Thèse de Doctorat

Université de Limoges

Université de Limoges
ED 653 : Sciences et Ingénierie
Faculté des Sciences et Techniques – Institut de Recherche XLIM

Thèse pour obtenir le grade de
Docteur de l'Université de Limoges
Sciences et technologies de l'information et de la communication

Présentée et soutenue par
Damien Boildieu
Le 15 décembre 2022

MÉTHODES DE RÉSOLUTION DE COURBES MULTIVARIÉES POUR LA MICROSPECTROSCOPIE CARS

Thèse dirigée par Philippe CARRÉ et Philippe LEPROUX

JURY :

Président du jury
M. Frédéric MORAIN-NICOLIER, Professeur – CReSTIC – Université de Reims Champagne-Ardenne

Rapporteurs
M. Olivier LALIGANT, Professeur – ImViA – Université de Bourgogne Franche-Comté
M. Olivier PIOT, Professeur – BioSpecT – Université de Reims Champagne-Ardenne

Examinateurs
Mme. Amandine MAGNAUDEIX, Maître de Conférences – IRCER – Université de Limoges
M. David HELBERT, Professeur – XLIM – Université de Poitiers
M. Philippe LEPROUX, Maitre de Conférences – XLIM – Université de Limoges
M. Philippe CARRÉ, Professeur – XLIM – Université de Poitiers

Remerciements

Je tiens tout d'abord à remercier aux membres du jury de cette thèse. Aux deux rapporteurs, Monsieur Olivier Laligant, professeur à l'université de Bourgogne Franche-Comté et Monsieur Olivier Piot, professeur à l'université Reims Champagne-Ardenne ainsi qu'au président du jury Monsieur Frédéric Morain-Nicolier, professeur à l'université Reims Champagne-Ardenne d'avoir accepté d'évaluer mes travaux et pour leurs retours constructifs.

Je tiens à remercier les différents membres de mon encadrement durant ces trois années de doctorat. Monsieur Philippe Carré pour m'avoir présenté, accepté pour ce sujet à la suite de mon Master en conception logicielle et la co-direction de cette thèse. Monsieur Philippe Leproux pour le temps passé à m'enseigner le fonctionnement du phénomène Raman et la spectroscopie CARS, le temps passé à l'étude des résultats ainsi que pour la co-direction de la thèse. Madame Amandine Magnaudeix pour les cours de biologie sur les cellules et les tissus, la création de la base de données de segmentation et le temps passé à étudier les résultats. Monsieur David Helbert pour son suivi et les conseils m'ayant aidé à faire la transition de l'informatique au traitement du signal.

Je remercie aussi mes collègues doctorants à XLIM et à l'ADIIS pour les moments passés ensemble, Alexandre Fenneteau, Clément Joubert, Thibault Lacharme, Adrien Raison, Gean Trindade et tous les autres. Sans vous ce doctorat aurait été bien moins plaisant.

Je remercie mes parents de m'avoir soutenu durant toutes ces années d'étude et d'avoir fait le déplacement pour ma soutenance.

Un merci tout particulier à Elsa Tamisier qui m'a supporté durant ce doctorat et gardé motivé dans les moments les plus durs, bon courage pour le tien.

Table des Matières

Liste des Figures

Liste des Tableaux

Liste des Algorithmes

Liste des Abbréviations

ACP – analyse en composantes principales.
ADN – acide desoxyribonucléique.
AE – auto-encodeur.
AEV – auto-encodeurs variationnels.
ARN – acide ribonucléique.
ARNm – acide ribonucléique messager.
ARNr – acide ribonucléique ribosomique.
ATP – Adénosine triphosphate.

BDNF – *brain-derived neurotrophic factor*.

CARS – *coherent anti-Stokes Raman scattering*.
CSV – Chan-Sandberg-Vese.
CV – Chan-Vese.

EQM – erreur quadratique moyenne.

FCLS – *fully constrained least squares*.

HSI – *hyperspectral imaging*.

IFD – immunofluorescence directe.
IFI – immunofluorescence indirecte.

LOF – *lack of fit*.

MCR – *multivariate curve resolution*.
MCR-ALS – *multivariate curve resolution - alternating least squares*.
MEC – matrice extracellulaire.
MEM – méthode de l'entropie maximale.

NNLS – *non-negative least squares*.

PCF – *photonic crystal fiber*.

RE – réticulum endoplasmique.
ReLU – *rectified linear unit*.

SAD – *spectral angle distance*.
SIMPLISMA – *simple-to-use interactive self-modelling mixture analysis*.
SLIC – *simple linear iterative clustering*.
SNR – *signal-to-noise ratio*.
SVD – *singular value decomposition*.

TDKK – *time domain Kramers-Kronig*.
TrkB – *Tropomyosin receptor kinase B*.

VCA – *vertex component analysis*.

Liste des Symboles

A_F – Premier coefficient de la méthode de l'entropie maximale.
A – Amplitude du mode vibrationnel.
B – Taille d'un lot.
C – Matrice des concentrations.
D – Matrice des données spectrales.
E_p – Champ électrique de l'onde pompe.
E_s – Champ électrique de l'onde Stokes.
E_{pr} – Champ électrique de l'onde sonde.
E – Champ électrique.
FN – Faux négatif.
FP – Faux positif.
F – Rang du système polynomial de la méthode de l'entropie maximale.
I_{CARS} – Intensité CARS.
I_{Raman} – Intensité Raman.
I – Matrice identité.
J – La matrice jacobienne.
K – Nombre de composants dans D.
N – Nombre de canaux de spectraux.
PV – Valeur principale de Cauchy.
$P^{(1)}$ – Polarisation d'ordre 1.
$P^{(3)}$ – Polarisation d'ordre 3.
P – Nombre de classes pour une classification ou segmentation.
R – Matrice de rotation.
S – Matrice des spectres.
U – Taille d'un filtre convolutif.
VP – Vrai positif.

V – Vecteurs propres.
Y – Données projetées.
Z – Ensemble des données encodées.
Γ – Largeur de bande du mode vibrationnel.
Λ – Valeurs propres.
Ω – Fréquence du mode vibrationnel.
Φ – Fonction de reconstruction de D à partir de C et S.
Σ – Matrice de variance-covariance.
ω_{as} – Fréquence de l'onde anti-Stokes.
$\bar{D}$ – Matrice D « augmentée ».
$\bar{S}$ – Matrice S « augmentée ».
$\bar{d}$ – Vecteur des moyennes de chaque colonne de D.
β – Deuxième coefficient de la méthode de l'entropie maximale.
$\breve{D}$ – D normalisé dans la méthode SIMPLISMA.
$\cdot$ – Opérateur de produit scalaire.
$\chi^{(1)}$ – Susceptibilité électrique d'ordre 1.
$\chi^{(3)}_R$ – Partie résonnante de $\chi^{(3)}$.
$\chi^{(3)}_{NR}$ – Partie non-résonnante de $\chi^{(3)}$.
$\chi^{(3)}$ – Susceptibilité électrique d'ordre 3.
ϵ – Paramètre de mise à l'échelle du modèle CNNAEU.
γ – Facteur de variabilité dans la méthode VCA.
$\hat{x}$ – Données décodées.
ι – Phase estimée par la méthode de l'entropie maximale.
λ – Paramètre de pondération d'une contrainte de la fonction de coût du modèle EndNet.
$\mathbb{B}$ – Ensemble composé de 0 et 1.
$\mathcal{A}$ – Matrice d'adjacence d'un graphe $\mathfrak{G}$.
$\mathcal{C}$ – Matrice des données doublement centrées dans l'algorithme isomap.
$\mathcal{D}$ – Un décodeur.
$\mathcal{E}$ – Un encodeur.
$\mathcal{E}$ – Matrice d'erreur.
$\mathcal{F}$ – Transformée de Fourier.
$\mathcal{I}$ – Image pouvant avoir un nombre de canaux quelconque. Utilisé aussi pour D sous la forme d'image avec $H \times W = M$.
$\mathcal{L}_{CE}$ – Entropie croisée.
$\mathcal{L}$ – Fonction de coût à optimiser.

$\mathcal{N}$ – Un bruit gaussien.
$\mathcal{O}$ – Matrice de corrélation autour de l'origine.
$\mathcal{P}$ – Métrique de précision.
$\mathcal{R}$ – Métrique de rappel.
$\mathcal{V}$ – Nombre de voisins utilisé pour le mélange de pixels.
κ – Pondération aléatoire utilisé pour le mélange de pixels.
$\mathfrak{C}$ – Courbe délimitant la segmentation dans la méthode de Chan-Vese.
$\mathfrak{D}$ – Dice score.
$\mathfrak{E}$ – Arêtes d'un graphe $\mathfrak{G}$.
$\mathfrak{G}$ – Un graphe quelconque.
$\mathfrak{N}$ – Fonction retournant le voisinage d'un pixel.
$\mathfrak{P}$ – Matrice de poids diagonale de l'analyse en composantes principales.
$\mathfrak{V}$ – Nœuds d'un graphe $\mathfrak{G}$.
$\mathfrak{W}^k$ – Poids utilisés pour trouver le k-ième spectre dans la méthode SIMPLISMA.
$\mathfrak{c}$ – Intensité moyenne au sein d'une classe de la segmentation de la méthode de Chan-Vese.
μ – Moyenne.
ν – Décalage Raman normalisé entre 0 et 1.
ω – Fréquence angulaire d'une onde.
ϕ – Erreur de phase dans la méthode de l'entropie maximale.
ω_p – Fréquence de l'onde pompe.
ψ – Opérateur défini par Liu *et al.* [1].
ρ – La probabilité qu'un neurone soit désactivé à l'apprentissage.
σ – Écart-type.
ω_s – Fréquence de l'onde Stokes.
$\star$ – Opérateur de corrélation croisée.
$\mathbf{SNR_{th}}$ – Seuil de ratio signal sur bruit pour la méthode de débruitage dans la méthode VCA.
θ – Phase d'un nombre complexe.
$\tilde{x}$ – Données bruitées.
υ – Paramètre de régularisation sur la longueur de la courbe de segmentation dans la méthode Chan-Vese.
$\varkappa$ – Paramètre de la méthode SIMPLISMA.
ϖ – Paramètre de régularisation sur l'aire de la courbe de segmentation dans la méthode Chan-Vese.

ϱ – Paramètre de pondération de l'homogénéité de l'intensité d'une classe dans la méthode de Chan-Vese.
ζ – Hyperparamètre de la contrainte de normalisation.
b – Biais d'un neurone artificiel.
d – Un spectre de D.
f – Une fonction dans le domaine temporel.
l – Facteur de l'ambiguïté d'intensité.
l – Point d'inflection de la sigmoïde.
p – Indice de « pureté » dans la méthode SIMPLISMA.
q – Coefficient d'inclinaison de la sigmoïde.
r_i^* – Conjugué du i-ème coefficient de la matrice de Toeplitz.
r_i – i-ème coefficient de la matrice de Toeplitz.
u – Fonction de Heaviside.
v – Un vecteur *one-hot*.
w – Paramètres d'un neurone artificiel.
x – Un vecteur quelconque.
y – Un second vecteur quelconque.
z – Données encodées.

