% et 0.07% de la variance dans les données.

En analysant la projection de l'ensemble des scalogrammes dans le plan CP1-CP2, on peut distinguer les groupes **Href** et **HEOM** des autres groupes le long de la CP1, ainsi que les groupes **NM** et **GR** des autres groupes le long de la CP2. En utilisant l'espace CP1-CP3, on peut également discriminer le groupe **B** le long de l'axe CP3. Le groupe **Thi** présente une variabilité relativement importante entre les différents échantillons qui le composent. Plus précisément, la classe **Thi 170** est fortement liée aux groupes **H**, **MG**, **Thi_Ra** et **BS**, tandis que les classes **Thi 163** et **Thi 166** sont légèrement moins proches mais restent relativement proches entre elles selon la CP1, et sont séparées selon la CP2. La classe **Thi 87** se rapproche le plus des groupes **Href** et **HEOM** selon la CP1. Cependant, cette espèce est mal représentée par la CP2, car ses '*scores*' sont presque nuls pour cette CP (Figure V.23a). L'analyse des '*loadings*' des CP1, CP2 et CP3 révèle les bandes vibrationnelles responsables des discriminations mentionnées précédemment. Les classes **Href** et **HEOM**, qui présentent les '*scores*' positifs les plus élevés en CP1, sont donc bien représentées par les bandes à 703 cm^{-1}, 1396 cm^{-1} et 1639 cm^{-1}. De plus, ces mêmes classes, qui ont des '*scores*' négatifs en CP2 et CP3, sont également bien représentées par les bandes négatives des CP2 et CP3, à savoir les bandes à 1056 cm^{-1}, 1419 cm^{-1}, 1527 cm^{-1} et 1634 cm^{-1}. Les classes **NM** et **GR**, qui ont les '*scores*' positifs les plus élevés en CP2, sont bien représentées par la bande de la CP2 à 712 cm^{-1}. En revanche, la classe **B** est très mal représentée par les bandes de la CP1 et de la CP2, puisque ses '*scores*' pour ces deux CP sont quasiment nuls. Cependant, ses '*scores*' pour la CP3 sont positifs et très élevés, ce qui indique que cette classe est plutôt bien représentée par les bandes à 602 cm^{-1}, 717 cm^{-1}, 1087 cm^{-1}, 1249 cm^{-1}, 1424 cm^{-1} et 1449 cm^{-1} (Figure V.23b).

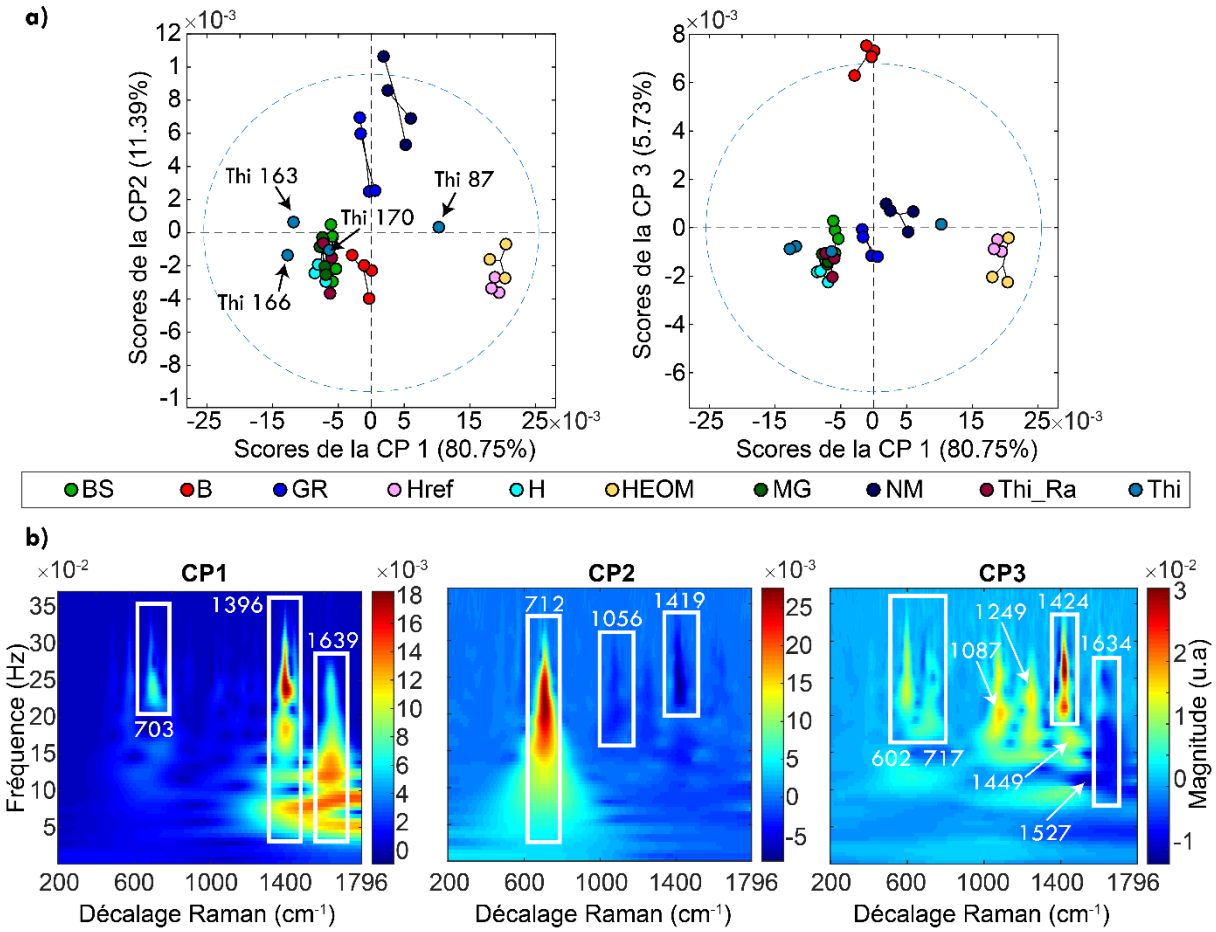


Figure V.23 : 'Scores' de la 1ère composante principale (CP1) versus la 2ème composante principale (CP2) dans le graphique de gauche et versus la 3ème composante principale dans le graphique de droite (a). De gauche vers la droite, 'loadings' de la 1ère, de la 2ème et de la 3ème composante principale (b). CP : composante principale.

Ensuite, les *'scores'* sur les axes CP1-CP4 montrent une discrimination entre les groupes **BS** et **H**, tant entre eux qu'en comparaison avec les groupes **Thi_Ra** et **MG** en sachant que ces groupes étaient mal discriminés suivant les 3 premières CP. Le groupe **BS** se caractérise par des *'scores'* positifs élevés en CP4, tandis que le groupe **H** présente des *'scores'* négatifs sur la même CP. Les groupes **MG** et **Href**, quant à eux, présentent des '*scores*' quasiment nuls pour la CP4 (Figure V.24a). Les bandes vibrationnelles positives du '*loading*' de la CP4, situées à 490 cm^{-1}, 575 cm^{-1}, 693 cm^{-1} et 1396 cm^{-1}, sont donc caractéristiques du groupe **BS**, tandis que les bandes vibrationnelles négatives, situées à 704 cm^{-1}, 1433 cm^{-1} et 1467 cm^{-1}, sont caractéristiques du groupe **H**. En revanche, l'ensemble de ces bandes de la CP4, qu'elles soient négatives ou positives, sont peu représentatives des groupes **MG** et **Thi_Ra** (Figure V.24b).

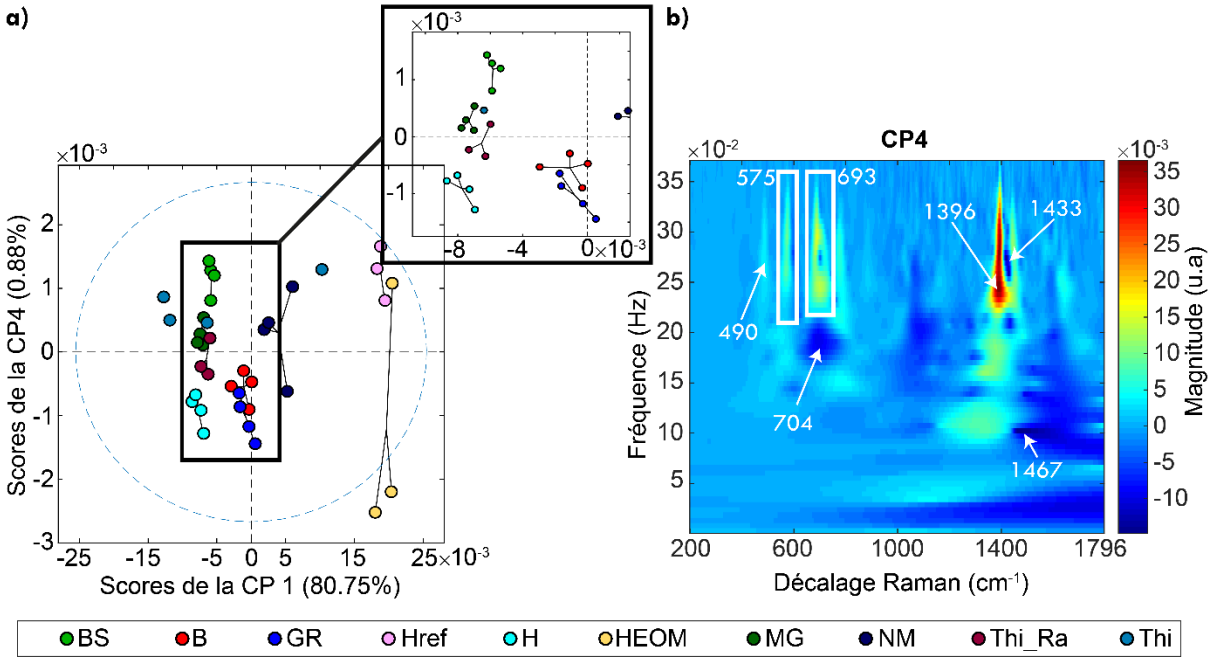


Figure V.24 : 'Scores' de la 1ère composante principale versus la 4ème composante principale (a). 'Loadings' de la 4ème composante principale (b). CP : composante principale.

En outre, les CP5 et CP6 nous permettent, de manière complémentaire, de discriminer entre les groupes **Href** et **HEOM** qui présentaient jusqu'alors une grande proximité en termes de '*scores*'. En effet, grâce à la projection des spectres dans les espaces CP1-CP5 et CP1-CP6 on observe des '*scores*' CP5 positifs pour le groupe **HEOM** et négatifs pour le groupe **Href** et à contrario des '*scores*' CP6 positifs pour le groupe **Href** et négatifs pour le groupe **HEOM** (Figure V.25a). Par conséquent, les bandes vibrationnelles principales caractéristiques du groupe **HEOM** sont situées à 581 cm^{-1} et 696 cm^{-1}. Ces bandes présentent des '*scores*' positifs en PC5 et négatifs en PC6. Les bandes vibrationnelles principales caractéristiques du groupe **Href** sont situées à 693 cm^{-1} à 711 cm^{-1}, 1408 cm^{-1}, 1419 cm^{-1} et 1617 cm^{-1}. La bande à 1396 cm^{-1} est un cas intéressant qui met en évidence toute la puissance de la transformée en ondelettes. En effet, cette bande présente des '*scores*' positifs dans les deux PC, mais avec une fréquence plus basse en PC5 par rapport à PC6. Cela indique que les deux groupes possèdent la bande à 1396 cm^{-1}, mais que dans le cas du groupe **HEOM**, cette bande est plus large (i.e. bande à plus basse fréquence) que dans le cas du groupe **Href** (i.e. bande à plus haute fréquence) (Figure V.25b).

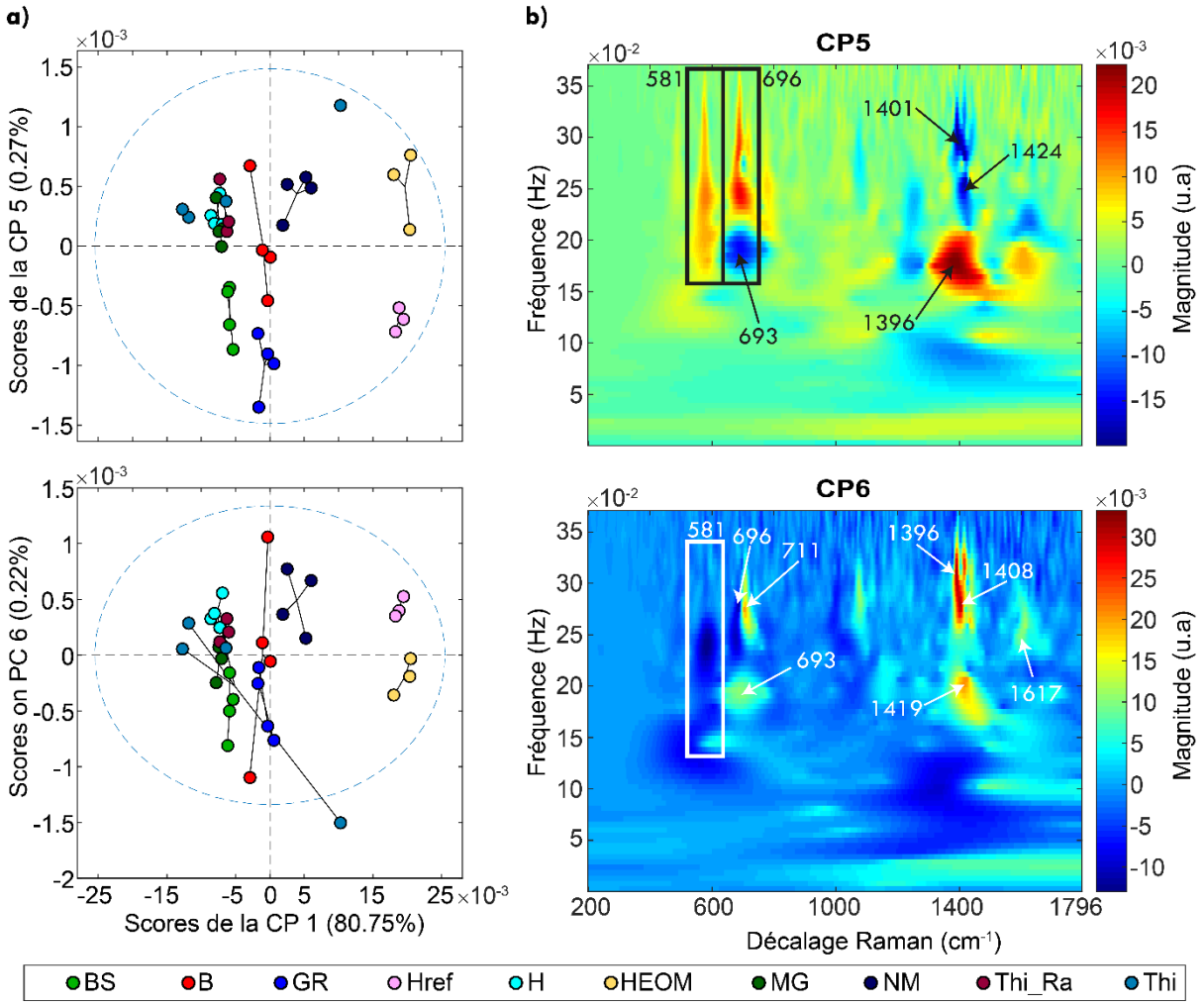


Figure V.25 : 'Scores' de la 1ère composante principale versus la 5ème composante principale dans le graphique de haut et versus la 6ème composante principale dans le graphique du bas (a). 'Loadings' de la 5ème composante principale en haut et de la 6ème composante principale en bas (b). CP : composante principale.

Et enfin, les groupes **H**, **BS**, **MG** et **Thi_Ra** peuvent être discriminés le long de la CP9 grâce à la projection de l'ensemble des scalogrammes dans l'espace CP1-CP9. En effet, les groupes **H** et **BS** présentent des '*scores*' CP9 positifs, avec des '*scores*' plus élevés pour le groupe **H**. En revanche, les groupes **MG** et **Thi_Ra** présentent des '*scores*' CP9 négatifs, avec les '*scores*' les plus intenses dans l'échelle négative pour le groupe **Thi_Ra** (Figure V.26a). Les bandes représentatives des groupes **H** et **BS** sont les bandes positives du '*loading*' de la CP9. Il s'agit des bandes à 575 cm^{-1}, 682 cm^{-1}, 1555 cm^{-1} et à 1610 cm^{-1}. Les bandes représentatives des groupes **MG** et **Thi_Ra** sont les bandes négatives du '*loading*' de la CP9. Il s'agit des bandes à 723 cm^{-1}, 1430 cm^{-1} et 1650 cm^{-1} (Figure V.26b). De plus, l'intensité des '*scores*' est liée à l'intensité des bandes vibrationnelles, ce qui indique que la bande à 575 cm^{-1} est plus intense pour le groupe **H** que pour le groupe **BS**.