Introduction générale

Dans son travail, le biologiste est régulièrement amené à visualiser les organismes qu'il étudie. Ces observations peuvent avoir de nombreuses raisons : étude de sanité d'une culture, compréhension de la physiologie des organismes ou encore étude d'éléments constitutifs des cellules, appelés constituants. Ces opérations sont principalement faites à l'aide de systèmes d'imagerie optique par fluorescence. Pour cela, des molécules intrinsèquement fluorescentes sont utilisées soit directement lorsqu'elles ont une affinité pour un constituant cellulaire précis (ex. : DAPI, BODIPY), soit couplées à un anticorps qui va avoir cette propriété quand la molécule fluorescente ne l'a pas et qu'il n'existe pas de molécule fluorescente ayant directement cette propriété. Une fois marquée, il est possible d'acquérir l'onde de fluorescence émise pour visualiser la molécule. Cette méthode possède cependant des limites. Tout d'abord, il faut trouver un marqueur compatible avec l'organisme étudié, ensuite, dans certains cas, la ou les cellules doivent être fixées, c'est-à-dire qu'elles sont figées et donc « tuéees » afin de les conserver dans leur état physiologique et de conserver leur morphologie. Pour finir, le marquage nécessite plusieurs étapes de modification des cellules pour permettre la fixation du marqueur aux molécules cibles et permettre une acquisition de qualité. Ces étapes modifient irrémédiablement l'échantillon et peuvent altérer la visualisation qui en est faite.

La microspectroscopie vibrationnelle est une alternative à l'imagerie par fluorescence reposant sur la mise en jeu de vibrations moléculaires pour obtenir de l'information sur la composition de l'échantillon acquis sans utiliser de techniques de marquage. En effet, lorsqu'excitées par une onde optique, les liaisons chimiques d'une molécule entrent en vibration et diffusent l'onde excitatrice. En connaissant le ou les types de vibration ainsi que la ou les liaisons impliquées dans celles-ci, il est possible de connaître la composition chimique du point de mesure. En prenant des mesures à plusieurs positions, nous obtenons une image composée de spectres possédant plusieurs centaines de canaux spectraux.

L'équipe biophotonique du laboratoire XLIM exploite ce phénomène physique par

le développement de nouveaux types de microspectroscopes dédiés à l'acquisition d'éléments biologiques. Parmi les appareils développés, l'un d'eux exploite la technologie dite CARS multiplex qui permet l'acquisition de spectres complets dans un intervalle de temps court. Les travaux présentés dans cette thèse font suite à ceux de CAPITAINE [2] et GUERENNE-DEL BEN [3] qui ont pu utiliser la microspectroscopie CARS pour étudier différents organismes biologiques allant de cellules cultivées *in vitro* au crâne de la souris. Les travaux de GUERENNE-DEL BEN *et al.* ont pu montrer que la technologie CARS pouvait servir à mettre en évidence des cellules cancéreuses dont le métabolisme a été modifié [4] mettant en évidence l'intérêt de ce type de microspectroscopie pour la recherche médicale.

Le projet CartData à l'origine de cette thèse est porté par le Laboratoire d'Excellence Σ-LIM, structure collaborative de recherche rassemblant les équipes de recherches pluridisciplinaires des deux unités mixtes de recherche CNRS XLIM et de l'IRCER sur des domaines allant de la science des matériaux et des procédés céramiques à la fabrication de composants électroniques et photoniques dédiés à des systèmes de communication intégrés, sécurisés et intelligents. Son objectif est développer une chaîne de traitement de données complète allant de l'acquisition des spectres CARS à l'analyse de ceux-ci. Les travaux de cette thèse portent sur l'analyse des données de cartographies CARS riches en informations spectrales pour fournir une information utile aux experts biologistes et physiciens amenés à utiliser ce type de microspectroscopie.

Le premier chapitre contient une contextualisation du projet de recherche avec une introduction à la biologie cellulaire et tissulaire, les éléments composant ces deux échelles du vivant ainsi que le cycle cellulaire participant à la prolifération des cellules et le développement des tissus. Suite à cette introduction, les différentes techniques de visualisation par lumière blanche et fluorescence sont présentées et leurs limites expliquées. Les phénomènes de diffusion Raman et de diffusion Raman anti-Stokes cohérente, CARS en anglais, sont introduits. Cette dernière dérive du phénomène de diffusion Raman mais utilise un mécanisme non linéaire pour acquérir l'information vibrationnelle. Ces données étant habituellement traitées par des méthodes de recouvrement de phase, les méthodes de traitement usuelles sont présentées et le choix de ne pas les utiliser dans ces travaux argumenté. Le chapitre se termine par une revue de l'état de l'art de l'utilisation du phénomène CARS pour la biologie et une présentation des cellules et tissus utilisés pour évaluer les méthodes développées durant la thèse.

Le second chapitre introduit les méthodes de projection en sous-espace avec l'analyse en composantes principales et l'algorithme isomap. Nous nous orientons ensuite vers les méthodes de résolution de courbes multivariées, *multivariate curve*

resolution (MCR) en anglais, qui permet de définir les composants principaux d'un jeu de données par leur signature spectrale et leur concentration en acquisition. La manière la plus courante de résoudre la MCR est l'utilisation de régression par moindres carrés alternés. Ainsi, cette méthode est présentée. Celle-ci nécessitant une initialisation, les méthodes d'initialisation *simple-to-use interactive self-modelling mixture analysis* (SIMPLISMA) et *vertex component analysis* (VCA) sont introduites. Ensuite, une discussion sur la sélection du nombre de composants recherchés est faite. Pour évaluer l'efficacité de la méthode, celle-ci est appliquée à différents jeux de données de cellules. La non-utilisation de méthode de recouvrement de phase sur le jeu de données est validée par l'application aux signatures spectrales des composants calculés par la MCR d'un algorithme de recouvrement de phase. Une proposition d'utilisation des spectres calculés comme base de projection pour comparer deux jeux de données de constitution similaire est faite et l'influence de la méthode d'initialisation est discutée. La MCR-ALS ne tient généralement pas compte de l'aspect spatial des données. Pour rajouter la prise en compte de la cohérence spatiale des cartographies et apporter de l'information supplémentaire à l'analyse, une contrainte de segmentation est développée et intégrée au sein de la MCR-ALS. L'utilisation de réseaux de neurones est devenue incontournable pour les tâches de classification et de segmentation : ceux-ci sont envisagés pour accomplir cette tâche. Cependant, la limite de cette approche est montrée par la difficulté à constituer une base d'apprentissage efficace pour entraîner un modèle. Finalement, la segmentation est intégrée par l'ajout d'une segmentation par contour actif en tant que contrainte dans la boucle de régression de la MCR.

Enfin, nous discutons de l'utilisation de réseaux de neurones, et plus précisément d'auto-encodeurs, pour appliquer la MCR. Activement utilisés en imagerie hyperspectrale pour le démélange, équivalent de la MCR dans ce domaine, où ils ont surpassé les approches traditionnelles, le potentiel des auto-encodeurs n'a cependant pas encore été étudié pour la microspectroscopie. Ils présentent pourtant l'avantage de pouvoir intégrer les contraintes par la structure même du réseau et l'optimisation par descente de gradient permet d'utiliser des fonctions de coût plus adaptées à l'analyse de spectres. Le chapitre commence par une formalisation de l'utilisation des auto-encodeurs pour effectuer la MCR et une revue de leur utilisation dans le domaine de l'imagerie hyperspectrale. Dans un second temps, deux modèles de l'état de l'art sont étudiés sur une image hyperspectrale puis l'un d'eux est évalué sur une cartographie CARS d'une cellule. Pour finir, une étude sur la construction d'un auto-encodeur pour appliquer la MCR à des données CARS est présentée. Celle-ci s'appuie sur l'utilisation d'un jeu de données artificielles simplifiant grandement la complexité d'une cartographie CARS

mais permettant d'évaluer quantitativement la qualité des résultats. L'intégration de contraintes de non-négativité et de l'utilisation de la structure spatiale sont discutées ainsi que les avantages et limites de l'approche.

1

Spectroscopie vibrationnelle et bioimagerie

Sommaire

VISUALISER les éléments qui composent le vivant est une opération extrêmement commune et importante pour les biologistes. Elle permet, par exemple, de suivre l'évolution d'une culture *in vitro*, observer le cycle de vie des cellules ou encore localiser des structures constituant l'échantillon, qu'il soit constitué de cellules en culture ou d'un tissu. Toutes ces informations permettent aux biologistes d'analyser l'état physiologique des échantillons observés. En biologie, six familles principales de molécules sont identifiées : l'eau, les sels minéraux, les glucides, les lipides, les acides nucléiques et les protéines. Ces familles seront présentées plus en détail dans la section 1.1. Pour analyser ces différents composants, des méthodes de spectroscopie peuvent être utilisées. Parmi elles, la diffusion Raman anti-Stokes cohérente, *coherent anti-Stokes Raman scattering* (CARS) en anglais, permet d'acquérir de l'information sur la composition chimique d'un échantillon en faisant vibrer les liaisons entre les différents atomes des molécules. Il est donc possible de visualiser les éléments composant une cellule ou un tissu dans une image grâce à l'information acquise. Dans ce chapitre, le phénomène physique CARS et l'imagerie d'échantillons biologiques sont introduits afin de contextualiser les méthodes qui sont appliquées aux données et permettre ainsi l'analyse des résultats.

L'imagerie pour la biologie et la description des entités biologiques présentes au sein des tissus et les cellules qui les composent sont introduites dans un premier temps. Le phénomène de diffusion Raman, son dérivé non linéaire CARS ainsi que la technique d'acquisition multiplex qui a permis d'acquérir nos jeux de données sont ensuite présentés. Le traitement numérique habituellement appliqué aux données CARS est décrit et les limites des méthodes actuelles mises en évidence. Le chapitre se clôture par une revue des constituants ayant pu être observés par imagerie CARS ainsi que la description des données analysées durant la thèse.

1.1 Imagerie pour la biologie

« Imager » des échantillons est une opération extrêmement courante dans le domaine de la biologie. Elle permet la visualisation des éléments biologiques à différentes échelles et, par ce biais, de s'informer sur l'état physiologique de l'échantillon. Dans le cadre de cette thèse, nous allons nous intéresser aux tissus et aux cellules qui les composent. La cellule est la plus petite unité fonctionnelle constituant le vivant et est composée de molécules comme les protéines ou les lipides, organisées entre-elles. Cette organisation est régulée spatialement et permet à la cellule de se développer et de se diviser. Dans l'organisme, les cellules sécrètent de la matière qui leur permet de s'as-

sembler et de s'organiser pour former la deuxième échelle qui nous intéresse : les tissus.

Nous présenterons ici la constitution des cellules et des tissus, le rôle de ces différents constituants ainsi que leur composition chimique. Ce dernier point est essentiel pour comprendre l'intérêt du phénomène CARS dans l'imagerie de cellules et tissus.

1.1.1 La biologie d'une cellule

Comme indiqué en section 1.1, la cellule représente la plus petite unité fonctionnelle des organismes vivants, l'ordre de grandeur est entre 10 et 40 µm en moyenne en fonction des types cellulaires. Les cellules sont constituées de sous-unités fonctionnelles analogues aux organes pour un organisme, généralement entourées d'une membrane, appelées organites. Chacun des organites a des rôles à remplir pour assurer la vie de la cellule. Les organites sont eux-mêmes constitués de molécules spécifiques qui leur permettent de se former par leurs propriétés chimiques.

1.1.1.1 Les constituants

Les molécules constituant le vivant peuvent être divisés en deux catégories : les constituants inorganiques et les constituants organiques. Au sein de la première catégorie, nous retrouvons l'eau H_2O, constituant principal du vivant. Le deuxième constituant inorganique est l'ensemble des sels minéraux sous forme d'ions qui participent à l'équilibre du fonctionnement de la cellule.

Les constituants organiques sont globalement formés de quatre familles : les glucides, les lipides, les acides nucléiques et les protéines.

Les glucides sont des molécules hydrophiles qui possèdent des propriétés énergétiques. Ils participent, entre autres, à l'adressage des protéines, c'est-à-dire le marquage pour acheminer la protéine là où elle sera utile dans la machinerie cellulaire, quand ils sont associés à celles-ci.

Les lipides jouent différents rôles au sein de l'organisme. Ils peuvent être des sources ou formes de stockage d'énergie, tout comme ils peuvent servir à la réponse d'une cellule à un signal ou encore participer à la structure de la cellule. Une catégorie de lipides qui va particulièrement nous intéresser est celle des phospholipides. Ils ont la particularité d'être amphipathiques : ils sont composés d'une partie hydrophile, la tête, et d'une partie hydrophobe, la queue. Cette particularité en fait le principal constituant

des membranes délimitant la cellule et ses organites. Au sein de la tête, nous trouvons principalement des groupements hydroxyles (-OH) ou amines ($-NH_2$). La queue, quant à elle, est composée d'une longue chaîne carbonée.

Les acides nucléiques sont des molécules hydrophiles. Ils régissent et conditionnent la vie de la cellule et parmi eux figurent les très connus acide ribonucléique (ARN) et acide desoxyribonucléique (ADN), ce dernier contient l'information génétique de la cellule. Les molécules d'ADN sont principalement situées dans le noyau, mais nous pouvons aussi en trouver dans les mitochondries (ADN mitochondrial).

Les protéines sont des molécules formées par des enchaînements d'acides aminés. Ces acides aminés sont liés par un groupement carboxyle (-COOH) d'un côté et d'un groupement amine ($-NH_2$) de l'autre. Les acides aminés sont des molécules ayant la particularité de posséder à la fois un groupement carboxyle ainsi qu'un groupement amine. Selon son type, une protéine peut être hydrophile, hydrophobe ou amphipathique. La protéine joue un rôle majeur dans tous les processus du vivant, sa structure et sa géométrie influent fortement sur sa fonction. La modification de la géométrie, ou conformation, d'une protéine est appelée repliement. Quatre niveaux de repliement protéiques sont décrits : Le repliement primaire correspond aux séquences d'acides aminés, le secondaire à un repliement régulier d'acides aminés, le repliement tertiaire correspond à la modification de la géométrie 3D d'une chaîne d'acides aminés et le quaternaire correspond aux liaisons entre les différentes chaînes d'une même protéine.

Ces différentes familles de molécules s'assemblent pour former les différents organites et compartiments de la cellule.

1.1.1.2 Les compartiments et les organites

Comme précisé au début de la section 1.1.1, les organites sont des structures membranaires qui accomplissent les rôles permettant à la cellule de vivre. Dans cette section, nous n'allons pas faire une liste exhaustive des différents organites, mais seulement présenter les principaux ainsi que quelques compartiments de la cellule [5]. Pour ce faire, nous allons nous appuyer sur la figure 1.1 qui présente un schéma simplifié d'une cellule.

La membrane plasmique matérialise les contours de la cellule et sépare son cytoplasme, composé du cytosol et des organites et composants intracellulaires, de

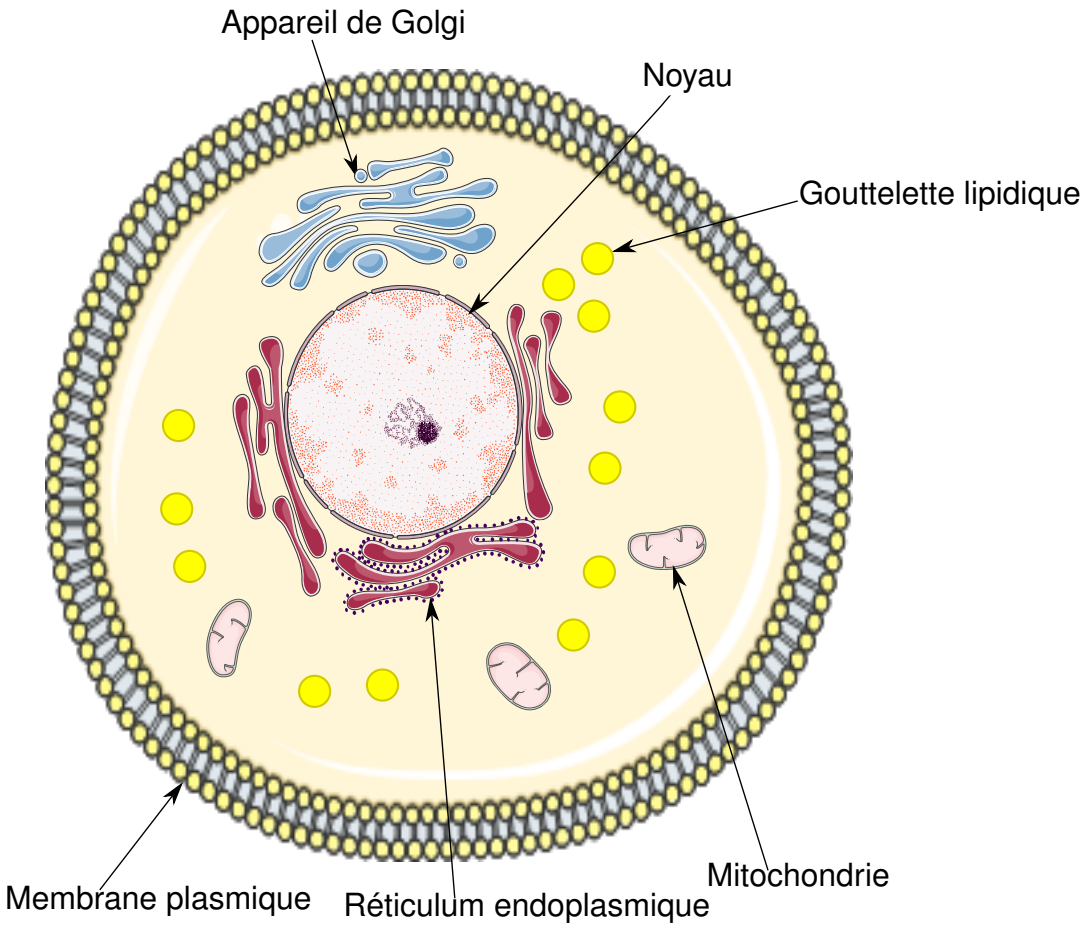


FIGURE 1.1 – Schéma simplifié d'une cellule avec ses principaux organites (modifié par Damien Boildieu, Servier Medical Art, CC BY 3.0).

l'environnement. Malgré qu'elle ne soit pas un organite, elle peut être assimilée à un compartiment de la cellule. Elle est structurée d'un bicouche de phospholipides due au caractère amphipathique des phospholipides qui la constituent. En effet, comme les queues des lipides sont hydrophobes et qu'à la fois l'environnement et le cytoplasme (environnement à l'intérieur de la cellule) sont des milieux aqueux, ceux-ci vont s'organiser en deux couches pour que les têtes hydrophiles soient en direction des milieux aqueux et ainsi les queues hydrophobes en sont isolées. Des protéines s'intègrent à différents endroits de la membrane selon leur composition. Les protéines composées d'acides aminés hydrophobes sont enchâssées au travers de la membrane, elles peuvent être intégralement membranaires ou peuvent effectuer des « passages » au travers de la membrane. Les séquences hydrophiles des protéines se retrouvent en dehors de la membrane.

Le noyau est le compartiment qui contient l'ADN. Ce dernier est stocké sous forme de chromosomes, visibles lors de la mitose, eux-mêmes composés de chromatine contient l'ADN. Il renferme aussi les nucléoles, sous-compartiment du noyau dans lequel l'ADN est transcrit en acide ribonucléique ribosomique (ARNr), structure qui participe à la traduction de l'ADN en acide ribonucléique messager (ARNm). L'ARNm permet la synthèse de protéines dans le cytoplasme par le mécanisme de transcription de l'ADN en ARN. Les constituants moléculaires principaux du noyau sont les protéines et les acides nucléiques.

Le réticulum endoplasmique (RE) est situé dans le cytoplasme, au prolongement du noyau. Il synthétise des lipides et protéines et assure la qualité de ceux-ci. Si un constituant synthétisé ne passe pas le contrôle de qualité, il est retenu ou dégradé au sein même du RE. Une fois les protéines synthétisées, le RE va modifier leur composition pour que le processus de repliement ait lieu. le RE rugueux est principalement impliqué dans la synthèse protéique tandis que le RE lisse est impliqué dans la synthèse de lipides. Parmi les lipides synthétisés par le RE lisse, des lipides de type stéroïdes comme la testostérone ou les œstrogènes jouent le rôle de molécules de communication.

À la suite du RE, les protéines et lipides passent par l'appareil de Golgi. Ses rôles sont d'assurer le bon adressage des protéines et lipides synthétisés pour qu'ils soient acheminés à leur emplacement final ainsi que la greffe de sucres aux protéines. Pour ce faire, des protéines résidentes de l'appareil de Golgi ajoutent des groupements différents aux constituants en fonction de leur composition. L'appareil de Golgi est composé de multiples enveloppes de membranes appelées saccules. L'appareil de Golgi est composé de trois compartiments : le *cis*, le médian et le *trans*. Le compartiment *cis* est le plus proche du RE et donc le premier traversé par les protéines et lipides. Il est suivi par médian et le *cis* est le dernier compartiment. Chaque compartiment va, ou non, modifier les molécules le traversant ou les retenir pour les adresser à la bonne destination.