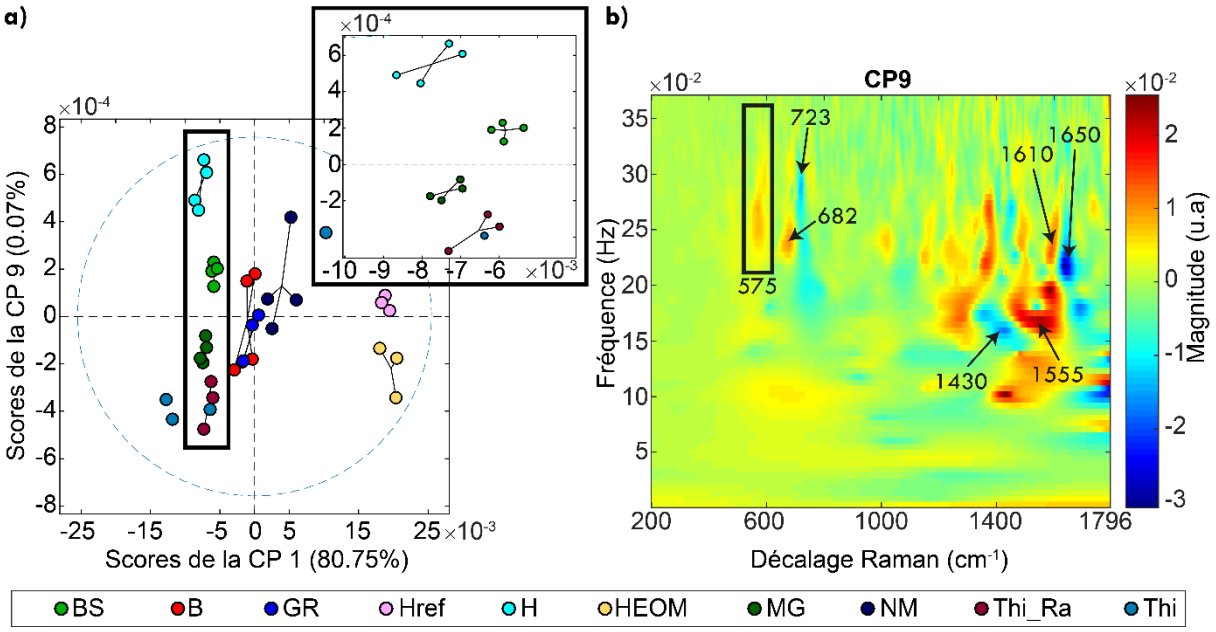


Figure V.26 : 'Scores' de la 1ère composante principale versus la 9ème composante principale (a). 'Loadings' de la 9ème composante principale (b). CP : composante principale.

Pour conclure, la transformée en ondelettes nous a permis d'effectuer une analyse plus fine des échantillons et d'obtenir une meilleure discrimination entre les groupes. Les informations obtenues à partir de l'analyse ACP classique et de l'analyse ACPM après la transformée en ondelettes sont complémentaires, car les bandes vibrationnelles responsables des séparations sont les mêmes, mais avec une plus grande finesse dans le cas de la transformée en ondelettes, ce qui permet une meilleure séparation. En effet, le domaine fréquentiel permet de distinguer les signaux qui se superposent dans le cas de l'analyse ACP classique. Un exemple parfait en est la bande à 1396 cm^{-1}, qui est plus large dans le groupe **HEOM** que dans le groupe **Href**, ce qui a pu être observé grâce à la distinction entre hautes et basses fréquences.

V.6 Conclusion

Dans le cadre de ce chapitre V, nous avons cherché à caractériser quantitativement les 16 HAP cibles présents dans les 37 extraits organiques des sols industriels. Pour ce faire, nous avons d'abord procédé à la quantification des 16 HAP cibles par GC-MS dans les extraits organiques de sols contaminés. Nous avons, ensuite, développé des modèles prédictifs de régression linéaire NPLS et non linéaire SVR afin d'établir les liens entre les informations spectrales et les informations quantitatives des HAP. Les résultats insatisfaisants obtenus lors de la modélisation linéaire NPLS suggèrent une relation non linéaire entre les données spectrales et les données quantitatives, malgré les précautions prises en amont pour respecter la condition de la loi de Beer-Lambert. Ceci met en évidence l'impact des phénomènes physico-chimiques sur la réponse en fluorescence et la difficulté à maintenir la linéarité des

signaux de fluorescence, même après les différentes corrections chimiométriques effectuées. En revanche, les résultats obtenus lors de la modélisation non linéaire SVR étaient satisfaisants en termes de performances de prédiction et de généralisation. Après optimisation et validation des modèles, nous avons pu tester avec succès la qualité de prédiction de ces modèles sur des extraits organiques qui n'avaient pas été inclus dans l'ensemble d'apprentissage des modèles (GR100R2 et HEOMR1).

Une fois cette étape réussie, nous avons étudié la réponse en fluorescence d'autres CAP que les HAP cibles, en nous référant aux connaissances disponibles dans la littérature. Il est apparu clairement, suite à l'évaluation des réponses en fluorescence des NSO-CAP, de leurs dérivés ainsi que des dérivés des HAP, que la caractérisation complète de la contamination des sols par les CAP est difficile à atteindre en se basant uniquement sur la spectroscopie de fluorescence. En revanche, l'analyse des réponses Raman de différentes catégories de CAP purs a démontré la pertinence de cette technique pour la détection de ces composés et sa complémentarité donc avec la spectroscopie de fluorescence.

De plus, après avoir analysé les extraits organiques des sols par spectroscopie Raman, nous avons réussi à discriminer la plupart des sols de notre étude. Nous avons également pu établir, en lien avec les ‘l*oadings*’, les bandes vibrationnelles responsables de cette discrimination. Dans un souci de perspective à ce travail de thèse, nous avons tester la transformée en ondelettes pour tenter d’améliorer la discrimination des données et tenter de comprendre les caractéristiques fréquentielles inhérentes à nos données Raman. Il faudra toutefois poursuivre l’étude afin d'assigner les bandes vibrationnelles responsables de cette discrimination entre les différents groupes aux vibrations moléculaires correspondantes, et de les lier à la nature de la pollution subie par chaque sol.

En outre, un travail de fusion de données et d’analyse multiblocs pourra être envisagée. Il s’agira alors de prendre en compte le mieux possible, via les différentes modalités spectrales (i.e. fluorescence et Raman) et/ou non spectrales (i.e. GC-MS), toute la complexité des systèmes physico-chimiques (i.e. extraits organiques des sols contaminés) et ce dans le but de mieux les décrire, les résoudre ou les modéliser. Pour ce faire, l’algorithme MT-SVD, pourra être étendu à l’analyse des transformées en ondelettes des spectres Raman, en vue de, par exemple, les débruiter en isolant les signaux de très hautes fréquences caractéristiques du bruit instrumental. En second lieu, il est possible de tester des algorithmes d’analyse multiblocs existants tels que l'Analyse en Composantes Communes et Poids Spécifiques ou encore l’analyse de co-inertie multiple[133]. Il est également envisageable de développer de nouvelles approches et algorithmes en se basant, par exemple, sur l’analyse de données topologique[134] ou encore, de considérer les matrices de données (i.e. MEEF et scalogrammes) comme des

images dotées de propriétés (e.g. contraste), liées à la chimie des échantillons mesurés. Cela ouvre la voie à des approches novatrices dans le domaine de la caractérisation des CAP via l'analyse multiblocs des données.

CONCLUSION GENERALE ET PERSPECTIVES

L'objectif principal de ce travail de recherche était de caractériser qualitativement et quantitativement, de manière aussi étendue que possible, la contamination des sols par les composés aromatiques polycycliques (CAP) et plus particulièrement par les hydrocarbures aromatiques polycycliques (HAP). Pour ce faire, différentes techniques analytiques ont été déployées, à savoir la spectroscopie de fluorescence 3D, la spectroscopie UV-Visible et la spectroscopie Raman, ainsi que la chromatographie gazeuse couplée à la spectrométrie de masse. Parallèlement, différentes méthodologies de chimiométrie ont été développées et appliquées aux données afin de corriger, d'explorer, d'analyser et enfin de modéliser les signaux obtenus.

Dans le but de pouvoir exploiter de manière plus complète les signaux chimiques d'intérêt mesurés par fluorescence 3D, nous avons développé un nouvel algorithme de correction des matrices d'excitation-émission de fluorescence (MEEF), que nous avons nommé MT-SVD pour '*Multi-Truncated Singular Value Decomposition*'. Son approche relativement innovante permet de gérer les effets de diffusion de la lumière, le bruit et les déficiences du rang matriciel, pouvant être observées dans les MEEF. Contrairement aux autres approches proposées dans la littérature, il se distingue par une stratégie avancée de troncature par SVD couplé à une analyse d'image pour extraire les informations chimiques pertinentes et en déduire une estimation rigoureuse du rang matriciel global à partir de l'étude de rangs locaux. Cet algorithme a été testé sur des cas modèles expérimentaux avant d'être appliqué à la correction de MEEF d'échantillons complexes, à savoir le goudron de houille et aussi aux extraits organiques issus de sols industriels pollués par des CAP.

Une fois ce travail méthodologique de correction des MEEF accompli, une approche originale de chimiométrie a été adoptée pour identifier les 16 HAP ciblés présents dans des extraits de sols industriels contaminés. Cette approche combine des connaissances a priori sur la contamination des sols par les HAP, le prétraitement MT-SVD et la décomposition '*PARallel FACtor*' (PARAFAC) des MEEF. Habituellement, les MEEF des échantillons, que nous souhaitons analysées, sont décomposées par PARAFAC. Les '*loadings*' et '*scores*' associées à cette décomposition sont généralement étudiés pour identifier les espèces pures présentes et déterminer leurs quantités relatives (suivant la spectroscopie employée). Notre approche diffère car le modèle PARAFAC est construit en décomposant le cube des MEEF des espèces de référence (i.e. les 16 HAP cibles) que nous souhaitons détecter, et non les MEEF des extraits organiques des sols, qui sont les échantillons que nous souhaitons analyser. Ensuite, ces échantillons que nous souhaitons caractériser sont projetés dans le modèle et les '*scores*' résultants sont étudiés. Cette approche présente l'avantage de fournir un modèle valide et généralisable à tout type de matrice contaminée par les HAP cibles. Cependant, elle limite la recherche des espèces pures uniquement à celles qui ont été utilisées

pour la construction du modèle. Les résultats de cette étude nous ont permis de mieux comprendre la composition et la distribution des 16 HAP cibles dans les sols contaminés, en relation avec leur abondance, leur origine, mais aussi en lien avec un aspect temporel de l'activité du site industriel.

Ensuite, nous avons quantifié la pollution des sols par ces 16 HAP grâce à la chromatographie en phase gazeuse couplée à la spectrométrie de masse. Ces informations ont été couplées aux informations spectrales de fluorescence pour modéliser le lien entre elles. Dans un premier temps, nous avons utilisé un algorithme de régression linéaire appelé '*Multiway Partial Least Squares Regression'* (NPLS) car les données de fluorescence, en respectant la condition de la loi de Beer-Lambert, présument une relation linéaire entre les intensités de fluorescence et les concentrations réelles des analytes (i.e. les 16 HAP). Cependant, nous avons constaté de faibles performances des modèles linéaires, suggérant une relation de nature non linéaire entre les intensités de fluorescence mesurées et les concentrations réelles des analytes. Par conséquent, nous avons opté pour la régression '*Support Vector Regression*' (SVR) non linéaire, une méthode d'apprentissage statistique, en tant que deuxième approche. Les résultats obtenus en termes de performances de modélisation et de généralisation des modèles SVR sont satisfaisants.

Ainsi, dans cette thèse, malgré la complexité de la pollution des sols étudiés et l'influence de divers phénomènes physico-chimiques (e.g. extinction de fluorescence) sur la réponse en fluorescence, l'utilisation d'une approche de décomposition adaptée et le couplage chimiométrique des données de différentes techniques analytiques nous ont permis d'obtenir une caractérisation qualitative et quantitative de la pollution des sols industriels par les HAP ciblés.

Néanmoins, nous avons cherché à approfondir la caractérisation de la pollution des sols industriels en ciblant d'autres CAP tels que les dérivés de HAP et les CAP hétérocycliques et leurs dérivés. Nous avons été toutefois confrontés aux limites de la spectroscopie de fluorescence. En effet, la présence d'un hétéroatome au sein d'un des cycles aromatiques d'un CAP ou dans le groupement fonctionnel de substitution d'un dérivé entraîne souvent une diminution du rendement de fluorescence ce qui rend leur détection difficile dans les MEEF. Pour pallier cet inconvénient, il est nécessaire d'utiliser d'autres techniques analytiques tels que la spectroscopie Raman qui présente l'avantage d'être non spécifique aux composés fluorescents. Cette approche permet également d'envisager des stratégies de traitement de données de type multi-blocs pour prendre en compte le mieux possible, via différentes modalités spectrales ou non spectrales, toute la complexité du système physico chimique étudié.

C’est pourquoi, la spectroscopie Raman a été utilisée pour décrire la pollution des sols par les CAP. Pour ce faire, des analyses ont été menées avec un équipement Raman adapté à ce type d’échantillons. Des solutions contenant des CAP purs et d’autres contenant des extraits organiques de sols ont été préparées et analysées. Les résultats obtenus sont prometteurs, car l’analyse des CAP purs démontre la faisabilité et l'adaptabilité de cette technique pour la détection des composés qui peuvent être difficiles à détecter par la spectroscopie de fluorescence. Elle montre également la richesse des spectres Raman et la spécificité des bandes vibrationnelles en relation avec les catégories des CAP. Par ailleurs, l’analyse chimiométrique des extraits organiques des sols pollués nous permet de discriminer la plupart des sols de l'étude et aussi d'établir, en lien avec les *‘loadings’*, les bandes vibrationnelles responsables de cette discrimination. Afin de pousser davantage cette discrimination et de préparer une future stratégie de type multi-blocs en chimiométrie, les spectres Raman ont été transformés en images appelées scalogrammes grâce à la transformée en ondelettes continue. Cette transformée nous a permis de mettre en évidence des caractéristiques dans les données Raman dans le domaine des fréquences. L'analyse chimiométrique de ces images nous a ainsi permis d'obtenir une discrimination plus fine et améliorée entre les différents sols contaminés.

Les résultats de cette thèse ouvrent ainsi de manière générale la voie, à d'autres travaux de recherche visant à réaliser une caractérisation qualitative et quantitative encore plus complète de la contamination des sols par les CAP avec le couplage de techniques et/ou une stratégie multi-blocs. De plus, les 16 modèles SVR qui ont été développés (un modèle par HAP) peuvent être déjà enrichis au fur et à mesure avec d'autres données pour les rendre encore plus robustes et étendre la gamme des concentrations ciblées. En outre, l’algorithme de prétraitement MT-SVD, qui a été développé, est une méthode adaptable pouvant être appliquée à tout type d'images, allant bien au-delà de la fluorescence 3D.

Pour finir, les méthodologies de construction du modèle PARAFAC et des modèles SVR, adoptées dans ce travail peuvent également être envisagées pour l'identification et la quantification d'autres composés organiques fluorescents que les CAP. En outre, notre travail de recherche s'est focalisé sur l'analyse d'une matrice environnementale spécifique, à savoir le sol, en mettant l'accent sur la fraction organique extractible par solvant et porteuse de la pollution par les CAP. Cependant, il est tout à fait envisageable d'ajuster nos méthodologies pour détecter et mesurer des composés hydrosolubles, ce qui contribuerait à compléter la caractérisation des composés polluants présents à la fois dans les matrices environnementales solides et aqueuses.

BIBLIOGRAPHIE

1. Payá Pérez, A. & Rodríguez Eugenio, N. Status of local soil contamination in Europe : revision of the indicator 'Progress in the management contaminated sites in Europe'. *Publ. Off. Eur. Union* (2018) doi:doi/10.2760/093804.

2. Prokop, G., Louwagie, G., Kibblewhite, M., Van Liedekerke, M. & Rabl-Berger, S. *Progress in the management of contaminated sites in Europe*. *Publications Office* https://data.europa.eu/doi/10.2788/4658 (2014) doi:doi/10.2788/4658.

3. Panagos, P., Van Liedekerke, M., Yigini, Y. & Montanarella, L. Contaminated Sites in Europe: Review of the Current Situation Based on Data Collected through a European Network. *J. Environ. Public Health* 158764 (2013) doi:10.1155/2013/158764.