L'organite produisant l'énergie de la cellule est la mitochondrie. Son rôle est de produire de l'énergie par la réalisation du cycle de Krebs et la chaîne respiratoire. Cette production se fait par la transformation du glucose ou de lipides en une molécule : l'Adénosine triphosphate (ATP). Cette molécule est ensuite utilisée dans de nombreux phénomènes. Elle fournit de l'énergie à la cellule en participant à des transferts de groupements phosphates à des molécules, participe à la synthèse d'acides nucléiques ainsi

qu'à la régulation de l'activité cellulaire et notamment la mort de certains constituants. La mitochondrie est un organite très particulier dans le fait qu'elle possède son propre ADN.

La gouttelette lipidique est un « réservoir » de lipides. Elle est majoritairement composée de lipides apolaires (ils ne possèdent pas de pôle électrique) dits neutres. Ces lipides sont stockés au sein des gouttelettes pour être utilisés plus tard pour les différentes activités de la cellule : production d'énergie, autres organites ou encore la signalisation cellulaire (système de communication des cellules).

Ces différents compartiments permettent à la cellule de se développer et vivre. La vie d'une cellule est régie par un cycle de différentes étapes appelé cycle cellulaire.

1.1.1.3 Le cycle cellulaire

La figure 1.2 présente un schéma du cycle cellulaire avec les différentes étapes qui composent l'interphase et la mitose. Ce cycle peut se décomposer en deux parties principales : l'interphase, une phase de « repos », ou quiescence, et la mitose qui correspond à la phase de division cellulaire.

L'interphase est l'état dans lequel la cellule va rester le plus longtemps et qui correspond à la phase où la cellule n'est pas en cours de division. Comme montré dans la figure 1.2, ces étapes sont au nombre de trois :

— la phase de croissance cellulaire (G1) : la cellule synthétise les protéines nécessaires pour la suite, son volume augmente. Les cellules différenciées, c'est-à-dire ayant acquis une fonction, comme les neurones restent en permanence à ce stade ;

— la phase de synthèse (S) : le matériel génétique de la cellule est répliqué donc la quantité d'ADN est doublée, c'est un processus long, la réplication doit être la plus fidèle possible en évitant le plus possible les mutations ;

— la phase de préparation à la division cellulaire (G2) : c'est la dernière étape de l'interphase, la cellule synthétise les protéines nécessaires à la division cellulaire et un contrôle de la qualité de la réplication du matériel génétique est effectué.

Le deuxième état est la mitose, l'état dans lequel la cellule se divise. Cet état englobe toutes les phases qui ont lieu pendant la division de la cellule en deux cellules filles. La mitose est bien plus courte que l'interphase mais bien plus complexe. La mitose se divise en cinq sous-phases :

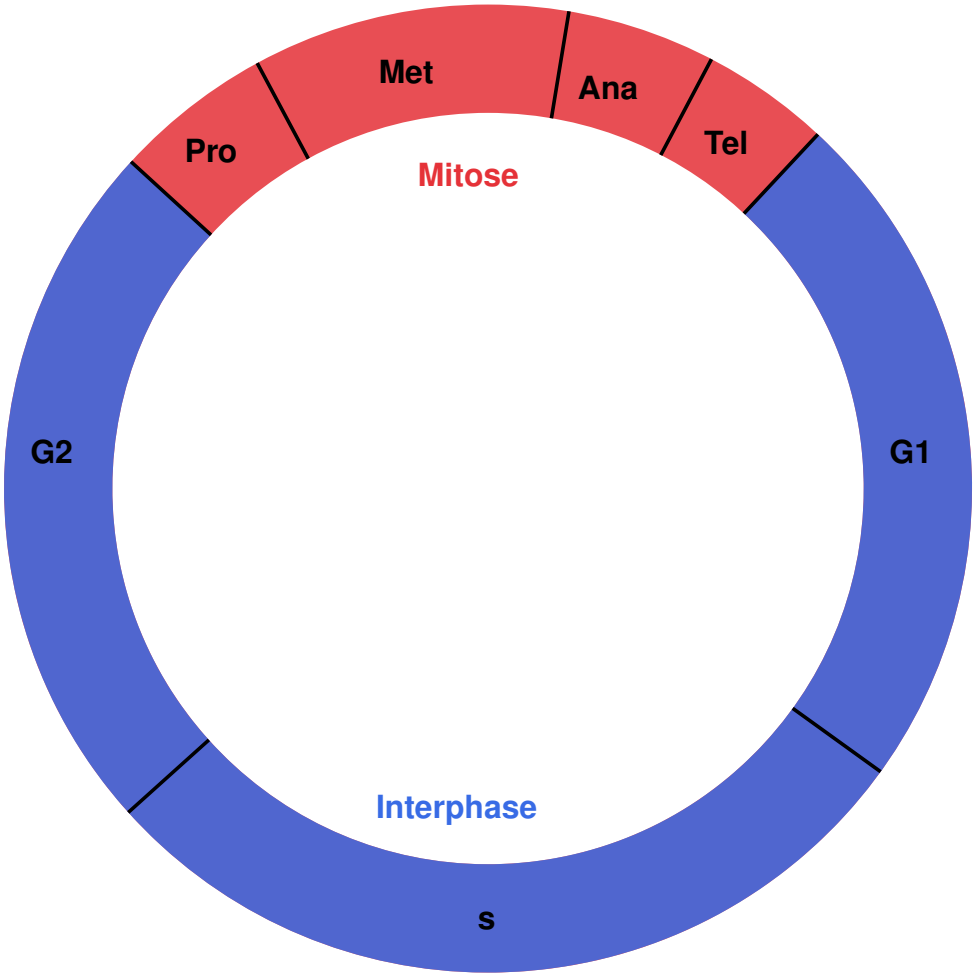


FIGURE 1.2 – Cycle de division d'une cellule : la prophase et la prométaphase sont regroupées au sein de Pro., Met. signifie métaphase, Ana. anaphase et Tel. télophase.

— la prophase, début de la mitose : la chromatine du noyau se condense, la membrane du noyau se déstructure et fusionne avec le RE, l'appareil de Golgi se décompose et les mitochondries se fragmentent [6] ;

— la prométaphase : les chromosomes sont fixés à des fibres, nommées kinétochores, en vue de la séparation du matériel répliqué, selon les auteurs, cette phase peut être intégrée au sein de la prophase comme c'est le cas dans la figure 1.2 [6] ;

— la métaphase : les chromosomes sont rassemblés à l'équateur de la cellule, appelé plaque équatoriale et les différents fragments de l'appareil de Golgi se séparent ;

— l'anaphase : les deux instances du matériel génétique sont séparées, la cellule se divise pour en former deux, les mitochondries se réassemblent ;

— la télophase : la membrane nucléaire se restructure, l'appareil de Golgi et le RE

en font de même, la chromatine se décondense. Deux cellules sont désormais formées.

Ce cycle se répète au besoin pour agrandir la population de cellules à condition que l'environnement soit viable [6]. La prolifération cellulaire est cependant contrôlée et continue en cas de croissance tissulaire. Chez l'adulte, elle est limitée au remplacement de cellule ou à des cas particuliers. Une prolifération continue non contrôlée est associée à des pathologies comme le cancer.

1.1.2 A l'échelle du tissu

Comme il a été évoqué en début de section 1.1, les cellules s'organisent entre-elles sous forme de tissu. Ce tissu se forme grâce à la sécrétion de certaines macromolécules qui lient les différentes cellules entre elles. Cet ensemble de macromolécules est appelé la matrice extracellulaire (MEC). Le tissu est donc l'ensemble composé de la MEC et des cellules formant l'échelle du vivant supérieure à la cellule. Un exemple de tissu est montré en figure 1.3, cet exemple est un épithélium, un tissu très présent chez les animaux au sein de la peau, des muqueuses ou encore de la paroi des organes.

La MEC donne la structure et la fonction du tissu au sein de l'organisme. Elle participe aussi au développement et à la régulation des cellules qui composent le tissu. Sans elle, la population de cellules forme un amas non fonctionnel pour l'organisme.

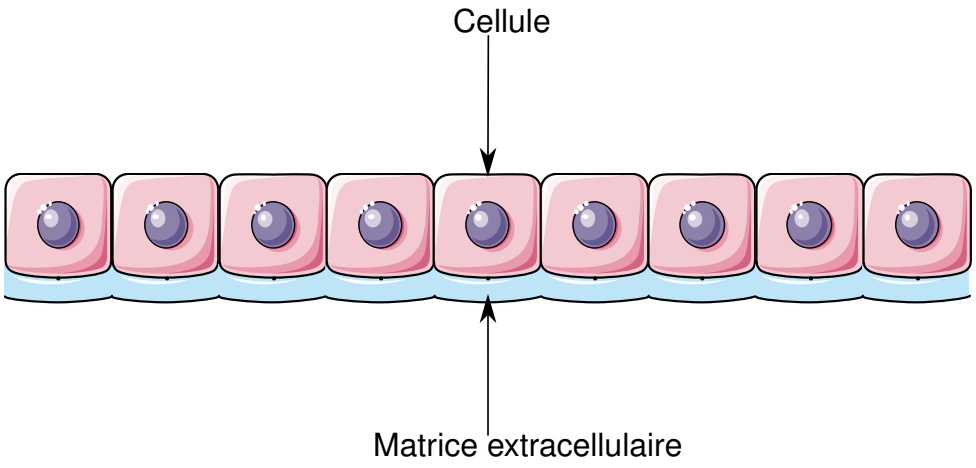


FIGURE 1.3 – Représentation schématique d'un épithélium simple.

Trois familles de molécules constituent les tissus. Les deux premières sont des molécules constituées d'acides aminés et de sucres : les glycosaminoglycanes et les protéoglycanes. Elles forment la substance fondamentale et participent à l'hydratation du tissu pour lui donner du volume en piégeant l'eau. Elles piègent et libèrent les facteurs

de croissance et les molécules de signalisation pour les cellules. Ce comportement est régi par les forces qui sont appliquées à la MEC ainsi que les différentes modifications (remodelage) qu'elle subit. La troisième famille de molécules est composée de deux catégories de protéines :

— des protéines fibreuses comme le collagène et l'élastine qui vont donner des propriétés physiques au tissu, par exemple, l'élastine est la protéine qui va conférer son élasticité au tissu ;

— des protéines plus petites participent à la communication des différentes cellules entre elles ainsi qu'à la croissance de la population pour assurer le bon fonctionnement de l'organisme. Avec certaines protéines fibreuses, elles contribuent à l'adhérence des cellules au tissu.

Il existe plusieurs types de MEC qui varient par leur structure et composition selon le type de tissu. La lame basale, par exemple, est très fine, entre 40 et 120 nm d'épaisseur. Elle polarise le tissu en formant un pôle basal matérialisant la limite du tissu. À l'opposé, les cellules adhèrent à la MEC, elles-mêmes polarisés : leur pôle basal faisant face à la lame basale et leur pôle apical, à l'opposé fait le plus souvent face à une lumière ou un autre tissu. La lame basale est présente chez tous les animaux et est associée aux tissus épithéliaux où elle sépare ce tissu des tissus conjonctifs. Elle est aussi localisée autour des cellules musculaires. La deuxième catégorie de MEC est celle des tissus conjonctifs. Celle-ci prend des formes beaucoup plus variées que la lame basale en fonction du tissu dont elle fait partie. Elle est en importante abondance au sein du tissu conjonctif, et même plus présente que les cellules qu'elle englobe. Cette MEC se retrouve dans les os, les dents, les carapaces des insectes et des crustacés, les tissus de soutien des organes ou encore les tissus adipeux. Ce dernier sera présenté plus en détail lors de l'introduction des jeux de données biologiques utilisées durant ces travaux de recherche en section 1.4.2.

Un moyen pour étudier toutes ces entités biologiques est de les visualiser à travers différents appareils d'acquisition. Parmi les différentes méthodes de visualisation, celles reposant sur des phénomènes physiques optiques sont regroupées sous le nom d'imagerie optique.

1.1.3 Techniques d'imagerie usuelles

1.1.3.1 Microscopie en champ clair

La plus ancienne et la plus simple des méthodes de microscopie est la vision par transmission de lumière blanche. L'échantillon est observé au travers d'un microscope composé de lentilles permettant de grossir l'objet visualisé. Cette méthode est rapide, simple et n'endommage pas l'échantillon. Cependant, elle présente de très importantes limites. Premièrement, elle ne permet pas de discriminer les organites de la cellule à l'exception du noyau. Deuxièmement, elle n'est pas capable d'importante résolution ; les meilleurs microscopes optiques ne pouvant aller au-delà d'un grossissement d'environ 2000. Elle sert donc principalement en routine de suivi d'évolution d'une culture.

La limite de discrimination peut être contournée en réalisant un marquage des cellules et, dans le cas des tissus, une coupe histologique. Une coupe histologique est une coupe très fine d'un tissu après inclusion de celui-ci dans une matière de dureté similaire (ex. : paraffine, résine PMMA). À la coupe, s'ajoute un marquage couleur pour mettre en évidence certains organites. Cette technique est cependant limitée à l'analyse de tissus fins et implique de nombreuses étapes qui modifient l'échantillon structurellement et chimiquement, dont la fixation des cellules du tissu. Une cellule fixée est une cellule qui a subi un procédé chimique fixant son état physiologique.

Un exemple d'acquisition en lumière blanche d'une cellule est présenté figure 1.4. Sur cette figure, nous pouvons délimiter les contours de la cellule mais guère plus, ce qui limite fortement l'analyse possible sur son état physiologique.

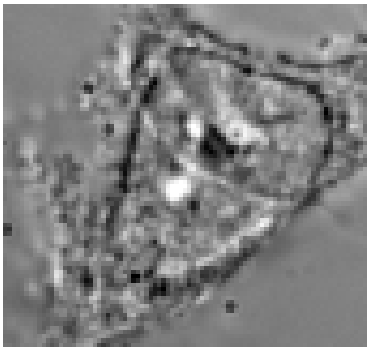


FIGURE 1.4 – Exemple d'acquisition par lumière blanche transmise d'une cellule [7].

1.1.3.2 Microscopie en fluorescence

La deuxième méthode d'imagerie optique la plus utilisée est l'imagerie par fluorescence, fréquemment mise en œuvre grâce à un microscope à épifluorescence ou, pour

une meilleure résolution spatiale, un microscope confocal à balayage laser. Ce type d'imagerie consiste à ajouter des marqueurs fluorescents dans l'échantillon qui vont se fixer aux organites ou aux molécules à observer. Pour faire ce marquage, plusieurs méthodes existent, la plus courante est l'immunofluorescence. L'immunofluorescence désigne le marquage par utilisation d'anticorps conjugués à une molécule fluorescente, fluorochrome, d'une protéine ciblée lorsqu'elle est excitée par un rayon émettant dans une plage de longueurs d'onde spécifiques. Il existe deux types d'immunofluorescence : l'immunofluorescence directe (IFD) et l'immunofluorescence indirecte (IFI). Les deux variantes reposent sur les interactions entre antigènes et anticorps.

L'IFD repose sur le marquage d'un anticorps choisi auquel un fluorochrome, une substance fluorescente, est rajouté. Pour permettre la visualisation par IFD, il faut tout d'abord réaliser un processus de marquage composé de quatre étapes :

1. La cellule à marquer est fixée ;
2. Les membranes sont perméabilisées à l'aide de substances tensio-actives appelées « détergents » ;
3. Une protéine autre qu'un anticorps est utilisée pour limiter les fixations aspécifiques, des fixations non désirées, cette étape est nommée la saturation ;
4. La cellule est incubée dans un bain avec l'anticorps primaire, conjugué à un fluorochrome, c'est l'étape du marquage.

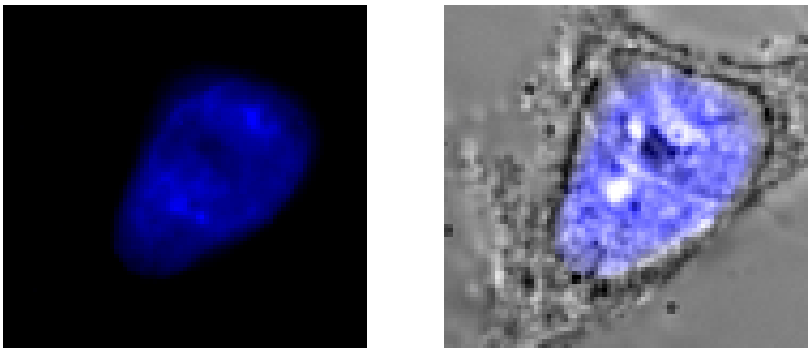
Une fois marqué, l'anticorps primaire est fixé à la molécule cible et celle-ci peut être visualisée au microscope en excitant l'échantillon par une longueur d'onde précise provoquant l'émission de l'onde fluorescente par le fluorochrome.

L'IFI est un peu plus complexe que l'IFD mais permet un signal de meilleure qualité. Dans cette méthode deux anticorps sont utilisés et non plus un seul. L'anticorps primaire, qui se fixe à la cible, n'est pas marqué et c'est un deuxième anticorps, dit secondaire, qui se lie de manière spécifique à l'anticorps primaire et qui l'est. Le processus de marquage pour l'IFI comporte les mêmes étapes que le marquage pour l'IFD mais avec une étape supplémentaire à la fin qui consiste en une nouvelle incubation avec l'anticorps secondaire pour qu'il se lie à l'anticorps primaire. L'avantage de la méthode d'IFI est de permettre de marquer plus de types d'échantillons et de produire une fluorescence plus importante que l'IFD. Elle est cependant plus complexe à mettre en œuvre que l'IFD.

Ces deux méthodes présentent toutefois des limites similaires. Les marqueurs fluorescents ne sont pas toujours compatibles avec la cellule ou l'antigène, la cible de

l'anticorps, qui doit être marqué. La fluorescence peut rentrer en conflit avec d'autres méthodes d'imagerie que nous pourrions souhaiter utiliser en complément. En effet, les longueurs d'onde utilisées pour la fluorescence sont principalement entre 350 et 800 nm. Les autres techniques d'acquisition utilisant ces fréquences, comme l'imagerie par infrarouge proche, peuvent subir une altération du signal acquis provoquée par la fluorescence. De plus, le phénomène de photo-blanchiment peut se produire lors de l'acquisition de l'image. Ce phénomène consiste en la disparition de la fluorescence quand le fluorochrome est trop longtemps excité. Mais le défaut le plus impactant est le processus de marquage qui implique de modifier chimiquement la cellule, introduisant ainsi des biais dans l'observation en plus de nécessiter une fixation de l'échantillon.

L'immunofluorescence n'est cependant pas la seule méthode d'imagerie par fluorescence. Il existe d'autres méthodes de fluorescence comme la modification génétique ou encore l'utilisation de sonde fluorescente. La figure 1.5 montre la même cellule qu'en figure 1.4 mais imagée en utilisant une sonde DAPI [8]. Cette sonde chimique intrinsèquement fluorescente illumine le noyau qui était difficilement visible dans la figure 1.4.



(a) **(b)**

FIGURE 1.5 – Cellule de la figure 1.4 imagée par fluorescence avec le marqueur DAPI [7]. (a) Fluorescence du noyau. (b) Superposition de l'image en lumière blanche avec la fluorescence.

Les méthodes utilisant des sondes ou des anticorps nécessitent de modifier l'échantillon. Lors de la modification génétique, la cellule est forcée à produire des protéines intrinsèquement fluorescentes qui peuvent être visualisées sans fixation. Ces modifications impliquent toutes un processus plus ou moins long ou coûteux. De plus, l'observation de tissus vivants par fluorescence est impossible dans la grande majorité des cas, ce qui induit la nécessité d'aller vers d'autres techniques d'imagerie pour obtenir de nouvelles informations.

1.2 Diffusions Raman spontanée et cohérente

Parmi les méthodes alternatives aux méthodes d'imageries optiques usuelles, la spectroscopie vibrationnelle et plus précisément la spectroscopie Raman et une de ses dérivées, la spectroscopie CARS, présentent des intérêts non négligeables. En effet, ces méthodes permettent d'obtenir de riches informations sur la composition chimique de l'objet analysé au travers de la vibration des différentes liaisons chimiques présentes au sein des molécules.

1.2.1 Microspectroscopie Raman

1.2.1.1 Le phénomène Raman

La spectroscopie Raman, comme son nom l'indique, repose sur l'effet Raman. L'effet Raman fut découvert en 1928 par RAMAN *et al.* Lors d'une de leurs expériences utilisant à la fois la lumière du soleil et une diode à vapeur de mercure pour illuminer des matériaux, RAMAN *et al.* remarquent que de nouvelles longueurs d'onde sont apparues dans le spectre de lumière. Plusieurs années plus tard, le phénomène est compris et baptisé Raman.

L'effet Raman repose sur la composition de la matière. Lorsqu'une onde traverse une molécule, les électrons qui relient les différents atomes peuvent être déplacés temporairement sous l'influence de celle-ci. À l'échelle de la molécule, cela conduit à une polarisation du milieu, dont les nuages électroniques se déforment de manière oscillatoire, et crée une onde électromagnétique. Cette polarisation est au cœur du phénomène Raman. Elle caractérise la densité des dipôles électriques qui se sont formés au passage de l'onde et s'exprime en $C \cdot m^{-2}$. À la polarisation s'associe un ordre variant selon le phénomène physique l'engendrant. Lorsque la polarisation est d'ordre 1, le phénomène est dit linéaire. Si la polarisation est d'un ordre supérieur à 1, le phénomène est alors non linéaire. Dans le cas du phénomène Raman, la polarisation est d'ordre 1.