4. Keith, L. H. & Telliard, W. A. Priority pollutants. I. A perspective view. *Environ. Sci. Technol.* **13**, 416–423 (1979).

5. Office of research and development. United states environmental protection agency. *Suspect Carcinogens in Water Supplies*. (1975).

6. Keith, L. H. The Source of U.S. EPA's Sixteen PAH Priority Pollutants. *Polycycl. Aromat. Compd.* **35**, 147–160 (2015).

7. Lundstedt, S. *et al.* First intercomparison study on the analysis of oxygenated polycyclic aromatic hydrocarbons (oxy-PAHs) and nitrogen heterocyclic polycyclic aromatic compounds (N-PACs) in contaminated soil. *TrAC - Trends in Analytical Chemistry* vol. 57 83–92 at https://doi.org/10.1016/j.trac.2014.01.007 (2014).

8. Achten, C. & Andersson, J. T. Overview of Polycyclic Aromatic Compounds (PAC). *Polycycl. Aromat. Compd.* **35**, 177–186 (2015).

9. Adeniji, A. O., Okoh, O. O. & Okoh, A. I. Analytical Methods for Polycyclic Aromatic Hydrocarbons and their Global Trend of Distribution in Water and Sediment: A Review. *Recent Insights Pet. Sci. Eng.* (2018) doi:10.5772/intechopen.71163.

10. Johansson, C. Oxydation par les ferrates d'un sol contaminé par du DNAPL en condition saturée : conséquences sur les Composés Aromatiques Polycycliques (HAP et CAP Polaires) : Expérimentations en batch et colonne. (Université de Lorraine, 2019).

11. Biache, C. Evolution des composants organiques d'un sol de cokerie en contexte d'atténuation naturelle. (Université Henri Poincaré, 2010).

12. Biache, C., Mansuy-Huault, L. & Faure, P. Impact of oxidation and biodegradation on the most commonly used polycyclic aromatic hydrocarbon (PAH) diagnostic ratios: Implications for the source identifications. *J. Hazard. Mater.* **267**, 31–39 (2014).

13. Suess, M. J. The environmental load and cycle of polycyclic aromatic hydrocarbons. *Sci. Total Environ.* **6**, 239–250 (1976).

14. Ahad, J. M. E. *et al.* Polycyclic aromatic compounds (PACs) in the Canadian environment: A review of sampling techniques, strategies and instrumentation. *Environ. Pollut.* **266**, 114988 (2020).

15. Paris, A., Gaillard, J.-L. & Ledauphin, J. Impact of biomass combustion on occurrence and distribution of aromatic hydrocarbons in apples. *Environ. Sci. Pollut. Res.* **27**, 3165–3172 (2020).

16. Mirnaghi, F. S. *et al.* Monitoring of polycyclic aromatic hydrocarbon contamination at four oil spill sites using fluorescence spectroscopy coupled with parallel factor-principal component analysis. *Environ. Sci. Process. Impacts* **21**, 413–426 (2019).

17. Qazi, F., Shahsavari, E., Prawer, S., Ball, A. S. & Tomljenovic-Hanic, S. Detection and identification of polyaromatic hydrocarbons (PAHs) contamination in soil using intrinsic fluorescence. *Environ. Pollut.* **272**, (2021).

18. Mastral, A. M. & Callén, M. S. A review on polycyclic aromatic hydrocarbon (PAH): Emissions from energy generation. *Environ. Sci. Technol.* **34**, 3051–3057 (2000).

19. Luthy, R. G. *et al.* Remediating tar-contaminated soils at manufactured gas plant sites. *Environ. Sci. Technol.* **28**, 266A-276A (1994).

20. Faure, P. Application des techniques de géochimie organique pétrolière à l'étude de problèmes environnementaux : polluants organiques, inertage et stockage des déchets. (Institut national polytechnique de Lorraine, 1999).

21. Titaley, I. A., Simonich, S. L. M. & Larsson, M. Recent Advances in the Study of the Remediation of Polycyclic Aromatic Compound (PAC)-Contaminated Soils: Transformation Products, Toxicity, and Bioavailability Analyses. *Environ. Sci. Technol. Lett.* **7**, 873–882 (2020).

22. Biache, C., Lorgeoux, C., Colombano, S., Saada, A. & Faure, P. Multistep thermodesorption coupled with molecular analyses as a quick, easy and environmentally friendly way to measure PAH availability in contaminated soils. *Talanta* **228**, 122235 (2021).

23. Biache, C., Lorgeoux, C., Saada, A., Colombano, S. & Faure, P. Fast method to quantify PAHs in contaminated soils by direct thermodesorption using analytical pyrolysis. *Talanta* **166**, 241–248 (2017).

24. Paris, A. Extractions et analyses des hydrocarbures aromatiques : approches méthodologiques et applications à des matrices fruitières. *Sci. Agric.* 268 (2017).

25. Boujday, S., de la Chapelle, M. L., Srajer, J. & Knoll, W. Enhanced vibrational spectroscopies as tools for small molecule biosensing. *Sensors (Switzerland)* **15**, 21239–21264 (2015).

26. Kumar, S., Negi, S. & Maiti, P. Biological and analytical techniques used for detection of polyaromatic hydrocarbons. *Environ. Sci. Pollut. Res.* **24**, 25810–25827 (2017).

27. Locquet, N., Aït-Kaddour, A. & Cordella, C. B. Y. *3D Fluorescence Spectroscopy and Its Applications*. *Encyclopedia of Analytical Chemistry* (2018). doi:10.1002/9780470027318.a9540.

28. Lakowicz, J. R. *Principles of fluorescence spectroscopy*. (Springer, 2006). doi:10.1007/978-0-387-46312-4.

29. Lavine, B. & Workman, J. Chemometrics. *Anal. Chem.* **80**, 4519–4531 (2008).

30. Afseth, N. K. & Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **117**, 92–99 (2012).

31. Cohen, J. E. & Bro, R. Nonnegative PARAFAC2: A Flexible Coupling Approach. in *Latent Variable Analysis and Signal Separation* (eds. Deville, Y., Gannot, S., Mason, R., Plumbley, M. D. & Ward, D.) 89–98 (Springer International Publishing, 2018).

32. Jalalvand, A. R. Engagement of chemometrics and analytical electrochemistry for clinical purposes: A review. *Chemom. Intell. Lab. Syst.* **227**, 104612 (2022).

33. Büchele, D., Chao, M., Ostermann, M., Leenen, M. & Bald, I. Multivariate chemometrics as a key tool for prediction of K and Fe in a diverse German agricultural soil-set using EDXRF. *Sci. Rep.* **9**, 17588 (2019).

34. Mishra, P. *et al.* Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends Anal. Chem.* **137**, 116206 (2021).

35. Simon, G. *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*. (Editions des archives contemporaines, 2020). doi:10.17184/eac.9782813002556.

36. Humbert, B., Mevellec, J.-Y., Grausem, J., Dossot, M. & Carteret, C. Spectrométrie d'absorption dans l'infrarouge. *Tech. l'ingénieur* **33**, 0–29 (2012).

37. Perraudeau, M. Lumière et couleur. *Tech. l'ingénieur* **33**, (2004).

38. Dalibart, M. & Servant, L. *Spectroscopie dans l'infrarouge*. (Techniques de l'ingénieur,

2020).

39. Di Benedetto, D. & Breuil, P. *Spectrophotométrie d'absorption dans l'ultraviolet et le visible*. vol. 33 (Techniques de l'ingénieur, 2014).

40. Ranaweera, R. K. R., Capone, D. L., Bastian, S. E. P., Cozzolino, D. & Jeffery, D. W. A Review of Wine Authentication Using Spectroscopic Approaches in Combination with Chemometrics. *Molecules* **26**, (2021).

41. Stokes, G. G. On the Change of Refrangibility of Light. *Philos. Trans. R. Soc. London* **142**, 463–562 (1852).

42. Demchenko, A. P. *Introduction to fluorescence sensing*. vol. 1 (Springer Cham, 2009).

43. Caland, F. Décomposition tensorielle de signaux luminescents émis par des biosenseurs bactériens pour l'identification de Systèmes Métaux-Bactéries. (Université de Lorraine, 2013).

44. Valeur, B. *Invitation à la fluorescence moléculaire*. (De Boeck, 2004).

45. Valeur, B. & Berberan-Santos, M. N. *Molecular Fluorescence: Principles and Applications*. (Wiley, 2012).

46. Guilbault, G. G. *Practical Fluorescence, Second Edition*. (CRC Press, 1990).

47. Luciani, X. Analyse numérique des spectres de fluorescence 3D issus de mélanges non linéaires. (Université du sud Toulon Var, 2007).

48. Hamla, S. Mid-infrared spectroscopy and confocal Raman imaging for the analysis of therapeutic proteins. (Université de Liège, 2023).

49. Mille, G., Guiliano, M. & Kister, J. Analysis and evolution of coals: UV fluorescence spectroscopy study (demineralized coals-oxidized coals). *Org. Geochem.* **13**, 947–952 (1988).

50. Schlepp, L. Réactivité des bitumes d'enrobage des déchets ultimes: comparaison avec un analogue naturel. (Institut national polytechnique de Lorraine, 2000).

51. Adeniji, A. O., Okoh, O. O. & Okoh, A. I. Analytical Methods for Polycyclic Aromatic Hydrocarbons and their Global Trend of Distribution in Water and Sediment: A Review. in (ed. Zoveidavianpoor, M.) Ch. 19 (IntechOpen, 2017). doi:10.5772/intechopen.71163.

52. Tommasini, M. & Zerbi, G. A Theoretical Raman Study on Polycyclic Aromatic Hydrocarbons of Environmental Interest. *Chem. Eng. Trans.* **22**, 263-268 SE-Research Articles (2010).

53. Chen, J., Huang, Y.-W. & Zhao, Y. Characterization of polycyclic aromatic hydrocarbons using Raman and surface-enhanced Raman spectroscopy. *J. Raman Spectrosc.* **46**, 64–69 (2015).

54. Cloutis, E., Szymanski, P., Applin, D. & Goltz, D. Identification and discrimination of polycyclic aromatic hydrocarbons using Raman spectroscopy. *Icarus* **274**, 211–230 (2016).

55. Izawa, M. R. M., Applin, D. M., Norman, L. & Cloutis, E. A. Reflectance spectroscopy (350–2500nm) of solid-state polycyclic aromatic hydrocarbons (PAHs). *Icarus* **237**, 159–181 (2014).

56. Pejcic, B. *et al.* Direct quantification of aromatic hydrocarbons in geochemical fluids with a mid-infrared attenuated total reflection sensor. *Org. Geochem.* **55**, 63–71 (2013).

57. Rivera-Figueroa, A. M., Ramazan, K. A. & Finlayson-Pitts, B. J. Fluorescence, Absorption, and Excitation Spectra of Polycyclic Aromatic Hydrocarbons as a Tool for Quantitative Analysis. *J. Chem. Educ.* **81**, 242 (2004).

58. Qazi, F., Shahsavari, E., Prawer, S., Ball, A. S. & Tomljenovic-Hanic, S. Detection and identification of polyaromatic hydrocarbons (PAHs) contamination in soil using intrinsic fluorescence. *Environ. Pollut.* **272**, 116010 (2021).

59. Murphy, K. R., Stedmon, C. A., Graeber, D. & Bro, R. Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Anal. Methods* **5**, 6557–6566 (2013).

60. Kowalski, B. R. Chemometrics: Views and Propositions. *J. Chem. Inf. Comput. Sci.* **15**, 201–203 (1975).

61. Cordella, C. L'analyse en composantes principales Une des techniques fondatrices de la chimiométrie. *l'actualité Chim.* **345**, 13–18 (2010).

62. Offroy, M., Moreau, M., Sobanska, S., Milanfar, P. & Duponchel, L. Pushing back the limits of Raman imaging by coupling super-resolution and chemometrics for aerosols characterization. *Sci. Rep.* **5**, 12303 (2015).

63. Bro, R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.* **38**, 149–171 (1997).

64. Ruckebusch, C. Résolution et modélisation chimiométrique en spectroscopie moléculaire. (Université des Sciences et Technologies de Lille, 2008).

65. Bro, R. Multiway calibration. Multilinear PLS. *J. Chemom.* **10**, 47–61 (1996).

66. Bi, L., Tsimhoni, O. & Liu, Y. Using the Support Vector Regression Approach to Model Human Performance. *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans* **41**, 410–417 (2011).

67. Blanchet, L. Méthodes de résolution dédiées à l'étude spectroscopique de processus photoinduits : adaptation aux spécificités des spectres résolus en temps. (Université des sciences et technologies de Lille, 2008).

68. Ho, C. N., Christian, G. D. & Davidson, E. R. Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer. *Anal. Chem.* **50**, 1108–1113 (1978).

69. Nie, J. F. *et al.* Simultaneous determination of 6-methylcoumarin and 7-methoxycoumarin in cosmetics using three-dimensional excitation-emission matrix fluorescence coupled with second-order calibration methods. *Talanta* **75**, 1260–1269 (2008).

70. Bro, R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics and Intelligent Laboratory Systems* vol. 46 133–147 at https://doi.org/10.1016/S0169-7439(98)00181-6 (1999).

71. Thygesen, L. G., Rinnan, Å., Barsberg, S. & Møller, J. K. S. Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area. *Chemom. Intell. Lab. Syst.* **71**, 97–106 (2004).

72. Rinnan, Å. & Andersen, C. M. Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation–emission data. *Chemom. Intell. Lab. Syst.* **76**, 91–99 (2005).

73. Bahram, M., Bro, R., Stedmon, C. & Afkhami, A. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *J. Chemom.* **20**, 99–105 (2006).

74. Yu, S., Xiao, X. & Xu, G. Eliminating Rayleigh and Raman Scattering in Three-Dimensional Fluorescence Spectroscopy by Kriging Interpolation. *J. Appl. Spectrosc.* **83**, 786–791 (2016).

75. Roger, J.-M. & Ecarnot, M. Chimiometrie chapitre 1/2 : les méthodes non supervisées, grain5:prétraitement. in *CheMOOCs* (2022).

76. Wise, B. M. *et al. Chemometrics Tutorial for PLS_Toolbox and Solo*. (Eigenvector Research, Inc, 2006).

77. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).

78. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).

79. Preys, S. Chimiométrie chapitre 1/2: les méthodes non supervisées, grain4: ACP2. in *CheMOOCs* (2022).

80. Jolliffe, I. *Principal Component Analysis*. (Springer Science & Business Media, 2002).

81. Vandeginste, B. G. M., Sielhorst, C. & Gerritsen, M. The NIPALS algorithm for the calculation of the principal components of a matrix. *TrAC Trends Anal. Chem.* **7**, 286–287 (1988).

82. Weinstein, R. Matrix Methods of Linear Algebra. https://github.com/MathWorks-Teaching-Resources/Matrix-Methods-of-Linear-Algebra/releases/tag/v1.0.4 (2023).

83. Kosanovich, K. A., Piovoso, M. J., Dahl, K. S., MacGregor, J. F. & Nomikos, P. Multi-way PCA applied to an industrial batch process. in *Proceedings of 1994 American Control Conference - ACC '94* vol. 2 1294–1298 vol.2 (1994).

84. Harshman, R. A. Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-model factor analysis. *UCLA Work. Pap. Phonetics* **16**, 1–84 (1970).

85. Carroll, J. & Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 283–319 (1970).

86. Bro, R. & Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* **17**, 274–286 (2003).

87. Rao, C. . & Mitra, S. . Generalized Inverse of Matrices and its Applications. *J. R. Stat. Soc. Ser. A* (1971) doi:10.2307/2344631.

88. Bro, R. Multi-way Analysis in the Food Industry. *Food Technol.* **41**, 3545–3564 (1998).

89. De Juan, A. & Tauler, R. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review. *Anal. Chim. Acta* **1145**, 59–78 (2021).

90. Kruskal, J. B. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18**, 95–138 (1977).

91. Stegeman, A. & Sidiropoulos, N. D. On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition. *Linear Algebra Appl.* **420**, 540–552 (2007).

92. Hanafi, M., Jaillais, B. & Qannari, E.-M. Chimiometrie chapitre 1/2 : les méthodes non supervisées, grain25:analyse des données 3 voies. in *CheMOOCs* (2022).

93. Harshman, R. A. Determination and proof of minimum uniqueness conditions for Parafac-1. *UCLA Work. Pap. Phonetics* **22**, 111–117 (1972).

94. Harshman, R. A. How can i know if it's 'real' ? a catalog of diagnostics for use with three-mode factor analysis and multidimensional scaling. *Res. methods multimode data Anal.* 566–591 (1984).

95. Sanchez, E. & Kowalski, B. R. Tensorial resolution: A direct trilinear decomposition. *J. Chemom.* **4**, 29–45 (1990).