Les nuages électroniques peuvent vibrer de différentes manières appelées modes vibrationnels correspondant aux caractéristiques mécaniques de la vibration. Il existe six types de vibration classés en trois catégories :

— l'élongation consiste en une oscillation dans le plan contenant les liaisons. Elle existe en mode symétrique qui indique que les oscillations des différentes liaisons sont symétriques, et asymétrique dans le cas contraire ;

— la déformation dans le plan est une rotation des liaisons en restant dans le même plan. Les deux modes sont le cisaillement qui correspond à des rotations symétriques et la bascule qui correspond aux rotations asymétriques ;

— la déformation hors du plan est une vibration sortant du plan contenant les liaisons. Comme les précédentes, elle existe en deux modes : balancement pour le symétrique, torsion pour l'asymétrique.

Le décalage de fréquence de l'onde diffusée dépend du mode de vibration ainsi que des groupements chimiques impliqués. Ainsi, il est possible d'obtenir de l'information sur la composition chimique du milieu observé en acquérant l'intensité de l'onde diffusée. Contrairement aux méthodes reposant sur la réflectance, la spectroscopie vibrationnelle Raman permet, dans une certaine limite, de faire l'acquisition des spectres à l'intérieur de l'échantillon et non seulement à la surface.

Dans la spectroscopie vibrationnelle Raman, deux ondes peuvent être utilisées pour récupérer l'information vibrationnelle : l'onde Stokes et l'onde anti-Stokes. L'onde Stokes correspond à des photons émis d'énergie inférieure à ceux absorbés alors que l'onde anti-Stokes correspond à des photons émis d'énergie supérieure. Dans la plupart des cas, c'est l'onde Stokes qui est mesurée, celle-ci a l'avantage d'être plus intense et facile à mesurer que l'onde anti-Stokes.

1.2.1.2 Description numérique du signal

Mathématiquement, les pics vibrationnels Raman prennent la forme d'une fonction lorentzienne définie de la manière suivante :

$$I_{Raman}(\omega) \propto \sum_j |\chi^{(1)}|^2 = \sum_j \frac{A_j \Gamma_j}{(\Omega_j - (\omega - \omega_s))^2 + \Gamma_j^2}. \tag{1.1}$$

Dans cette équation, ω correspond à la fréquence angulaire de l'onde excitatrice, ω_s à la fréquence angulaire de l'onde Stokes et $\chi^{(1)}$ est la susceptibilité électrique d'ordre 1 qui est fonction de la polarisation. A, Γ et Ω correspondent respectivement à l'amplitude, la largeur de bande et la fréquence du mode vibrationnel j. Lors d'une acquisition Raman, l'intensité I_{Raman} mesure le module au carré de la polarisation :

$$I_{Raman} \propto |P^{(1)}(\omega_s)|^2, \tag{1.2}$$

avec $P^{(1)}(\omega_s)$ la polarisation correspondant à l'onde Stokes.

Le spectre Raman montre l'évolution de l'intensité I_{Raman} en fonction de la dif-

férence de fréquence entre l'onde incidente et l'onde diffusée. Le nombre d'onde (relatif) ou décalage Raman, exprimé en cm^{-1}, est couramment utilisé en abscisses lors de la représentation de ce spectre.. Un spectre Raman peut être découpé en trois zones distinctes [10] :

— la zone d'empreinte digitale comprise entre 400 et 1800 cm^{-1} comprend de nombreux modes vibrationnels caractéristiques des constituants moléculaires comme les protéines, les lipides ou encore, les acides nucléiques ;

— la zone blanche entre 2000 et 2500 cm^{-1} contenant des modes vibrationnels qui n'existent pas au sein du vivant ;

— la zone CH/OH comprise entre 2800 et 3800 cm^{-1} correspond essentiellement aux modes d'étirement carbone-hydrogène (CH_2 et CH_3) et de l'eau (OH).

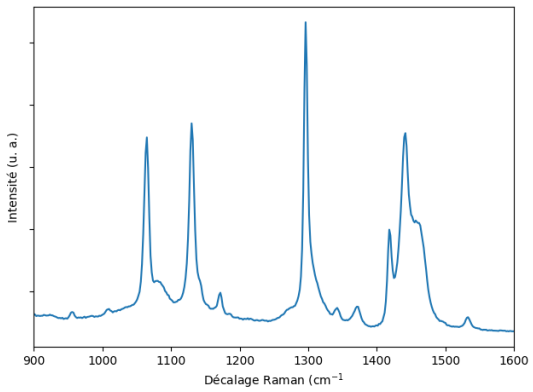


FIGURE 1.6 – Spectre Raman d'un échantillon de polyéthylène.

Un exemple de spectre Raman est montré figure 1.6. Ce spectre correspond à une acquisition de polyéthylène, un plastique très commun, entre 900 et 1600 cm^{-1} dans la zone d'empreinte digitale. Sur ce spectre, les pics Raman sont facilement identifiables et suivent des fonctions lorentziennes, le pic à 1300 cm^{-1} est un très bon exemple de pic Raman « idéal ».

Comme pour toute mesure physique, un bruit est présent au sein des données. Dans le cas du Raman, ce bruit est principalement un bruit additif et indépendant entre les différents décalages Raman. Il peut être dû à plusieurs phénomènes comme les raies cosmiques, des raies intenses dues à des photons de haute énergie qui noient l'information, ou encore une variance du flux de photons [11].

Le Raman possède toutefois ses limites. L'une d'elles, non négligeable lorsqu'appliqué à des échantillons biologiques, est sa tendance à rentrer en conflit avec la fluorescence. En effet, lorsqu'un échantillon émet de la fluorescence, l'énergie générée

par celle-ci va aussi apparaître sur le spectre Raman. Au mieux, une ligne de base qui pourra être corrigée s'ajoute au spectre, au pire de l'information vibrationnelle est perdue, masquée par une fluorescence trop intense. De fait, le Raman est souvent inutilisable si les échantillons ont été marqués ou sont fluorescents. Il faut alors se tourner vers d'autres phénomènes vibrationnels qui ne présentent pas cette limite. Parmi eux, la microspectroscopie CARS qui a été utilisée pour les données étudiées dans cette thèse.

1.2.2 Microspectroscopie CARS

La spectroscopie par diffusion Raman anti-Stokes cohérente, *coherent anti-Stokes Raman scattering* (CARS) en anglais, repose sur un phénomène physique analogue au Raman mais avec deux principales différences. Ce n'est pas l'onde Stokes qui est mesurée mais l'onde anti-Stokes avec une polarisation d'ordre 3, une polarisation qui implique trois champs, faisant du phénomène CARS un phénomène non linéaire. La spectroscopie CARS fut proposé pour la première fois par MAKER *et al.* en 1965 [12] mais le terme *Coherent Anti-Stokes Raman Spectroscopy* n'apparaît qu'en 1974 avec BEGLEY *et al.* [13]. Il est rendu possible par l'utilisation de lasers plus puissants qui contrebalancent la faible intensité de l'onde anti-Stokes émise. La diffusion CARS permet de faire des acquisitions sur une plage spectrale ne se superposant pas avec l'émission de fluorescence.

1.2.2.1 Principe

La diffusion CARS est un phénomène non linéaire résultant d'une polarisation d'ordre 3 $P^{(3)}$ suite à un mélange à quatre ondes. En effet, pour générer l'onde vibrationnelle anti-Stokes à la fréquence ω_{as}, trois autres sont nécessaires, définies par leur champ électrique E et leur fréquence ω : l'onde pompe $E_p(\omega_p)$, l'onde sonde $E_{pr}(E_{pr})$ et l'onde Stokes $E_s(\omega_s)$. La polarisation d'ordre 3 $P^{(3)}$ se définit alors comme suit :

$$P^{(3)}(\omega_{as}) = E_p E_{pr} E_S \chi^{(3)}(-\omega_{as}; \omega_p, -\omega_s, \omega_{pr}), \tag{1.3}$$

et l'équation de l'intensité I_{CARS} mesurée, de manière analogue à l'équation 1.2, devient :

$$I_{CARS} \propto |P^{(3)}(\omega_{as})|^2. \tag{1.4}$$

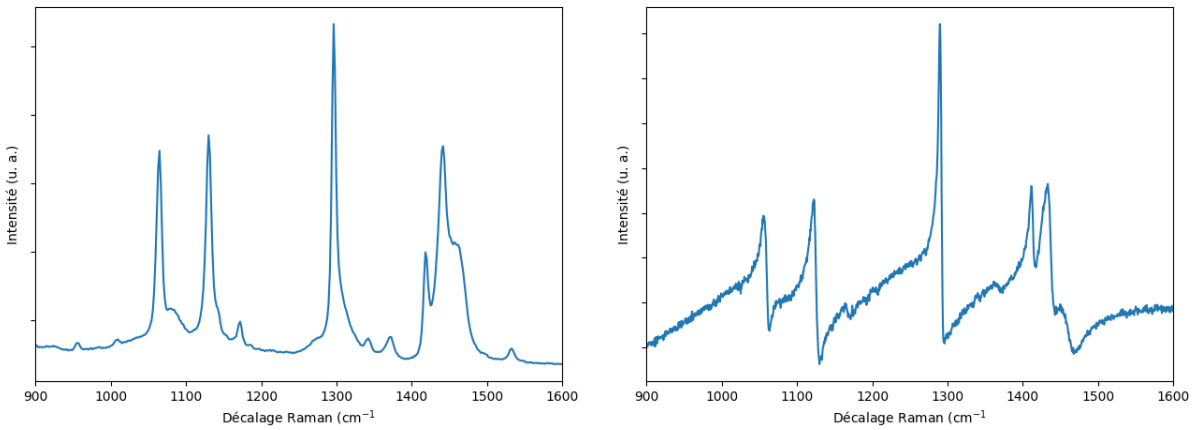
Dans le cas de la diffusion CARS, $\chi^{(3)}$, la susceptibilité d'ordre 3, est un nombre complexe composé de deux parties :

$$\chi^{(3)} = |\chi^{(3)}(\omega_{as})|e^{i\theta(\omega_{as})}, \tag{1.5}$$

$$\chi^{(3)} = \chi_R^{(3)} + \chi_{NR}^{(3)}, \tag{1.6}$$

où $\chi_R^{(3)}$ est la partie résonnante complexe contenant l'information vibrationnelle et $\chi_{NR}^{(3)}$ la partie non-résonnante réelle pouvant être assimilée à du bruit.