96. Nazmi Liyana Mohd Napi, N. *et al.* Multiple Linear Regression (MLR) and Principal Component Regression (PCR) for Ozone (O3) Concentrations Prediction. *IOP Conf. Ser. Earth Environ. Sci.* **616**, 12004 (2020).

97. Rutledge, D. Chimiométrie chapitre 1/2: les méthodes non supervisées, grain8: Régression PLS. in *CheMOOCs* (2022).

98. Wold, H. Estimation of Principal Components and Related Models by Iterative Least Squares. *Multivar. Anal.* 391–420 (1966).

99. Wold, H. Path Models with Latent Variables: The NIPALS Approach. in *Quantitative Sociology International Perspectives on Mathematical and Statistical Modeling* (eds. Blalock, H. M., Aganbegian, A., Borodkin, F. M., Boudon, R. & Capecchi, V. B. T.-Q. S.) 307–357 (Academic Press, 1975). doi:https://doi.org/10.1016/B978-0-12-103950-9.50017-4.

100. Hamla, S. *et al.* A new alternative tool to analyse glycosylation in pharmaceutical proteins based on infrared spectroscopy combined with nonlinear support vector regression. *Analyst* **147**, 1086–1098 (2022).

101. Boser, B. E., Guyon, I. M. & Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 144–152 (Association for Computing Machinery, 1992). doi:10.1145/130385.130401.

102. Laouti, N. Diagnostic de défauts par les Machines à Vecteurs Supports : application à différents systèmes mutivariables nonlinéaires. (Université de Lyon, 2012).

103. Hamla, S. Mid-infrared spectroscopy and confocal Raman imaging for the analysis of therapeutic proteins. (Université de Liège, 2022).

104. Hamla, S. *et al.* A New Alternative Tool to Analyse Glycosylation in Monoclonal Antibodies Based on Drop-Coating Deposition Raman imaging: A Proof of Concept. *Molecules* vol. 27 at https://doi.org/10.3390/molecules27144405 (2022).

105. Saim, N., Dean, J. R., Abdullah, M. P. & Zakaria, Z. An Experimental Design Approach for the Determination of Polycyclic Aromatic Hydrocarbons from Highly Contaminated Soil Using Accelerated Solvent Extraction. *Anal. Chem.* **70**, 420–424 (1998).

106. Buco, S., Moragues, M., Sergent, M., Doumenq, P. & Mille, G. An experimental design approach for optimizing polycyclic aromatic hydrocarbon analysis in contaminated soil by pyrolyser-gas chromatography-mass spectrometry. *Environ. Res.* **104**, 209–215 (2007).

107. Lamparczyk, H., Ochocka, R. J., Grzybowski, J., Halkiewicz, J. & Radecki, A. Classification of marine environment samples based on chromatographic analysis of hydrocarbons and principal component analysis. *Oil Chem. Pollut.* **6**, 177–193 (1990).

108. Rojo-Nieto, E., Sales, D. & Perales, J. A. Sources, transport and fate of PAHs in sediments and superficial water of a chronically polluted semi-enclosed body of seawater: linking of compartments. *Environ. Sci. Process. Impacts* **15**, 986–995 (2013).

109. Bosco, M. V & Larrechi, M. S. PARAFAC and MCR-ALS applied to the quantitative monitoring of the photodegradation process of polycyclic aromatic hydrocarbons using three-dimensional excitation emission fluorescent spectra: Comparative results with HPLC. *Talanta* **71**, 1703–1709 (2007).

110. Ferretto, N. *et al.* Identification and quantification of known polycyclic aromatic hydrocarbons and pesticides in complex mixtures using fluorescence excitation–emission matrices and parallel factor analysis. *Chemosphere* **107**, 344–353 (2014).

111. Driskill, A. K., Alvey, J., Dotson, A. D. & Tomco, P. L. Monitoring polycyclic aromatic hydrocarbon (PAH) attenuation in Arctic waters using fluorescence spectroscopy. *Cold Reg. Sci. Technol.* **145**, 76–85 (2018).

112. Feng, Y. *et al.* Raman-infrared spectral fusion combined with partial least squares (PLS) for quantitative analysis of polycyclic aromatic hydrocarbons in soil. *Anal. Methods* **12**, 1203–1211 (2020).

113. Li, H. *et al.* Quantitative analysis of phenanthrene in soil by fluorescence spectroscopy coupled with the CARS-PLS model. *RSC Adv.* **13**, 9353–9360 (2023).

114. Li, M. *et al.* Quantitative analysis of polycyclic aromatic hydrocarbons in soil by infrared spectroscopy combined with hybrid variable selection strategy and partial least squares. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **257**, 119771 (2021).

115. Asghari, A., Haj Hosseini, A. & Ghajarbeygi, P. Fast and non-destructive determination of histamine in tuna fish by ATR-FTIR spectroscopy combined with PLS calibration method. *Infrared Phys. Technol.* **123**, 104093 (2022).

116. Carlos Cobas, J., Bernstein, M. A., Martín-Pastor, M. & Tahoces, P. G. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *J. Magn. Reson.* **183**, 145–151 (2006).

117. Brunton, S. L. & Kutz, J. N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. (Cambridge University Press, 2019). doi:DOI: 10.1017/9781108380690.

118. Offroy, M. Développement de la super-résolution appliquée à l'imagerie des spectroscopies vibrationnelles. (Université de Lille 1, 2012).

119. Donoho, D., Gavish, M. & Romanov, E. *ScreeNOT*: Exact MSE-optimal singular value thresholding in correlated noise. *Ann. Stat.* **51**, 122–148 (2023).

120. Gavish, M. & Donoho, D. L. The Optimal Hard Threshold for Singular Values is 4/sqrt (3). *IEEE Trans. Inf. Theory* **60**, 5040–5053 (2014).

121. Gonzalez, R. C., Woods, R. E. & Eddins, S. L. *Digital Image processing using MATLAB*. (Pearson/Prentice Hall, 2004).

122. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man. Cybern.* **9**, 62–66 (1979).

123. Tauler, R. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *J. Chemom.* **15**, 627–646 (2001).

124. Bertrand, O. Enregistrement moléculaire de changements d'usage des sols et de pressions anthropiques : l'exemple d'un étang piscicole (Lansquenet, Lorraine). (Université de Lorraine, 2012).

125. Bertrand, O. *et al.* A possible terrigenous origin for perylene based on a sedimentary record of a pond (Lorraine, France). *Org. Geochem.* **58**, 69–77 (2013).

126. Boulangé, M., Lorgeoux, C., Biache, C., Saada, A. & Faure, P. Fenton-like and potassium permanganate oxidations of PAH-contaminated soils: Impact of oxidant

doses on PAH and polar PAC (polycyclic aromatic compound) behavior. *Chemosphere* **224**, 437–444 (2019).

127. Boulangé, M. Mobilisation et transfert des composés aromatiques polycycliques (HAP et CAP polaires) dans les sols historiquement contaminés par des goudrons de houille : expérimentations au laboratoire et in situ. (Université de Lorraines, 2017).

128. Ballabio, D. & Consonni, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal. Methods* **5**, 3790–3798 (2013).

129. Afseth, N. K., Segtnan, V. H. & Wold, J. P. Raman Spectra of Biological Samples: A Study of Preprocessing Methods. *Appl. Spectrosc.* **60**, 1358–1367 (2006).

130. Alsberg, B. K., Woodward, A. M. & Kell, D. B. An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemom. Intell. Lab. Syst.* **37**, 215–239 (1997).

131. Truchetet, F. *Ondelettes pour le signal numérique*. (Hermès, 1998).

132. Daubechies, I. *Ten Lectures on Wavelets*. (Society for Industrial and Applied Mathematics, 1992).

133. Hanafi, M., Mazerolles, G., Dufour, E. & Qannari, E. M. Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques. *J. Chemom.* **20**, 172–183 (2006).

134. Offroy, M. & Duponchel, L. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Anal. Chim. Acta* **910**, 1–11 (2016).

135. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **50**, 1–18 (2000).

136. Böke, J. S., Popp, J. & Krafft, C. Optical photothermal infrared spectroscopy with simultaneously acquired Raman spectroscopy for two-dimensional microplastic identification. *Sci. Rep.* **12**, 18785 (2022).

137. Kruse, F. A. *et al.* The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote Sens. Environ.* **44**, 145–163 (1993).

138. Sajdak, M. Application of chemometrics to identifying solid fuels and their origin. *Open Chem.* **11**, 151–159 (2013).

139. Xie, S., Lawnizak, A., Lio, P. & Krishnan, S. Feature Extraction by Multi-Scale Principal Component Analysis and Classification in Spectral Domain. *Engineering* **5**, 268–271 (2013).

140. Chen, X., Han, Y. & Wu, J. Burst Noise Measuring on the Basis of Wavelet and Fourier Transform. in *2010 International Conference on Measuring Technology and Mechatronics Automation* vol. 1 769–771 (2010).

141. Torrésani, B. *Analyse continue par ondelettes*. (EDP Sciences, 2012).

ANNEXES

Listes des annexes

Il existe plusieurs méthodes permettant la mesure de la polarité d'un solvant. Les données de polarité fournies dans le

Tableau A.1 : se base sur le calcul de l'énergie de transfert (E^T), considérée comme constante de polarité, d'un colorant solvaté, le 2,6-diphényl-4-(2,4,6-triphényl-1-pyridinio) phénoxyde, par les différents solvant en mesurant l'énergie correspondant à la bande d'absorption à la plus forte longueur d'onde (λ_{max}) du colorant dans les différents solvants :

$$E^T = 1{,}196 \times 10^{-2} \lambda_{max} \quad \text{(A.1)}$$

Tableau A.1 : constantes de polarité des différents solvants organiques usuellement utilisés pour l'extraction et/ou la solvatation des HAP. Ordre croissant de polarité.

Nom du solvant	**Formule brute**	**Constante de polarité** (E^T) $\mathrm{kJ.mol^{-1}}$
Hexane	C_6H_{14}	129,16
Cyclohexane	C_6H_{12}	130,42
Toluène	C_7H_8	141,71
Chloroforme	$CHCl_3$	163,44
Dichlorométhane	CH_2Cl_2	171,80
Acétone	C_3H_6O	176,39
Acétonitrile	C_2H_3N	192,28
Alcool isopropylique	C_3H_8O	203,15
Méthanol	CH_3OH	231,99

B.1) Introduction

L'identification des '*loadings*' est une opération principalement réalisée de manière visuelle. Cependant, lorsque les signatures spectrales des '*loadings*' sont très similaires entre elles ou qu'un '*loading*' est une combinaison de plusieurs signaux sources, il est toujours utile d'avoir à disposition un outil informatique ergonomique capable d'aider rapidement et de manière fiable pour cette tâche. C'est pourquoi, une interface graphique utilisateur '*Graphical User Interface*' (GUI) a été développée sous l'environnement MATLAB® dans le cadre de cette thèse pour effectuer le calcul de distances métriques ou d'angles spectraux entre une MEEF de type '*loading*' inconnu et des MEEF d'espèces de références (CAP) préalablement disposées dans une base de données. Les distances métriques (resp. angles spectraux) sont des mesures mathématiques utilisées pour calculer une longueur (resp. un angle) séparant deux vecteurs dans un espace vectoriel. Plus la distance (resp. l'angle) est faible entre deux vecteurs, plus ces vecteurs se ressemblent.

B.2) Présentation de l'interface graphique utilisateur (GUI)

La GUI est structurée en 5 fenêtres de navigation, à savoir une fenêtre dédiée à l'importation de la MEEF du '*loading*' inconnu, une fenêtre pour son exploration visuelle, une autre pour le calcul des distances métriques entre la MEEF inconnue et les MEEF de référence, une pour fournir des informations pratiques sur la distance métrique choisie, et enfin la dernière pour l'affichage des MEEF de référence (Figure B.1).

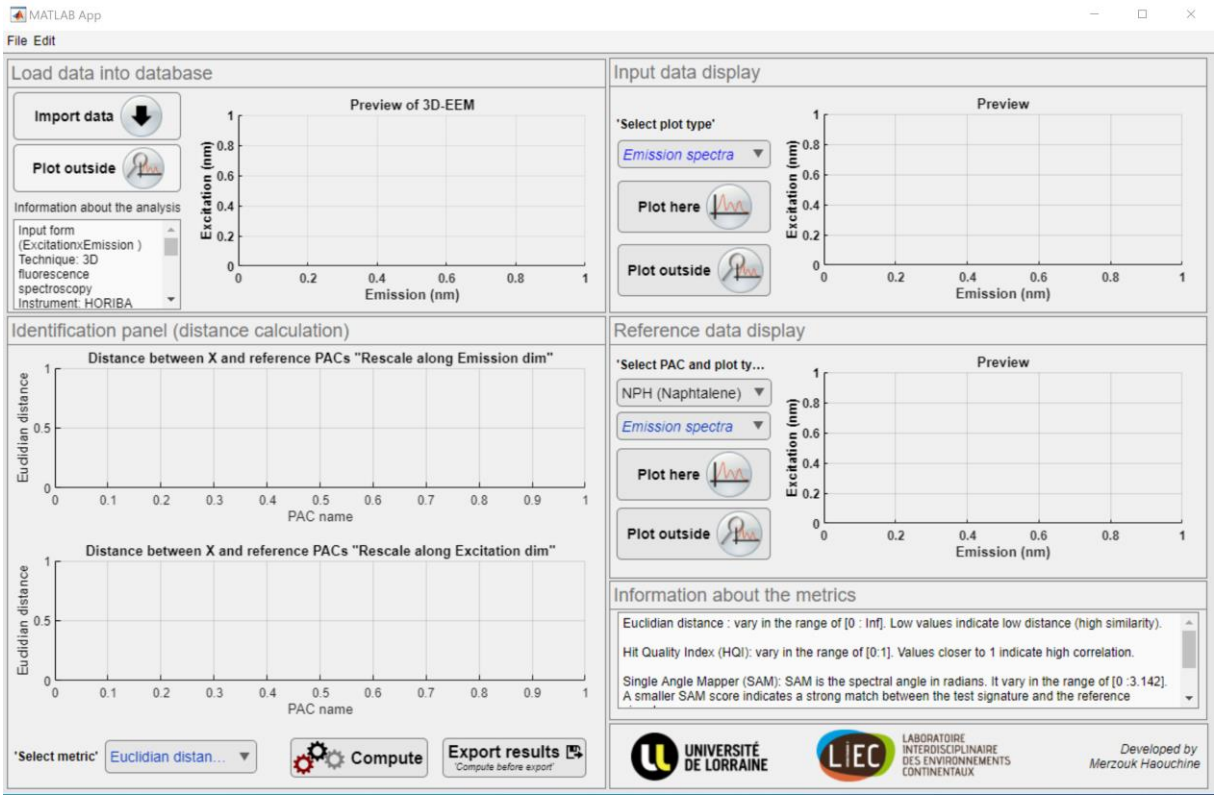


Figure B.1 : Page d'accueil de l'interface graphique utilisateur (GUI).

B.2.1) Importation de la MEEF du '*loading*' inconnu

Avant toute utilisation de la GUI, il est nécessaire de lire d'abord le panneau intitulé '***Information about the analysis***'. Ce panneau fournit à l'utilisateur les informations nécessaires pour une bonne utilisation de la GUI. En effet, il liste le nom de la technique analytique et de l'instrument utilisés, ainsi que les conditions d'acquisition des MEEF (nature, type et taille des données, temps d'intégration, gamme et résolution spectrale en émission et en excitation) qui doivent être respectées. Il mentionne également le prétraitement utilisé pour la correction des MEEF, à savoir la MT-SVD, ainsi que les formats de données acceptés (i.e. les fichiers .dat, .txt et .mat) (Figure B.2).

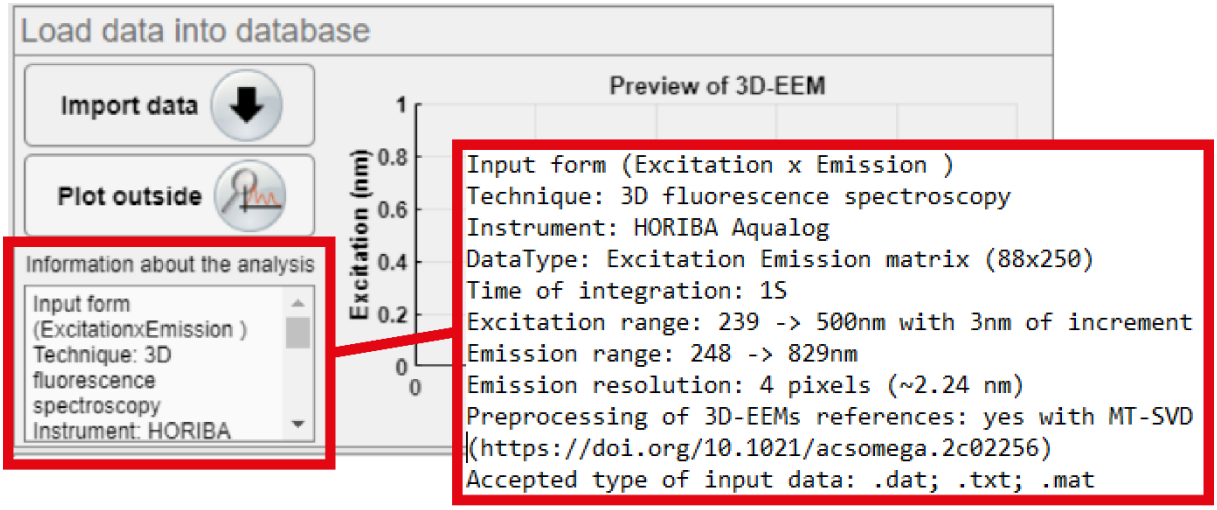


Figure B.2 : Panneau 'information about the analysis'.

Ensuite, en appuyant sur le bouton ***'Import data***', une fenêtre s'ouvre, nous permettant d'accéder à l'endroit de l'ordinateur où est stockée la donnée (MEEF) que nous souhaitons importer. Dans l'exemple illustré dans la Figure B.3a, nous sélectionnons le '*Loading1*' comme donnée d'entrée. Un aperçu en 3D ('*mesh plot*') de la MEEF du loading1 est affiché alors dans la section '***Preview of 3D-EEM***' (Figure B.3b). Enfin, il est également possible d'actionner le bouton '***Plot outside***' pour visualiser le '*mesh plot*' dans une fenêtre indépendante de la GUI.

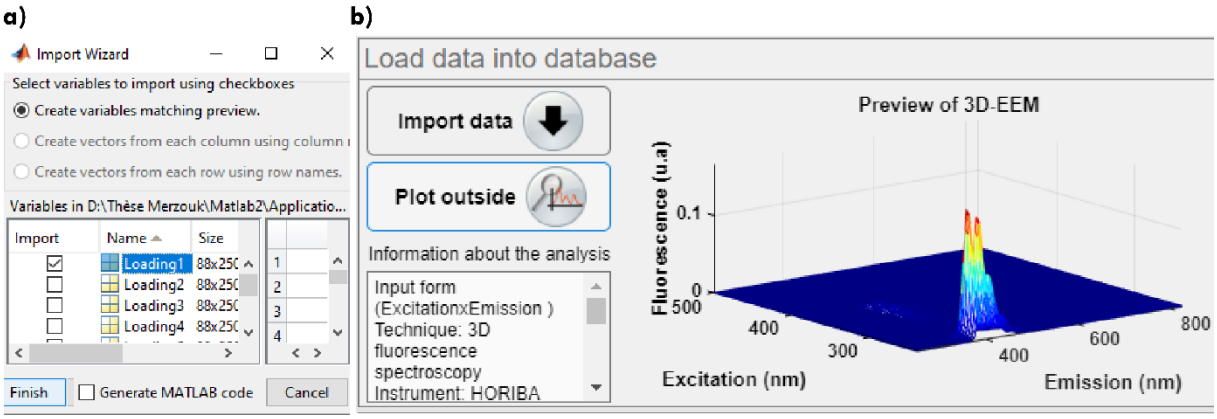


Figure B.3 : Assistant d'importation des données (a). Prévisualisation de la MEEF d'entrée (Loading1) (b).

B.2.2) Exploration visuelle de la MEEF d'entrée

Une fois la MEEF d'intérêt (Loading1) importée et prévisualisée, il est possible d'explorer davantage cette donnée grâce à la deuxième fenêtre. Cela est réalisé via le menu déroulant '***select plot type***', qui propose différents types de tracés. On peut ainsi choisir le tracé des spectres d'émission, le tracé des spectres d'excitation, le tracé du spectre d'émission moyen et médian, ou du spectre d'excitation moyen et médian. On peut également afficher l'image pixélisée de la MEEF ou tracer les contours de l'image (Figure B.4). Il est également possible d'actionner le bouton '***Plot outside***' pour visualiser le tracé choisi dans une fenêtre

indépendante de la GUI. Cet affichage de la MEEF d'entrée est utile au début de l'analyse pour mieux comprendre les données d'entrée, mais il peut également être utilisée à la fin de l'identification pour confirmer visuellement les résultats des calculs qui seront décrits par la suite.

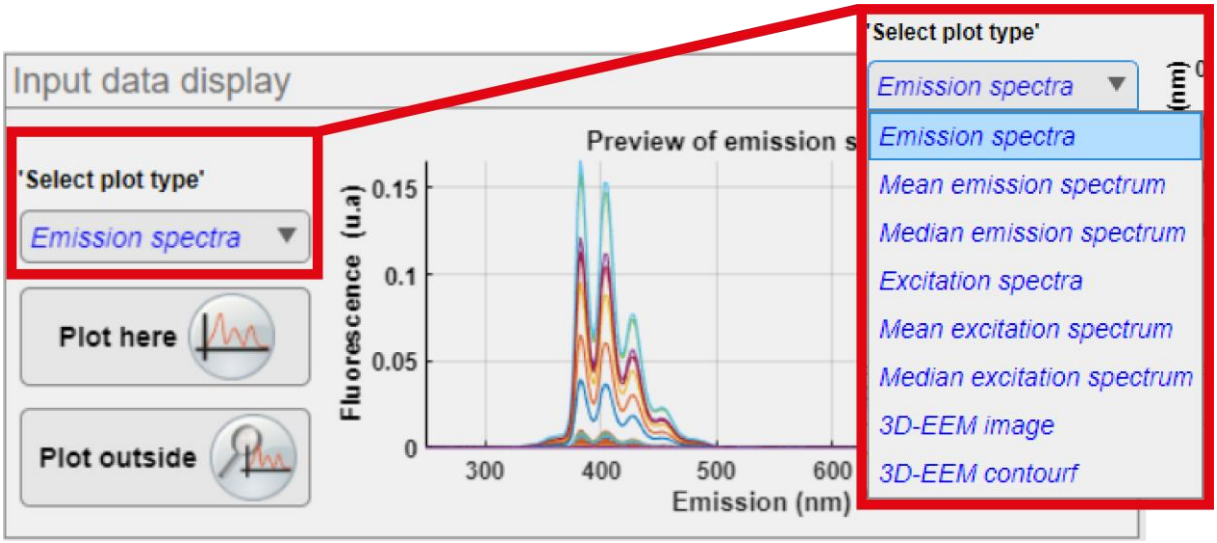


Figure B.4 : Exploration graphique de la MEEF d'entrée (Loading1).

B.2.3) Calcul des distances métriques entre la MEEF inconnue et les MEEF de référence

Pour rappel, une MEEF est une image (i.e. une matrice) de dimensions $Excitation \times Emission$, mais c'est aussi une collection de spectres (i.e. des vecteurs) d'émission et de spectres d'excitation. Dans les calculs qui vont suivre, l'intérêt est porté sur le caractère spectral (i.e. vectoriel) de la MEEF, mais comme les deux collections de spectres sont considérées, l'intérêt est aussi porté indirectement sur le caractère image de la MEEF.

Comme mentionné dans le panneau intitulé '***Information about the analysis***', la GUI est conçue pour recevoir des MEEF de taille $(88{,}250)$, comprenant 88 spectres d'émission et 250 spectres d'excitation. Avant tout calcul, chaque spectre d'excitation et d'émission est préalablement normalisé à l'aide d'une normalisation min-max de la MEEF d'entrée (normalisation valable également pour toutes les MEEF de référence). Cette normalisation permet d'avoir une échelle d'intensités commune entre la MEEF de '*loading*' inconnu et la MEEF de référence à laquelle elle sera comparée, puisqu'elle consiste à mettre à l'échelle l'ensemble des intensités dans un intervalle $[a, b]$ via l'équation :

$$\mathbf{MEEF}_{i,j}^{\text{normalisée}} = a + \left(\frac{\mathbf{MEEF}_{i,j} - \min(\mathbf{MEEF})}{\max(\mathbf{MEEF}) - \min(\mathbf{MEEF})} \right) \times (b - a) \qquad \text{(B.1)}$$

Dans notre cas, $a = 0$ et $b = 1$.

Maintenant que les données sont préparées, les algorithmes de calcul des distances métriques entre vecteurs, à savoir la distance euclidienne, l'indice de qualité de détection '*Hit Quality Index* '(HQI) et le mappage angulaire spectral '*Spectral Angle Mapper*' (SAM) peuvent être déployés. L'utilisateur peut choisir l'algorithme qu'il souhaite utiliser grâce au menu déroulant '***Select metric***'. La distance euclidienne est définie par défaut. Ensuite, il lance le calcul en cliquant sur le bouton '***Compute***', et les résultats s'affichent dans le graphique supérieur pour les calculs relatifs aux 88 spectres d'émission et dans le graphique inférieur pour les calculs relatifs aux 250 spectres d'excitation. Pour chacun des graphiques, nous remarquons en abscisse le nom de la molécule de référence et en ordonnée la valeur de la distance entre celle-ci et la MEEF du '*loading*' inconnu. Les résultats des calculs peuvent ensuite être exportés en cliquant sur le bouton '***Export results***' (Figure B.5a). De plus, l'utilisateur dispose des informations pratiques sur la distance métrique choisie dans la fenêtre du même nom (Figure B.5b).

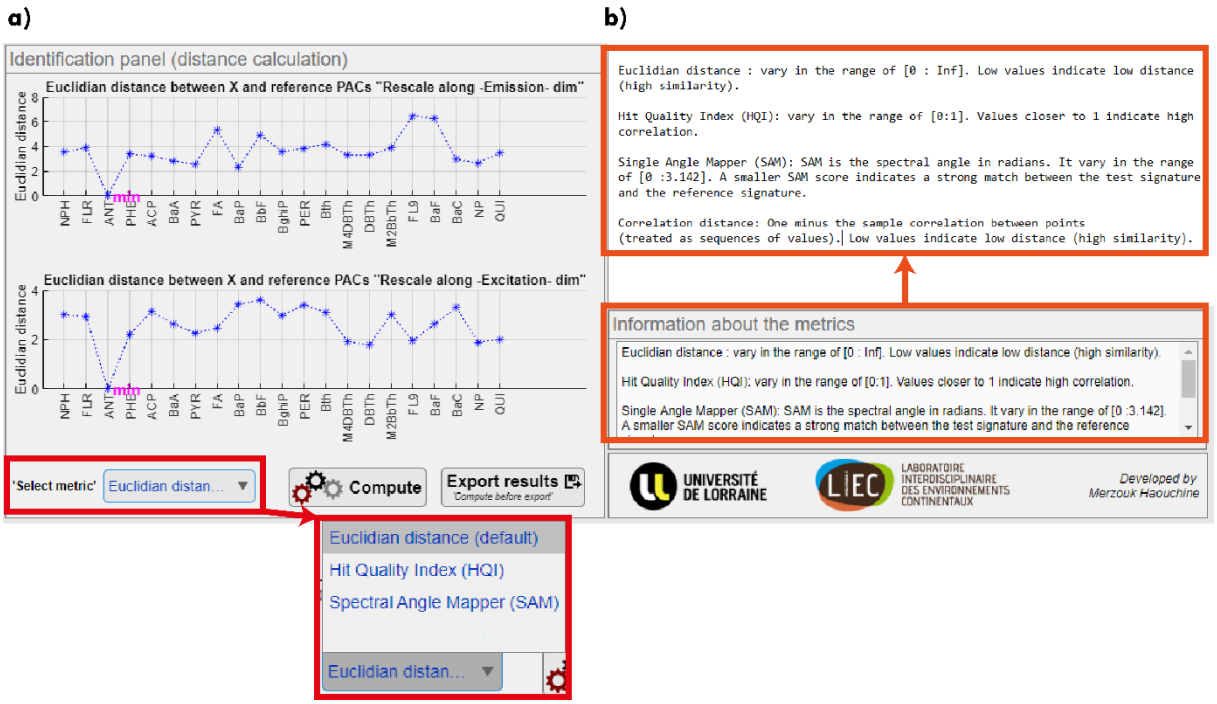


Figure B.5 : Identification de la MEEF du 'loading' inconnu (a) et informations pratiques concernant les algorithmes des distances métriques (b).

B.2.3.1) Distance euclidienne

La distance euclidienne mesure la longueur du segment reliant deux vecteurs dans l'espace vectoriel qu'ils forment. Elle varie dans l'intervalle $[0 : +\infty]$. Les faibles valeurs indiquent une faible distance et donc une grande similarité entre les vecteurs. La distance euclidienne entre chaque spectre d'excitation du '*loading*' $\mathbf{s}_{ex}^{\text{load}}$ et spectre d'excitation de la MEEF de référence $\mathbf{s}_{ex}^{\text{réf}}$ est définie par la relation[135] :

$$d_{\mathbf{s}_{ex}^{\text{load}}\mathbf{s}_{ex}^{\text{réf}}} = \sqrt[2]{\sum_{ex=1}^{88}(s_{ex}^{\text{load}} - s_{ex}^{réf})^2} \quad \text{(B.2)}$$

Où s_{ex}^{load} est la valeur du spectre d'excitation à la longueur d'onde ex. Aussi, la distance euclidienne entre chaque spectre d'émission $\mathbf{s}_{em}^{\text{load}}$ du '*loading*' et spectre d'émission de la MEEF de référence $\mathbf{s}_{em}^{\text{réf}}$ est définie par la relation :

$$d_{\mathbf{s}_{em}^{\text{load}}\mathbf{s}_{em}^{\text{réf}}} = \sqrt[2]{\sum_{em=1}^{250}(s_{em}^{\text{load}} - s_{em}^{réf})^2} \quad \text{(B.3)}$$

où s_{em}^{load} est la valeur du spectre d'émission à la longueur d'onde em.

Au total, pour une MEEF, 250 distances euclidiennes sont calculées entre les spectres d'excitation du '*loading*' et ceux de la MEEF de référence, et 88 distances euclidiennes sont calculées entre les spectres d'émission du '*loading*' et ceux de la MEEF de référence. Ce calcul est réalisé autant de fois qu'il y a de MEEF de références. Ensuite, la valeur maximale pour chacune des dimensions est sélectionnée car elle représente la plus grande dissimilarité observée au sein de l'ensemble des spectres. Enfin, le programme de la GUI compare les différentes distances en fonction des espèces de références et repère la valeur minimale qu'il indique comme étant la correspondance optimale. Par exemple, dans le cas du '*loading1*', nous observons que l'anthracène (ANT) est l'espèce qui lui correspond aussi bien en excitation qu'en émission (Figure B.5a).

B.2.3.2) Indice de qualité de détection 'Hit Quality Index'

L'indice HQI est une mesure d'orthogonalité entre deux vecteurs (i.e. spectres). Il se base sur le calcul du cosinus de l'angle formé par ces vecteurs. Ainsi, sa valeur varie entre 0 pour un angle à 90° et 1 pour un angle à 0°. Deux spectres similaires correspondent à deux vecteurs parallèles ayant une forte dépendance linéaire, avec un HQI égal à 1. En revanche, deux spectres très différents correspondent à deux vecteurs orthogonaux linéairement indépendants, avec un HQI qui se rapproche de 0[136]. Dans notre cas, l'algorithme HQI a été implémenté pour calculer la ressemblance entre le spectre moyen d'excitation du '*loading*' $\mathbf{s}_{ex}^{\text{load}-\text{moy.}}$ et le spectre moyen d'excitation de la MEEF de référence $\mathbf{s}_{ex}^{\text{réf}-\text{moy.}}$ par la relation :

$$\text{HQI}\left(\mathbf{s}_{ex}^{\text{load}-\text{moy.}}, \mathbf{s}_{ex}^{\text{réf}-\text{moy.}}\right) = \frac{(\mathbf{s}_{ex}^{\text{load}-\text{moy.}} \cdot \mathbf{s}_{ex}^{\text{réf}-\text{moy.}})^2}{(\mathbf{s}_{ex}^{\text{load}-\text{moy.}} \cdot \mathbf{s}_{ex}^{\text{load}-\text{moy.}}) \times (\mathbf{s}_{ex}^{\text{réf}-\text{moy.}} \cdot \mathbf{s}_{ex}^{\text{réf}-\text{moy.}})} \quad \text{(B.4)}$$

Où l'opérateur · signifie le produit scalaire. Par ailleurs, le même calcul est réalisé pour estimer la ressemblance entre le spectre moyen d'émission du '*loading*' et le spectre moyen

d'émission de la MEEF de référence. Enfin, le programme de la GUI compare les différents HQI en fonction des espèces de références et repère la valeur maximale qu'il indique comme étant la correspondance optimale. Concernant la MEEF du Loading1 qui nous sert d'exemple, on remarque que l'anthracène (ANT) est l'espèce qui lui concorde aussi bien en excitation qu'en émission. On note également une valeur HQI élevée pour le phénanthrène (PHE) sur le graphique de l'excitation ce qui est cohérent puisque le PHE est un isomère de l'ANT. Cependant, ce résultat démontre l'intérêt de travailler sur les deux dimensions en parallèle, puisque, à l'instar du PHE et de l'ANT, il arrive que deux composés présentent des profils similaires sur une dimension et dissimilaires sur une autre (Figure B.6).