Contrairement au Raman, l'erreur présente au sein d'un spectre CARS n'est pas seulement additive. Cette erreur, appelée bruit de fond non-résonnant, comporte à la fois l'erreur d'acquisition mais aussi l'information non-vibrationnelle $\chi_{NR}^{(3)}$ qui modifie l'allure des spectres. En raison de la nature complexe de l'information vibrationnelle, les pics CARS ne forment pas une fonction lorentzienne mais présentent une première phase dispersive avant une brusque augmentation de l'intensité pour finir par une deuxième phase descendante similaire au Raman. Cependant, ce n'est pas la seule modification dans le spectre apportée par le bruit non-résonnant. Celui-ci modifie aussi le spectre sur son allure générale avec une variation lente formant une ligne de base qui peut être corrigée par des méthodes d'ajustement de courbes.



(a) **(b)**

FIGURE 1.7 – Spectres d'un échantillon de polyéthylène [7]. (a) Spectre Raman de l'échantillon. (b) Spectre CARS brut de l'échantillon.

La figure 1.7b montre une acquisition CARS du même échantillon de polyéthylène que dans la figure 1.6, aussi disponible en figure 1.7a. Dans la figure 1.7b, l'aspect dispersif des pics CARS en première partie, par exemple à 1300 cm-1 et la descente lente dans la figure 1.7b sont très présents. Au même décalage Raman, le signal Raman provoque un pic à l'allure d'une fonction lorentzienne dans la figure 1.7a.

Le bruit de fond non-résonnant étant particulièrement problématique pour l'analyse de spectres CARS, de nombreux travaux portent sur l'extraction du signal résonnant du spectre acquis. Les méthodes plus couramment utilisées pour effectuer cette opération seront présentées dans la section 1.3.

Une autre complexité de la microspectroscopie CARS est son temps d'acquisition. En effet, avec un système d'acquisition classique, à la fois l'acquisition spectrale et l'acquisition spatiale sont séquentielles, il faut changer la fréquence des lasers utilisés pour chaque échantillon spectral et recommencer la séquence pour chaque point spatial à acquérir. Cela entraîne un temps d'acquisition très long pour obtenir de bonnes résolutions spectrales et spatiales. Pour remédier à cette limitation, de nouvelles techniques d'acquisition ont été développées, l'une d'elles est la diffusion CARS multiplex [14].

1.2.2.2 Approche multiplex

La diffusion CARS multiplex [14] consiste en l'utilisation d'une onde monochromatique et d'une onde à large bande pour acquérir un spectre CARS complet simultanément. Un mécanisme pour obtenir une onde à large bande est la génération de supercontinuum [15].

Une méthode pour générer un supercontinuum repose sur l'utilisation d'une fibre à cristal photonique, *photonic crystal fiber* (PCF) en anglais. Lorsque l'onde émise par le laser passe au travers de la fibre, elle est modifiée et son spectre d'émission étendu, permettant ainsi de couvrir une large bande de longueurs d'onde. Ainsi, le phénomène CARS est appliqué sur toute la largeur d'émission de l'onde et il devient alors possible de récupérer un spectre complet en une seule mesure. L'acquisition spatiale est cependant faite séquentiellement pouvant entraîner un effet de décalage entre les différentes lignes acquises.

Le système d'acquisition utilisé pour acquérir les données de cellules étudiées durant cette thèse est un système CARS multiplex présenté en figure 1.8.

1.3 Traitement numérique des spectres CARS

Comme le montre l'équation 1.4, le signal CARS acquis est le module au carré d'un nombre complexe $\chi^{(3)}$. Ce nombre complexe est lui-même en deux parties, une associée à l'information vibrationnelle $\chi_R^{(3)}$ et l'autre dite non-résonnante $\chi_{NR}^{(3)}$. L'information

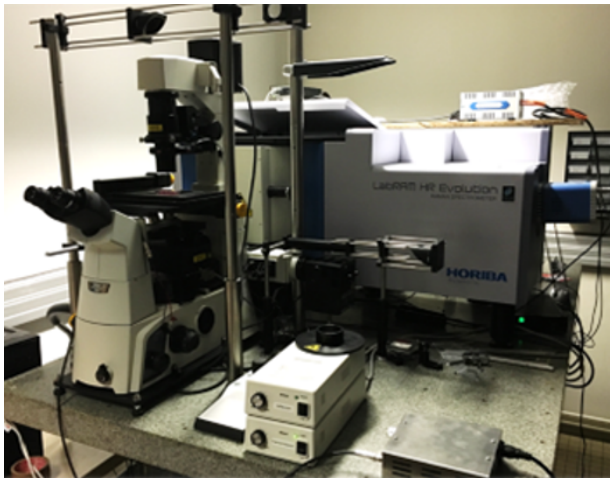


FIGURE 1.8 – Système d'acquisition utilisé pour obtenir les jeux de données des cellules étudiées.

vibrationnelle d'un pic CARS peut être définie de la manière suivante :

$$\chi_R^{(3)} = \sum_j \frac{A_j}{\Omega_j - (\omega_p - \omega_s) - i\Gamma_j}, \tag{1.7}$$

avec ω_p et ω_s respectivement la fréquence de l'onde pompe et celle de l'onde Stokes. A, Γ et Ω sont respectivement l'amplitude, la largeur de bande et la fréquence du mode vibrationnel j. Cette équation est très proche de la formulation du signal Raman en l'équation 1.1 et la partie imaginaire de $\chi_R^{(3)}$ permet d'obtenir un signal comparable au signal Raman. C'est dans le but d'obtenir des spectres qui peuvent être comparés à ceux acquis en Raman que la majeure partie des travaux concernant le traitement du signal CARS porte sur le calcul de la phase de $\chi^{(3)}$. En effet, en retrouvant la phase de $\chi^{(3)}$, nous pouvons reconstruire le nombre complexe et en récupérer la partie imaginaire :

$$\mathrm{Im}(\chi_R^{(3)}(\omega_{as})) = |\chi^{(3)}(\omega_{as})| \sin(\theta(\omega_{as})) \tag{1.8}$$

Parmi les différentes méthodes existantes, deux sont bien plus utilisées que toutes les autres : La méthode de l'entropie maximale (MEM) [16] et la méthode TDKK reposant sur le domaine temporel de Kramers-Kronig [1], que nous allons présenter dans les paragraphes suivants.

1.3.1 Méthode de l'entropie maximale

La méthode de l'entropie maximale (MEM) [17] appliquée au CARS fut présentée pour la première fois par VARTIAINEN [16] en 1992. Cet algorithme utilise la méthode de l'entropie maximale pour reformuler l'intensité CARS dans un système polynomial de rang F :

$$|\chi^{(3)}(\nu)|^2 = \frac{b_0}{|1 + \sum_{p=1}^{F} a_p e^{-i2\pi p\omega_{as}}|^2} = \left|\frac{\beta}{A_F(\omega_{as})}\right|^2, \tag{1.9}$$

avec ν le décalage Raman normalisé entre 0 et 1, $F \leq N/2$ le nombre de coefficients de corrélation, N le nombre de canaux mesurés, A_F et β sont des nombres complexes correspondant aux coefficients de l'entropie maximale.

Ce système peut ensuite être résolu en utilisant une matrice de Toeplitz :

$$\begin{bmatrix} r_0 & r_1^* & \cdots & r_F^* \\ r_1 & r_0 & \cdots & r_{F-1}^* \\ \vdots & \vdots & \ddots & \vdots \\ r_F & r_{F-1} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_F \end{bmatrix} = \begin{bmatrix} |\beta|^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.10}$$

Les coefficients r_i et leurs conjugués r_i^* sont déterminés par la transformée de Fourier du spectre traité. Une fois le système résolu, comme A_F est un nombre complexe, nous pouvons écrire :

$$A_F(\nu) = |A_F(\nu)|e^{i\iota(\nu)} \tag{1.11}$$

En considérant que l'erreur de phase $\phi(\nu)$, la différence entre la phase réelle $\theta(\nu)$ et la phase estimée par la résolution du système $\iota(\nu)$, correspond à un signal de fond sans information ne variant pas ou peu, elle peut être estimée en la considérant comme nulle dans une zone non-vibrationnelle. Ainsi, la phase de l'information vibrationnelle est retrouvée par addition :

$$\theta(\nu) \approx \phi(\nu) + \iota(\nu). \tag{1.12}$$

Une fois la phase obtenue, il est possible de calculer la partie imaginaire et de tracer un spectre comparable à un spectre Raman en utilisant l'équation 1.8.

Bien que l'article original de la méthode n'utilise pas de mesure de référence [16], de plus récentes implémentations en utilisent une [18]. Cette mesure se fait par l'acqui-

sition d'un échantillon ne comportant pas d'information vibrationnelle ou par l'acquisition du support dans lequel sera mis l'échantillon. Cette acquisition est utilisée pour estimer $\chi^{(3)}_{NR}$ et servira à normaliser les spectres mesurés en divisant les différentes acquisitions spectrales par celles de la mesure de référence. Cette mesure de référence permet d'estimer le bruit lié au système d'acquisition pour rendre possible une meilleure comparaison de spectres de différentes acquisitions [19]. L'impact de cette mesure de référence sera reprécisé dans la section 1.3.2.

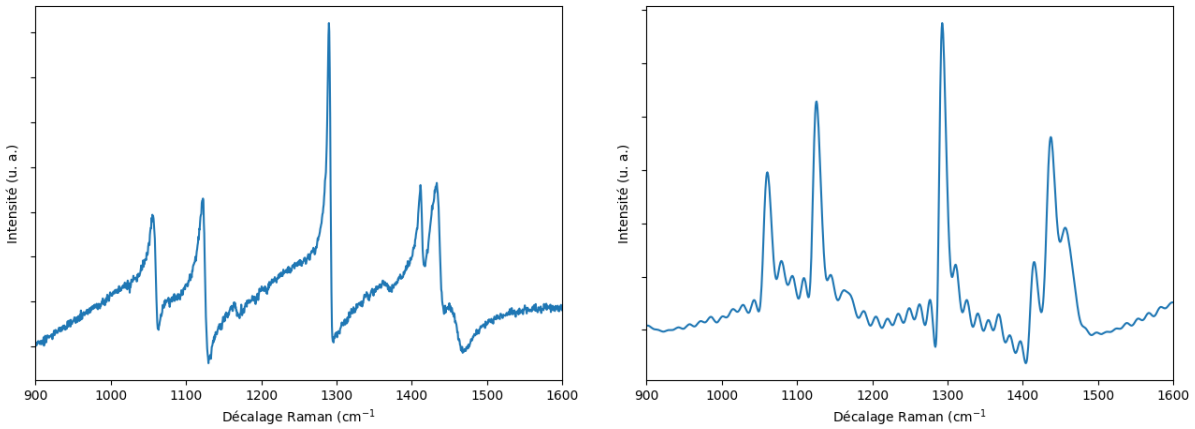


FIGURE 1.9 – Spectres d'un échantillon de polyéthylène [7]. (a) Spectre CARS brut. (b) Spectre CARS traité par la MEM et ajusté de sa ligne de base.