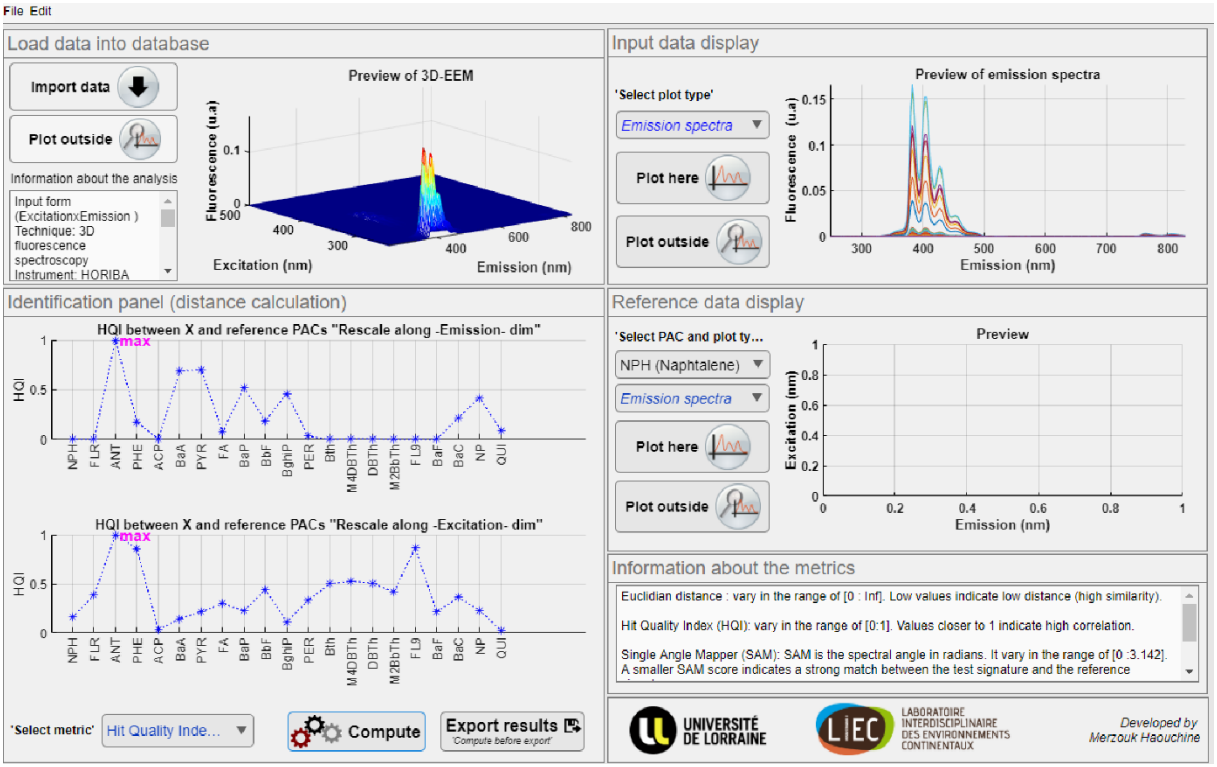


Figure B.6 : Résultats des indices HQI sur la MEEF du Loading1.

B.2.3.3) 'Spectral Angle Mapper'

Le SAM est un autre algorithme mesurant l'orthogonalité entre deux vecteurs (i.e. spectres). Sa différence avec le HQI se situe dans le fait qu'il mesure l'angle et non le cosinus de l'angle entre les deux vecteurs. Ainsi, sa valeur se situe entre 0° (=0 radians) et 180° (=π radians). Plus l'angle entre deux vecteurs est petit, plus la ressemblance entre les deux spectres est grande, ce qui indique une forte correspondance. Comme pour le HQI, l'algorithme SAM a été implémenté pour calculer la ressemblance entre le spectre moyen d'excitation du '*loading*' $\mathbf{s}_{ex}^{\mathrm{load-moy.}}$ et le spectre moyen d'excitation de la MEEF de référence $\mathbf{s}_{ex}^{\mathrm{réf-moy.}}$ par la relation[137] :

$$\mathrm{SAM}\left(\mathbf{s}_{ex}^{\text{load}-\text{moy.}}, \mathbf{s}_{ex}^{\text{réf}-\text{moy.}}\right) = \frac{\mathbf{s}_{ex}^{\text{load}-\text{moy.}} \cdot \mathbf{s}_{ex}^{\text{réf}-\text{moy.}}}{\left\|\mathbf{s}_{ex}^{\text{load}-\text{moy.}}\right\| \times \left\|\mathbf{s}_{ex}^{\text{réf}-\text{moy.}}\right\|} \quad \text{(B.5)}$$

Où l'opérateur · signifie le produit scalaire. Les termes $\left\|\mathbf{s}_{ex}^{\text{load}-\text{moy.}}\right\|$ et $\left\|\mathbf{s}_{ex}^{\text{réf}-\text{moy.}}\right\|$ signifient la norme euclidienne du vecteur $\mathbf{s}_{ex}^{\text{load}-\text{moy.}}$ et du vecteur $\mathbf{s}_{ex}^{\text{réf}-\text{moy.}}$, respectivement. La norme euclidienne du vecteur $\mathbf{s}_{ex}^{\text{load}-\text{moy.}}$ est donnée par la relation suivante[138] :

$$\left\|\mathbf{s}_{ex}^{\text{load}-\text{moy.}}\right\| = \sqrt[2]{\sum_{ex=1}^{88} (s_{ex}^{\text{load}-\text{moyen}})^2} \quad \text{(B.6)}$$

Où s_{ex}^{load} est la valeur du spectre d'excitation à la longueur d'onde ex.

B.2.4) Affichage des MEEF de référence

Maintenant que le Loading1 a été identifié comme étant correspondant à l'anthracène, nous pouvons visualiser graphiquement la MEEF de référence de cette espèce et la comparer à celle du '*loading*'. L'affichage en mode image, que nous avons choisi dans la Figure B.7 nous confirme bien que notre Loading1 correspond à l'anthracène, car les longueurs d'ondes d'émission et d'excitation sont identiques, entre 380 nm et 450 nm pour l'émission et entre 250 nm et 400 nm pour l'excitation. Nous notons également que malgré la différence d'ordre de grandeur en termes d'intensités (barre de couleur à droite), nous arrivons à bien identifier ce '*loading*' avec les calculs de l'étape précédente, car nous avons normalisé nos MEEF.

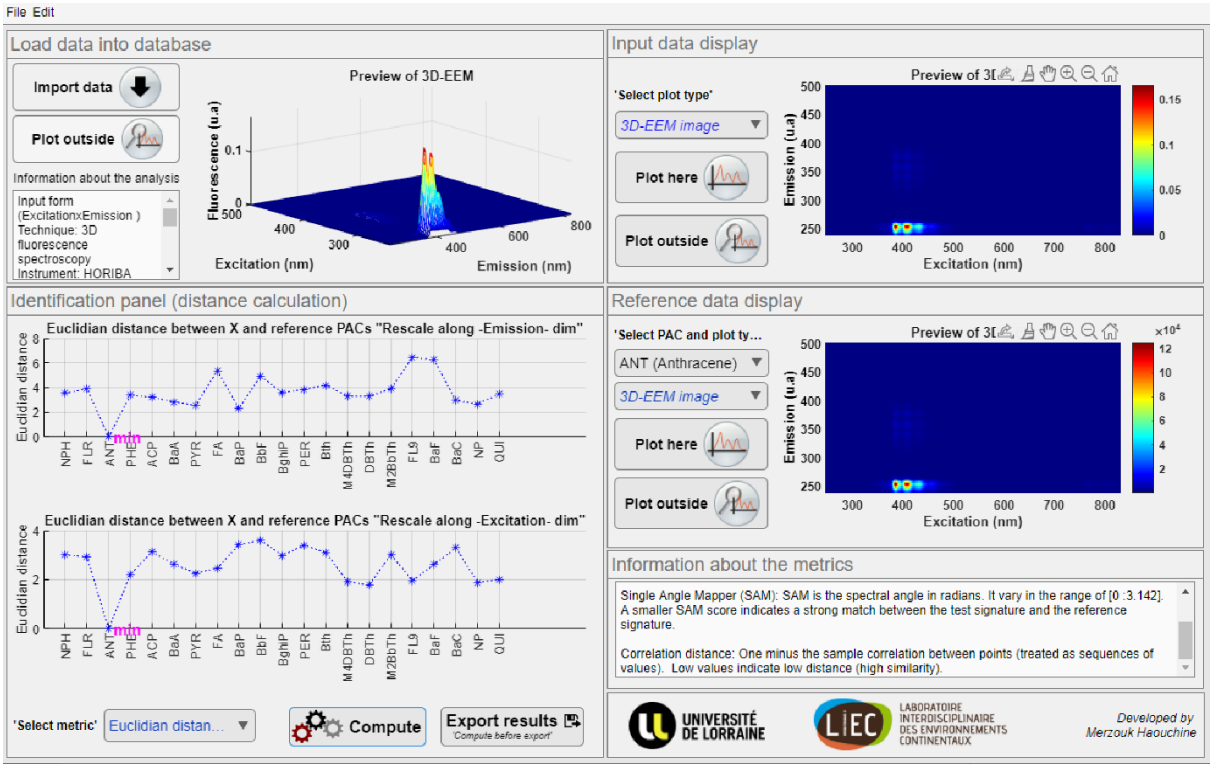


Figure B.7 : Résultats de l'identification du Loading1 grâce l'interface graphique utilisateur.

B.3) Base de données incluse dans la GUI

La base des données de références comprend pour le moment 25 CAP dont 16 HAP, 2 céto-HAP, 1 nitro-HAP, 2 S-CAP, 2 S-CAP alkylés et 2 N-CAP (Tableau B.2). Chaque MEEF a été acquise à sa concentration optimale (i.e. meilleur rapport signal/bruit) puis a été prétraitée par MT-SVD pour corriger les effets de diffusion de la lumière et du bruit (toutes les informations sont disponibles sur le panneau '***Information about the analysis***' (cf. **B.2.1) Importation de la MEEF du '*loading*' inconnu**).

Tableau B.2 : Base des données de références incluse dans l'interface graphique utilisateur.

Catégorie	Sous-catégorie	Nom de la molécule	Abréviation du nom
HAP et leurs dérivés	HAP	Anthracène	ANT
		Chrysène	CHR
		Phénanthrène	PHE
		Naphtalène	NPH
		Acénaphtène	AC
		Fluorène	FLR
		Dibenz(*a,h*)anthracène	DBahA
		Benz[α]anthracène	BaA
		Pyrène	PYR

		Benz[α]pyrène	BaP
		Pérylène	PER
		Benz[*k*]fluoranthène	BkF
		Benz(*g,h,i*)perylène	BghiP
		Fluoranthène	FA
		Benz[*b*]fluoranthène	BbF
		Indénopyrène	IP
	Céto-HAP	9H-Fluorén-9-one	9FL
		1H-Benzo[a]fluorén-1-one	BaF
	Nitro-HAP	1-Nitropyrène	NP
S-CAP (Thiazaarènes) et leurs dérivés	S-CAP	1-benzothiophene	BTh
		Dibenzothiophene	DBTh
	S-CAP alkylés	2-Méthyl-1-benzothiophène	2MBTh
		4-Méthyldibenzo[*b,d*]thiophène	4MDBTh
N-CAP (Azaarènes)	/	Quinoléine	QUI
		Benzo[α]carbazole	BaC

L'algorithme itératif '*Asymmetric Least Squares*', basé sur le filtre de Whittaker (WLS) est une méthode de correction de la ligne de base. Elle consiste à estimer une ligne de base avec un polynôme en l'ajustant par itération sur l'ensemble des spectres. Une fois la ligne de base est ajustée, elle est utilisée pour corriger chaque spectre. Cette ligne de base notée l est ajustée en minimisant la fonction[103] :

$$S = \sum_i p(x_i - l_i)^2 + \lambda \sum_i (\Delta^2 l_i)^2 \qquad \text{(C.1)}$$

Avec $\Delta^2 l_i = l_i - 2l_{i-1} + l_{i-2}$. Ce terme traduit l'ajustement de la ligne de base sur le spectre brut x_i de l'échantillon i. Le paramètre $0 < p < 1$ correspond au degré d'asymétrie requis pour l'ajustement. Des valeurs plus grandes permettent davantage de régions à pente négative. Des valeurs plus petites n'autorisent pas les régions à pente négative. Le paramètre λ est un paramètre similaire à l'ordre du polynôme car il contrôle le degré de courbure autorisée pour la ligne de base. Plus la valeur de λ est petite, plus il y a de courbure autorisée dans l'ajustement de la ligne de base. Ces deux paramètres sont à optimiser lors de la correction[76]. Dans notre cas, p est fixé à 0.001 et λ à 1000.

Un signal temporel continu est une représentation mathématique d'un phénomène physique continu qui varie en fonction du temps. Il peut s'agir d'un signal électrique, sonore, optique ou de tout autre type de signal continu qui évolue au fil du temps. Graphiquement, un signal temporel peut être visualisé comme une courbe qui évolue en fonction du temps (le temps en abscisse et l'intensité de la grandeur physique en ordonnée). Un signal temporel continu peut ensuite être discrétisé (i.e. échantillonné) à des instants de temps spécifiques, et à chacun de ces morceaux est associé un nombre. Ce signal continu devient donc un signal discret. Cette discrétisation est souvent nécessaire pour la manipulation et le traitement numérique du signal. Un spectre électromagnétique (e.g. spectre Raman) est un signal qui est mesuré et par conséquent, un signal discret. Cela signifie que l'on mesure une série d'intensités représentant des variables spectrales avec un pas défini par la résolution spectrale de l'appareil de mesure. Cependant, ces variables spectrales sont généralement fortement corrélées les unes aux autres, ce qui signifie qu'elles sont étroitement liées ou dépendantes les unes, des autres. Cette corrélation permet de considérer ce signal, initialement discret, comme un signal continu. Donc, par analogie à un signal temporel continu, on considère en spectroscopie, le signal spectral comme une courbe qui évolue en fonction des variables spectrales (par analogie ici au temps) et dont les ordonnées représentent les intensités correspondantes[75]. À partir de là, des transformations mathématiques développées dans le cadre du traitement des signaux temporels continus, permettant la transition du domaine temporel vers le domaine fréquentiel, peuvent être envisagées pour passer d'un signal spectral à un signal fréquentiel. Ces transformations permettent d'analyser les différentes fréquences présentes dans le signal et d'obtenir des informations sur les composantes fréquentielles du phénomène étudié[75].

En chimiométrie, ce type de transition permet, par exemple, une meilleure discrimination entre des groupes d'individus[139], une compression des données[130], l'étude de la nature du bruit contenu dans les données[140] ou encore le débruitage des signaux[117]. La plus populaire de ces méthodes est sans doute la transformée de Fourier, qui offre une résolution fréquentielle maximale. Cependant, cette dernière ne permet pas d'établir une correspondance directe entre l'information fréquentielle et l'information temporelle (par analogie, l'information spectrale) car elle est caractérisée par une résolution temporelle (par analogie, résolution spectrale) nulle (Figure D.1).

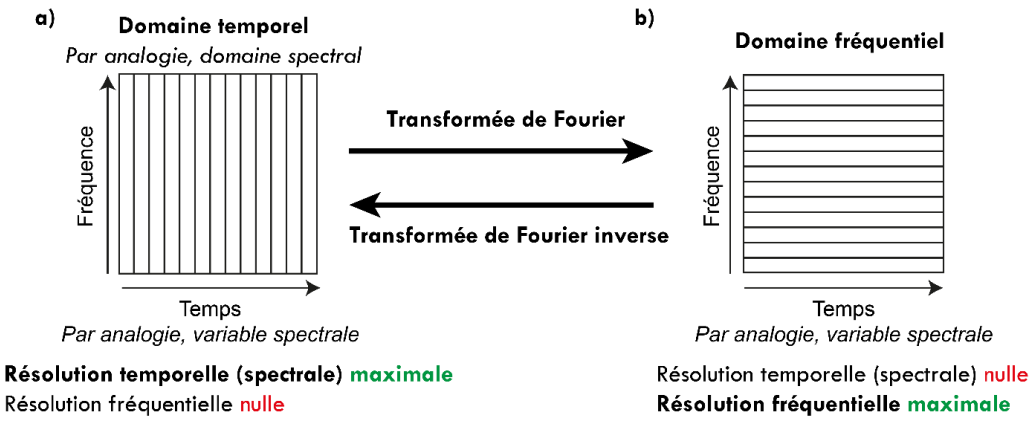


Figure D.1 : Illustration de la dualité entre la résolution temporelle (spectrale) et la résolution fréquentielle dans le domaine temporel (a) et le domaine fréquentiel (b).

Pour mieux illustrer ce point, prenons le cas simple d'un signal '*chirp'* simulé qui évolue dans le temps depuis les basses fréquences (5 Hz) vers les hautes fréquences (50 Hz) avec une amplitude entre -1 et 1 u.a, et le cas d'un autre '*chirp*' qui évolue inversement dans le temps (i.e. de 50 Hz à 5 Hz) (Figure D.2a). La transformée de Fourier de ces deux signaux nous permet d'identifier leur plage fréquentielle, qui est entre 5 Hz et 50 Hz. Cependant, elle ne nous permet pas de déduire une évolution temporelle inverse entre les deux cas, car les résultats des deux transformations sont identiques (Figure D.2b). N.B. Pour faciliter la lecture, dans la suite du texte, nous allons parler uniquement de domaine spectral en admettant l'analogie avec le domaine temporel.

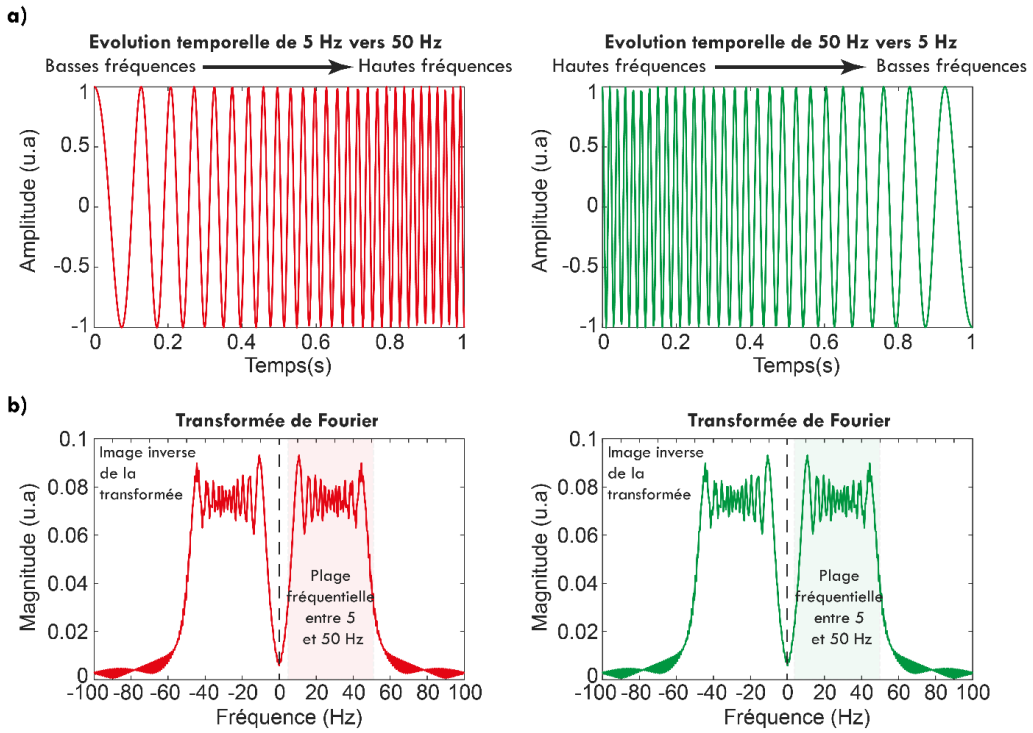


Figure D.2 : Deux signaux 'chirp' qui évolue inversement dans le temps. Depuis les basses fréquences vers les hautes fréquences pour celui de gauche et depuis les hautes fréquences vers les basses fréquences pour celui de gauche (a). Résultats de la transformée de Fourier des deux signaux 'chirp' (b).

Lorsque nous voulons par exemple, déterminer l'implication d'une bande vibrationnelle dans la discrimination observée dans le domaine fréquentiel entre deux individus, cela est irréalisable avec la transformée de Fourier. Ainsi, d'autres transformations mathématiques ont été développées pour tenter de relier les fréquences observées à leurs domaines spectraux correspondants. L'une d'entre elles est la transformée de Fourier fenêtrée[141], qui consiste en un échantillonnage préalable du signal spectral par une fenêtre d'apodisation de taille déterminée, puis la réalisation de la transformée de Fourier sur chaque partie du signal délimitée par la fenêtre. Cette méthode offre une première solution pour localiser dans le domaine spectral l'information fréquentielle. Cependant, elle implique intrinsèquement que lorsque la résolution fréquentielle augmente, celle du domaine spectral diminue, et vice versa (fait appelé principe d'incertitude). Or, dans le cas de signaux complexes tels que les signaux spectroscopiques, l'évolution du signal est un mélange de signaux de basses fréquences (i.e. bandes vibrationnelles larges) et de hautes fréquences (i.e. bandes vibrationnelles fines). Les signaux de basses fréquences nécessitent une résolution fréquentielle élevée pour être bien caractérisés, mais ne nécessitent pas une résolution spectrale élevée, car ils évoluent lentement en fonction des variables spectrales. En revanche, les signaux de hautes fréquences ne nécessitent pas une résolution fréquentielle élevée, mais nécessitent une résolution spectrale élevée, car ils évoluent très rapidement avec les variables spectrales. Ainsi, par exemple, la caractérisation d'une bande vibrationnelle fine peut être ratée si on décide de fixer une résolution fréquentielle élevée et donc une résolution spectrale faible. À l'inverse, la caractérisation d'une bande vibrationnelle large peut être ratée si on décide de fixer une résolution spectrale élevée et donc une résolution fréquentielle basse (Figure D.3a). C'est pourquoi la transformée en ondelettes continue a été développée. Elle est bien adaptée aux signaux complexes tels que les signaux spectroscopiques, car elle offre un bon compromis entre la résolution spectrale et la résolution fréquentielle. En effet, lorsque le signal spectral évolue rapidement (i.e. bandes vibrationnelles fines), la résolution spectrale est augmentée. La résolution fréquentielle est quant à elle abaissée, car les bandes vibrationnelles fines se caractérisent par des hautes fréquences qui ne nécessitent pas une résolution fréquentielle élevée pour être caractérisées. En revanche, lorsque le signal spectral évolue lentement (i.e. bandes vibrationnelles larges), la résolution spectrale est réduite tandis que la résolution fréquentielle est augmentée, car ces signaux se caractérisent par des basses fréquences qui nécessitent une bonne résolution fréquentielle pour être bien caractérisées (Figure D.3b).

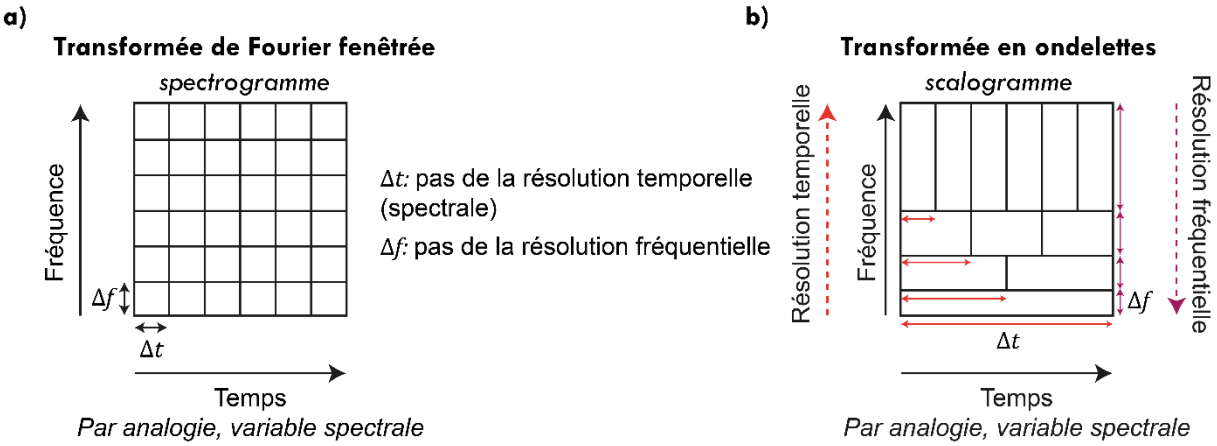


Figure D.3 : Illustration de la dualité entre la résolution temporelle (spectrale) et la résolution fréquentielle. Transformée de Fourier fenêtrée (a). Le résultat d'une transformée de Fourier fenêtrée est appelé spectrogramme. Transformée en ondelettes (b). Le résultat d'une transformée en ondelettes est appelé scalogramme.

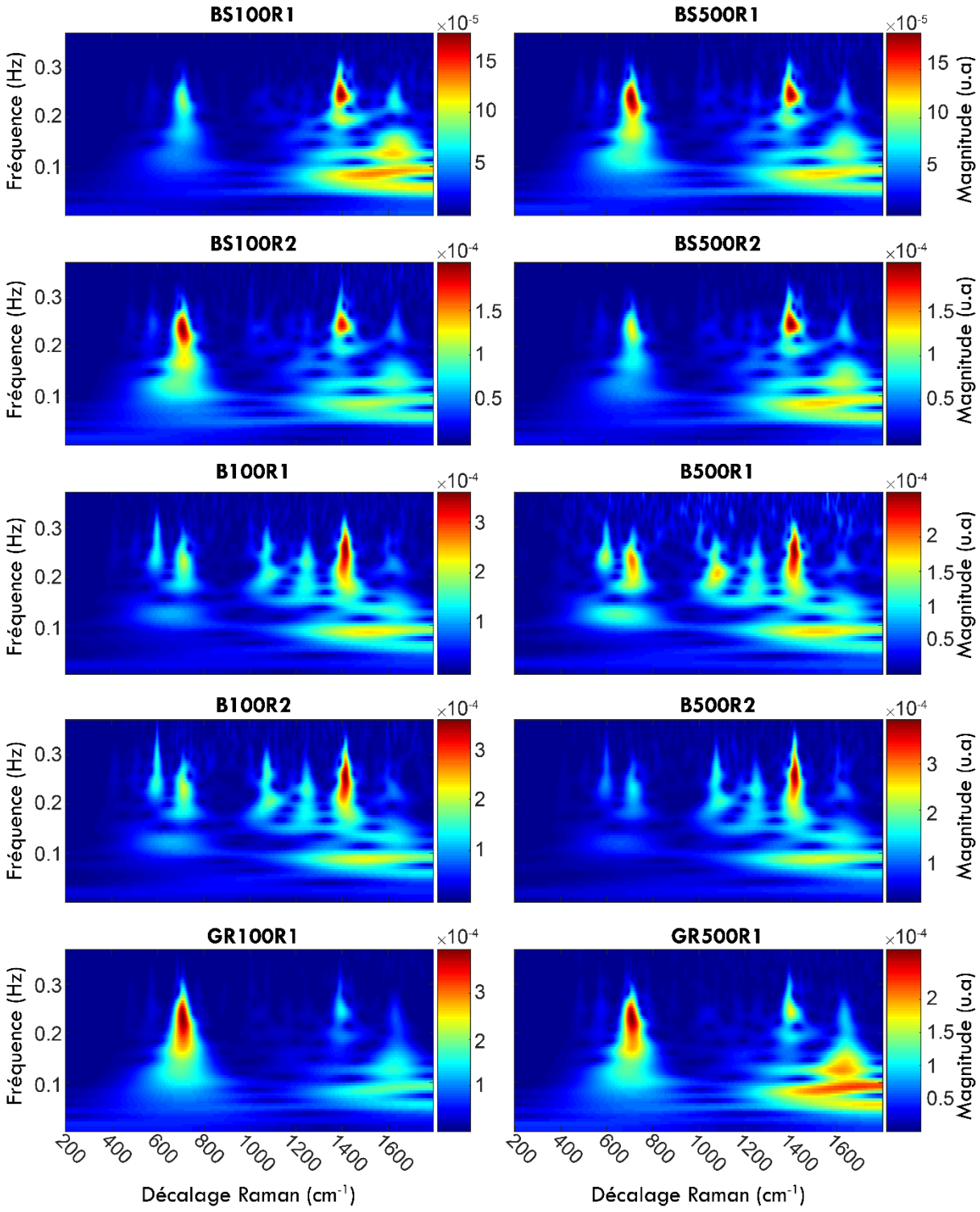


Figure E.1 : Scalogrammes des extraits organiques des sols BS, B et GR

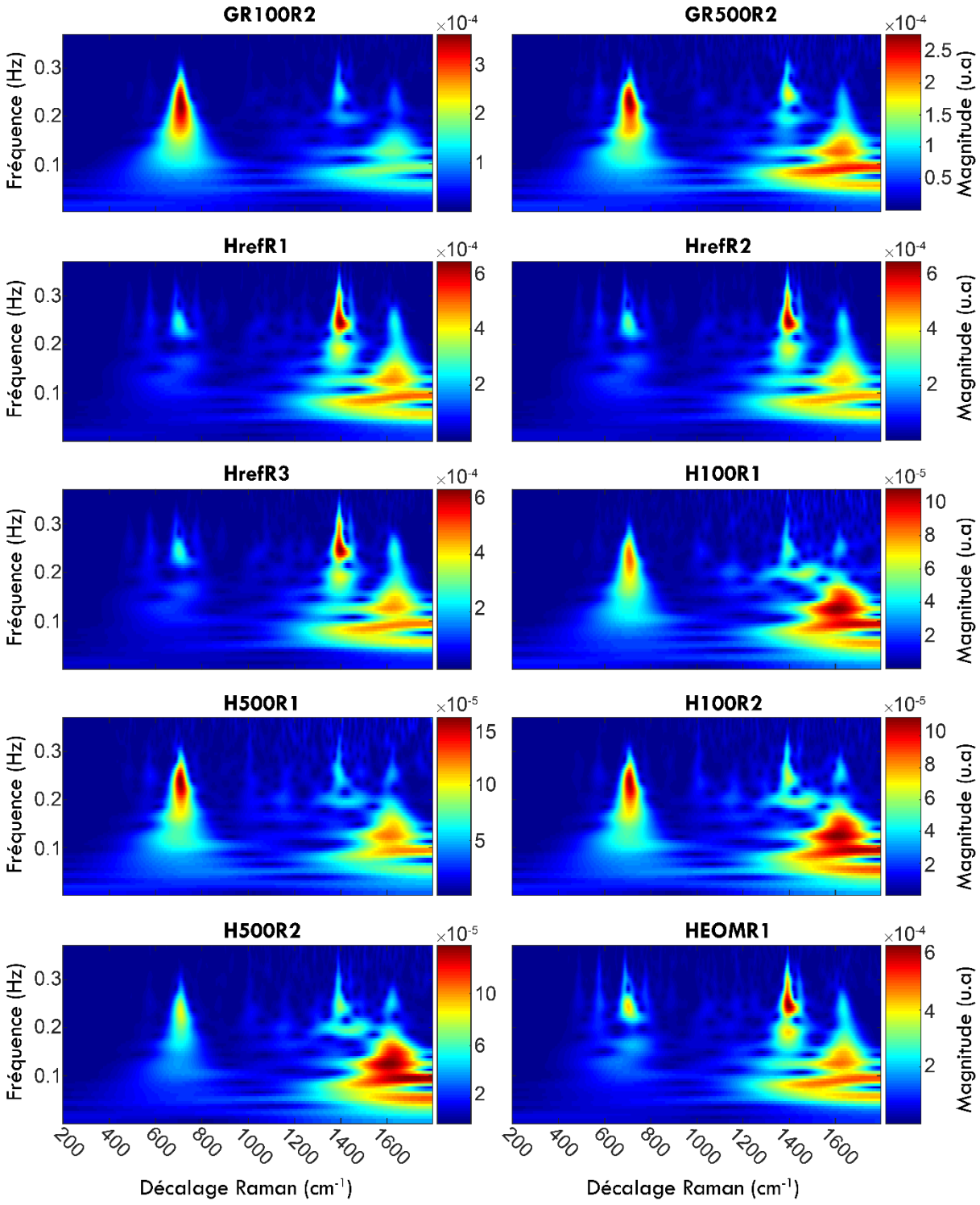


Figure E.2 : Scalogrammes des extraits organiques des sols GR, Href, H et HEOM.

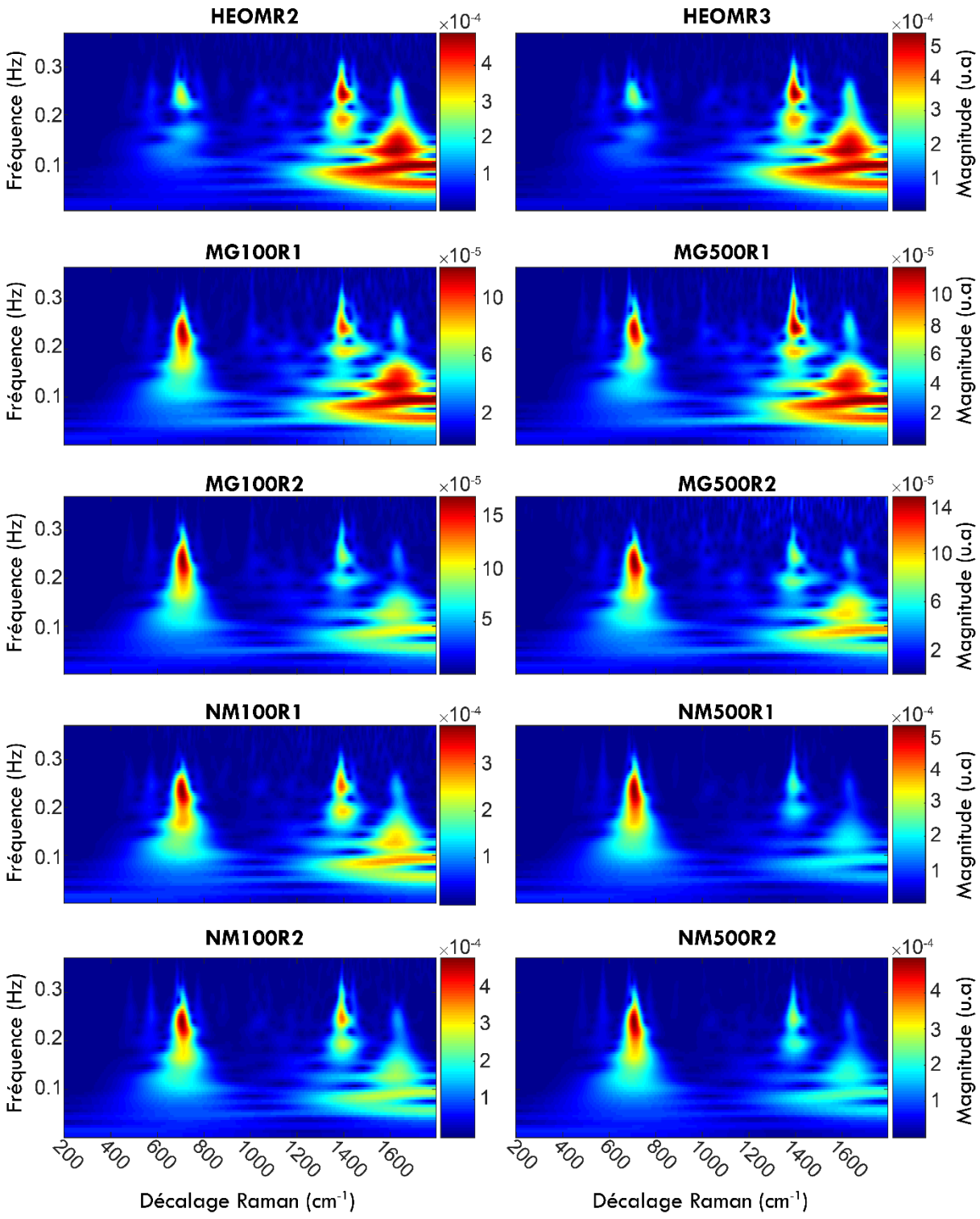


Figure E.3 : Scalogrammes des extraits organiques des sols HEOM, MG et NM.

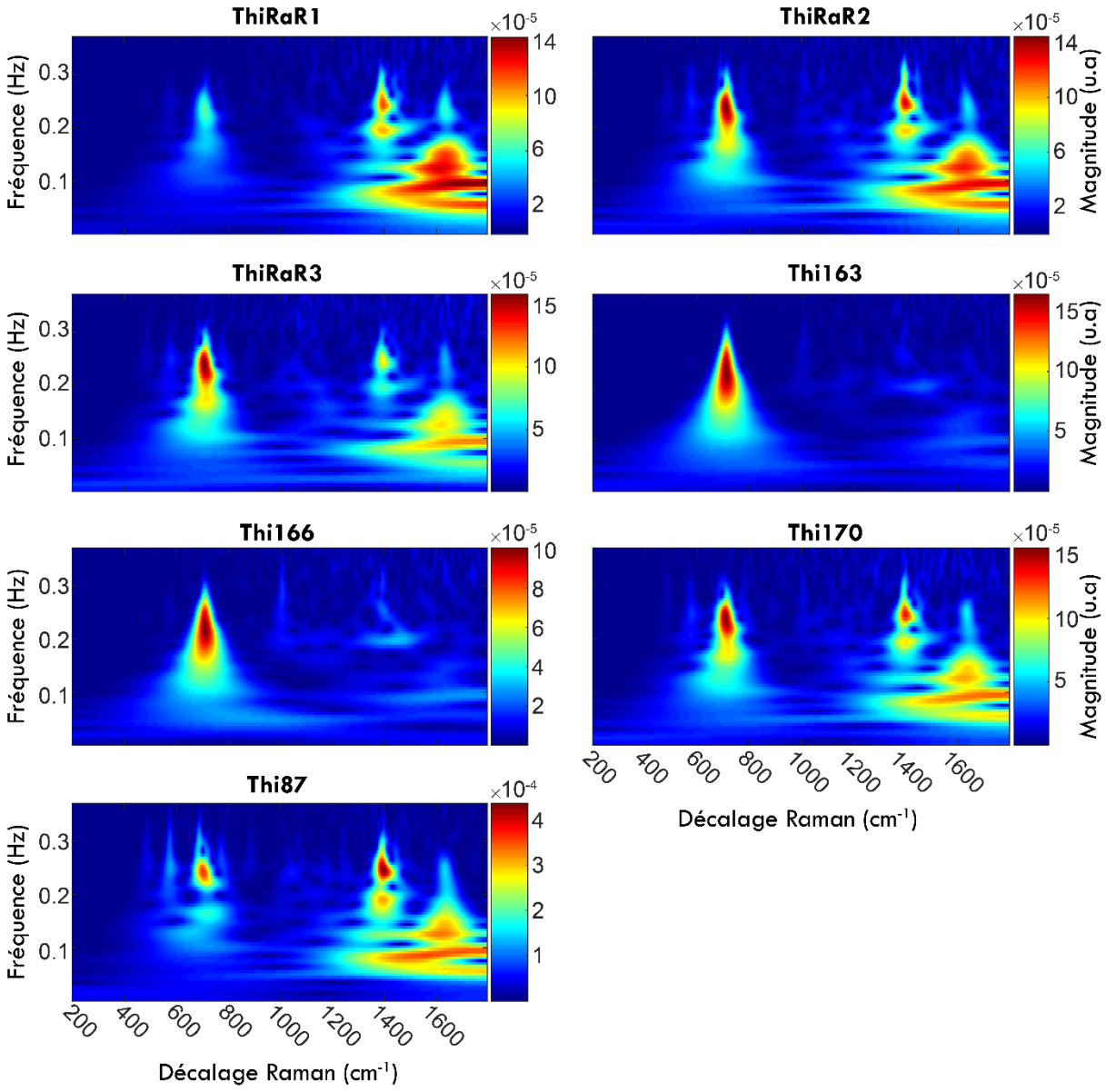


Figure E.4 : Scalogrammes des extraits organiques des sols ThiRa et Thi.

Communications scientifiques

Publications

1. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. Handle Matrix Rank Deficiency, Noise, and Interferences in 3D Emission–Excitation Matrices: Effective Truncated Singular-Value Decomposition in Chemometrics Applied to the Analysis of Polycyclic Aromatic Compounds. ACS Omega 2022 7 (27), 23653-23661, https://doi.org/10.1021/acsomega.2c02256.
2. M. Marchetti, G. Gouadec, M. Offroy, **M. Haouchine,** A. Djerbi, O. Omikrine-Metalssi, J-M. Torrenti, J-M. Mechling, G. Simon, P. Turcry; P. Barthélémy, O. Amiri. Raman identification of CaCO3 polymorphs in concrete prepared with carbonated recycled. Materials and Structures, Springer. *In peer review.*

Posters

1. H. Huot, **M. Haouchine**, C. Caillet, M. Offroy. Exploratory study of infrared spectral signatures of a range of forest, agricultural and anthropized soils from the North-East of France. International conference of the IUSS Working Group on Soils of Urban, Industrial, Traffic, Mining and Military Areas, Santiago de Compostela (Spain) on 4th-7th September 2023.
2. H. Huot, **M. Haouchine**, C. Caillet, M. Offroy. Exploratory study of infrared spectral signatures of a range of forest, agricultural and anthropized soils from the North-East of France. 29th GFSV scientific days, Deauville (France), from May 25 to May 26, 2023.
3. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. Gérer la déficience du rang matriciel, le bruit et les interférences dans les matrices d'émission-excitation de fluorescence grâce à la chimiométrie : un nouvel algorithme appliqué à l'analyse des composés aromatiques polycycliques. Congrès '*French Researchers in organic geochemistry*', FROG V, Rennes (France), du 04 au 06 juillet 2022.
4. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. A new chemometrics preprocessing (MT-SVD) based on effective information truncation to handle matrix rank deficiencies as well as the effects of noise and light scattering in 3D excitation emission fluorescence matrices. Congrès du groupe français de chimiométrie, Brest (France), 7 juin 2022.
5. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. Handle matrix rank deficiency, noise and interferences in 3D emission-excitation matrices using chemometrics: A new algorithm applied to the analysis of polycyclic aromatic hydrocarbons. 28th GFSV scientific days, Nivelles (Belgium), from May 18 to May 20, 2022.
6. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. Outils de chimiométrie pour le suivi de contamination aux composés aromatiques polycycliques. Séminaire annuelle de l'école doctorale SIRENA, Nancy (France), 26 mai 2021.

Communications orales

1. M. Marchetti, M. Offroy, **M. Haouchine,** G. Simon. Raman identification of CaCO3 polymorphs in concrete prepared with carbonated recycled. 29th GFSV scientific days, Deauville (France), from May 25 to May 26, 2023.
2. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. Multi-Truncated Singular Value Decomposition (MT-SVD) : "A New Chemometrics Algorithm to Preprocess 3D-Excitation-Emission Fluorescence Matrices : Application to the Analysis of Polycyclic

Aromatic Compounds (PACs)", journée annuelle des doctorants du LIEC, Metz (France), 17 juin 2022.

3. **M. Haouchine**, C. Biache, C. Lorgeoux, P. Faure, M. Offroy. Gérer la déficience du rang matriciel, le bruit et les interférences dans les matrices d'émission-excitation 3D de fluorescence grâce à la chimiométrie: nouvel algorithme appliqué à l'analyse des hydrocarbures aromatiques polycycliques (CAP). E-Séminaire école doctorale SIRENA, Nancy (France), 08 mars 2022.

Résumé-Abstract

Développement d'outils de chimiométrie pour le suivi de contamination aux composés aromatiques polycycliques dans des matrices environnementales complexes

Résumé : Les composés aromatiques polycycliques (CAP), incluant les hydrocarbures aromatiques polycycliques (HAP), sont des milliers de molécules organiques se caractérisant par une grande diversité chimique et une origine à la fois naturelle mais surtout anthropique. Ils suscitent des préoccupations en raison de leur toxicité et de leur persistance dans l'environnement, notamment dans les sols. Leur caractérisation revêt donc une importance cruciale tant sur le plan environnemental que pour la santé publique. Elle permet en effet une meilleure compréhension de l'évolution de cette pollution et, par conséquent, une gestion améliorée des sites contaminés. Des efforts de recherche sont actuellement fournis pour atteindre cet objectif, en utilisant diverses techniques analytiques avancées, notamment les systèmes chromatographiques couplés à divers détecteurs. Néanmoins, la complexité de ces approches analytiques de référence, le nombre et la variabilité chimique des CAP, ainsi que la complexité des matrices environnementales, limitent la caractérisation qualitative et quantitative exhaustive de ces composés. Dans cette perspective, les techniques spectroscopiques, exploitant l'interaction de la lumière avec la matière, apparaissent comme prometteuses et complémentaires en raison de leur facilité de mise en œuvre et de leur rapidité. Cependant, les quantités importantes de données qu'elles génèrent sont complexes, car ce sont des signatures globales des échantillons analysés. De plus, elles sont souvent perturbées par des bruits instrumentaux et des signaux interférents issus de phénomènes physiques concurrents. C'est là que la chimiométrie intervient, offrant des outils de traitement mathématique et statistique avancés, essentiels pour une exploitation pertinente des données spectroscopiques. Nos travaux de recherche proposent dans ce contexte des approches novatrices de chimiométrie appliquées aux données de spectroscopie de fluorescence et Raman dans le but d'améliorer les étapes de diagnostic des sols contaminés par les CAP. Nous présenterons en premier lieu un nouvel algorithme de prétraitement des matrices d'excitation-émission de fluorescence 3D (MEEF), visant à réduire le bruit tout en identifiant les informations chimiques pertinentes qu'elles renferment. Nous exposerons ensuite une méthodologie originale de décomposition PARAFAC de MEEF d'extraits organiques de sols industriels contaminés, permettant une caractérisation qualitative de 16 HAP cibles. Puis, nous associerons les MEEF à des données quantitatives de chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS) pour caractériser de manière quantitative la pollution par les 16 HAP à l'aide d'algorithmes de régression ('*support vector regression*' SVR). Enfin, nous discuterons la perspective d'élargir la caractérisation à d'autres CAP (e.g. dérivés de HAP) grâce à une analyse multiblocs des différents ensembles de données provenant de diverses techniques analytiques, notamment la spectroscopie Raman que nous démontrerons comme complémentaire à la spectroscopie de fluorescence, ainsi qu'à la GC-MS.

Mots-clés : Chimiométrie, CAP, HAP, spectroscopie, fluorescence, Raman, MEEF, MT-SVD, PARAFAC, GC-MS, algorithmes de régression, SVR.

Résumé-Abstract

Development of Chemometrics tools for monitoring contamination by polycyclic aromatic compounds in complex environmental matrices.

Abstract: Polycyclic aromatic compounds (PACs), including polycyclic aromatic hydrocarbons (PAHs), represent thousands of organic molecules characterized by significant chemical diversity and an origin that is both natural and predominantly anthropogenic. They raise concerns due to their toxicity and persistence in the environment, notably in soils. Therefore, their comprehensive characterization holds crucial importance, both environmentally and for public health. Indeed, it enables a better understanding of the evolution of this pollution and, consequently, an improved management of contaminated sites. Current research efforts are dedicated to achieving this goal through the use of various advanced analytical techniques, especially chromatographic systems coupled with diverse detectors. However, the complexity of these reference analytical approaches, the number and chemical variability of PACs, and the intricacies of environmental matrices limit the exhaustive qualitative and quantitative characterization of these compounds. In this context, spectroscopic techniques, exploiting the interaction of light with matter, emerge as promising and complementary, owing to their ease of implementation and speed. Nevertheless, the substantial quantities of data they generate are complex, as they represent global signatures of analyzed samples. Moreover, these data are frequently disrupted by instrumental noise and interfering signals stemming from concurrent physical phenomena. This is where chemometrics intervenes, offering advanced mathematical and statistical tools essential for the relevant exploitation of spectroscopic data. In this context, our research proposes innovative chemometric approaches applied to fluorescence and Raman spectroscopy data to enhance the diagnostic stages of soils contaminated by PACs. We will firstly introduce a novel preprocessing algorithm for 3D excitation-emission fluorescence matrices (3D EEMs), aimed at reducing noise while identifying pertinent chemical information within them. Subsequently, we will present an original PARAFAC decomposition methodology of EEMs from organic extracts of contaminated industrial soils, enabling the qualitative characterization of 16 target PAHs. Furthermore, we will combine EEMs with quantitative gas chromatography-mass spectrometry (GC-MS) data to quantitatively characterize the pollution by the 16 target PAHs using regression algorithms (‘*support vector regression*’ SVR). Lastly, we will discuss the prospect of extending the characterization to other PACs (e.g., PAH derivatives) through a multiblock analysis of datasets originating from various analytical techniques, including Raman spectroscopy, which we will demonstrate as complementary to fluorescence spectroscopy, as well as GC-MS.

Keywords: Chemometrics, PACs, PAHs, spectroscopy, fluorescence, Raman, 3D EEMs, MT-SVD, PARAFAC, GC-MS, regression algorithms, SVR.