La figure 1.9 présente la comparaison entre un spectre CARS brut de polyéthylène en figure 1.9a avec un spectre CARS de polyéthylène traité par la MEM et ajusté de sa ligne de base. Pour estimer la ligne de base, une interpolation à l'aide de splines cubiques a été utilisé. Sur le spectre traité de la figure 1.9b, les pics sont plus proches d'une fonction lorentzienne et le spectre plus proche du spectre Raman. Cependant, des artefacts sont présents dans le spectre et les pics ont été décalés par rapport à leur position dans le spectre brut de la figure 1.7b.

1.3.2 Méthode TDKK

La méthode TDKK est la deuxième méthode la plus utilisée pour extraire la phase à partir de spectres CARS [1]. Au lieu d'utiliser la méthode de l'entropie maximale, cette méthode utilise la relation de Kramers-Kronig dans le domaine temporel dont son nom est tiré *time domain Kramers-Kronig* (TDKK) en anglais. Celle-ci est utilisée dans le domaine de l'électromagnétisme pour exprimer une relation entre la partie imaginaire et

réelle d'un nombre complexe. La relation de Kramers-Kronig, présentée plus en détail dans l'annexe A.1 définit la phase de la susceptibilité $\chi(\omega)$ de la manière suivante :

$$\theta(\omega) = -\frac{PV}{\pi}\int_{-\infty}^{\infty}\frac{\ln|\chi(\omega')|}{\omega'-\omega}d\omega', \tag{1.13}$$

avec PV la valeur principale de Cauchy.

La relation de Kramers-Kronig n'est cependant valable que pour les systèmes linéaires, ce que n'est pas la diffusion CARS en raison du bruit de fond non-résonnant. Liu *et al.* proposent une variante de cette méthode qui tient compte des spécificités de la diffusion CARS. Tout d'abord, un opérateur ψ est défini tel que :

$$\psi(f(\omega)) = \mathcal{F}[u(t)\mathcal{F}^{-1}[f(\omega)]], \tag{1.14}$$

avec f une fonction dans le domaine temporelle, u la fonction de Heaviside, $\mathcal{F}$ la transformée de Fourier et $\mathcal{F}^{-1}$ son inverse. En utilisant le théorème de la convolution, il est possible de reformuler ψ :

$$\begin{aligned}\psi(f(\omega)) &= \frac{1}{\sqrt{2\pi}}(\mathcal{F}[u(t)] * \mathcal{F}[\mathcal{F}^{-1}[f(\omega)]]) \\ &= \frac{1}{\sqrt{2\pi}}(\mathcal{F}[u(t)] * f(\omega)).\end{aligned} \tag{1.15}$$

En explicitant la convolution et à l'aide d'un changement de variable (voir annexe A.2), l'équation 1.15 peut s'écrire :

$$\psi(f(\omega_{as})) = \frac{1}{2}\left(-\frac{i}{\pi}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega' + f(\omega)\right). \tag{1.16}$$

Cette équation très proche de l'équation 1.13 peut être utilisée en remplaçant $f(\omega)$

par $\ln|\chi(\omega)|$ pour calculer la phase $\theta(\omega)$:

$$\frac{1}{2}\left(-\frac{i}{\pi}PV\int_{-\infty}^{+\infty}\frac{\ln|\chi(\omega')|}{\omega'-\omega}d\omega'+\ln|\chi(\omega)|\right)=\psi(\ln|\chi(\omega)|) \tag{1.17}$$

$$-\frac{i}{\pi}PV\int_{-\infty}^{+\infty}\frac{\ln|\chi(\omega')|}{\omega'-\omega}d\omega'+\ln|\chi(\omega)|=2\psi(\ln|\chi(\omega)|) \tag{1.18}$$

$$-\frac{i}{\pi}PV\int_{-\infty}^{+\infty}\frac{\ln|\chi(\omega')|}{\omega'-\omega}d\omega'=2\psi(\ln|\chi(\omega)|)-\ln|\chi(\omega) \tag{1.19}$$

$$-\frac{PV}{\pi}\int_{-\infty}^{+\infty}\frac{\ln|\chi(\omega'|)}{\omega'-\omega}d\omega'=\mathsf{Im}[2\psi(\ln|\chi^{(3)}(\omega_{as})|)-\ln(|\chi^{(3)}(\omega_{as})|)] \tag{1.20}$$

$$\theta(\omega_{as})=2\mathsf{Im}\left[\psi(\ln(|\chi^{(3)}(\omega_{as})|))-\frac{\ln|\chi^{(3)}(\omega_{as})|}{2}\right] \tag{1.21}$$

Pour pouvoir l'utiliser, la fonction de Heaviside pose des conditions sur le phénomène qui ne sont pas respectées dans le contexte de la diffusion CARS. En effet, le bruit de fond non-résonnant implique $\mathcal{F}^{-1}[t]\neq 0$ pour $t<0$. Liu *et al.* posent comme hypothèse que pour $t<0$, seule la partie non-résonnante émet du signal et remplacent $\psi(f(\omega))$ par :

$$\psi(f(\omega))=\mathcal{F}[\eta(t;f(\omega))] \tag{1.22}$$

$$\eta(t;f(\omega))=\begin{cases}\mathcal{F}^{-1}[f(\omega)] & \text{si } t\geq 0,\\ \mathcal{F}^{-1}[f_{NR}(\omega)] & \text{sinon}\end{cases}, \tag{1.23}$$

où $f_{NR}=\chi^{(3)}_{NR}$ est le signal non-résonnant. Cette reformulation permet de calculer la phase θ de $\chi^{(3)}$ à partir $ln|\chi^{(3)}(\omega_{as})|$:

$$\theta(\omega_{as})=2\mathsf{Im}(\mathcal{F}[\eta(t;\ln(|\chi^{(3)}(\omega_{as})|)]-\frac{\ln(|\chi^{(3)}(\omega_{as})|)}{2}). \tag{1.24}$$

Pour calculer $\eta(t;\ln(|\chi^{(3)}(\omega_{as})|))$, il est nécessaire d'avoir une mesure de référence du bruit de fond non-résonnant. En 2016, Camp *et al.* [20] modifient la méthode pour obtenir une meilleure estimation de $\chi^{(3)}_{NR}$ et ainsi améliorer le calcul de $\mathsf{Im}(\chi^{(3)}_{R}(\omega_{as}))$. Cependant, une mesure de référence est toujours requise et des erreurs demeurent. Les données étudiées n'en possédant pas, cette méthode ne peut être utilisée sur celles-ci.

1.3.3 Nouvelles approches

Plus récemment, de nouvelles méthodes ont été développées dans le but d'obtenir des méthodes toujours plus précises, extraire de nouvelles informations ou enlever

la contrainte de mesure de références. Ces méthodes se basent sur des modèles statistiques [21] ou des réseaux de neurones [22], [23].

1.3.3.1 Modèle bayésien

En 2020, HÄRKÖNEN *et al.* utilisent un modèle bayésien pour faire de l'analyse quantitative de spectres CARS [21]. Dans cette méthode, les spectres CARS sont modélisés comme un modèle statistique mélangé à un bruit additif. Ce modèle est ensuite reformulé sous forme de distribution de probabilité *a posteriori* non normalisée à l'aide d'ondelettes. Les paramètres de cette distribution sont finalement trouvés par la méthode de Monte-Carlo. Cette méthode nécessite cependant l'utilisation d'une autre méthode de recouvrement de phase afin d'initialiser la distribution pour l'échantillonnage par la méthode de Monte-Carlo.

1.3.3.2 Approches par réseaux de neurones

Comme pour une grande partie des opérations faites en traitement du signal, les réseaux de neurones sont devenus des modèles très présents et utilisés. Le recouvrement de phases de spectres CARS ne fait pas exception et diverses architectures ont été expérimentées : convolutives [24], [25], récurrentes [22] ou encore auto-encodeur [23]. Ces différentes architectures seront présentées en section 3.1.

Bien qu'obtenant des résultats expérimentaux prometteurs et ne nécessitent pas de mesure de référence, ces méthodes présentent des difficultés lorsqu'appliquées à des données fortement bruitées. De plus, ces méthodes présentent des problèmes de généralisation très limitants. En effet, ce sont toutes des méthodes basées sur un apprentissage supervisé à partir de jeux de données synthétiques. Ces jeux de données doivent être générés selon le système d'acquisition pour obtenir des résultats convaincants, ce qui limite très fortement la portée de ces méthodes [23].

1.3.3.3 Non-usage de méthodes de recouvrement de phase

Le recouvrement de la phase est un domaine très actif dans le traitement de signal CARS. Cependant, le traitement du bruit de fond non-résonnant est complexe et nécessite la plupart du temps une mesure de référence. Dans les méthodes de référence comme la MEM et le *time domain Kramers-Kronig* (TDKK), le signal obtenu conserve ou introduit de nouveaux pics dispersifs. Lorsqu'un pic a une intensité bien plus élevée que les autres dans le spectre, ces derniers peuvent disparaître lors du

calcul de Im$(\chi_R^{(3)}(\omega_{as}))$. Les méthodes plus récentes par réseau de neurones obtiennent de meilleurs résultats en levant la contrainte de la mesure de référence mais elles sont très sensibles au paramétrage et nécessitent de générer un jeu de données synthétique correspondant au système d'acquisition. De plus, l'absence de mesure de référence dans les données utilisées durant la thèse rend plus complexe encore l'utilisation de ces méthodes.

Pour ces raisons, aucune méthode de recouvrement de phase ne sera utilisée dans les travaux présentés, les spectres CARS bruts seront directement analysés. Les analyses des spectres obtenus par les méthodes développées seront comparées avec celles qui peuvent être faites à partir de spectres traités par la MEM afin de s'assurer de la validité de ce choix.

Maintenant que le fonctionnement des échelles du vivant étudiées ainsi que le fonctionnement du phénomène CARS et son traitement numérique ont été présentés, l'utilisation de ce phénomène dans le contexte de la microscopie d'éléments biologiques va être présentée ainsi que les jeux de données utilisés pour appliquer les méthodes développées.

1.4 Application de la microscopie CARS à la bioimagerie

1.4.1 État de l'art

Depuis le début des années 2000, les spectroscopies Raman et CARS ont été de plus en plus utilisées pour imager et étudier des éléments biologiques. Une majeure partie du travail a été d'identifier les pics vibrationnels pouvant être associés à des constituants moléculaires ou des organites. En les identifiant, il est possible d'intégrer le spectre sur la plage spectrale souhaitée pour obtenir une image qui mettra en évidence l'élément désiré. C'est l'opération la plus communément appliquée pour générer des images à partir des différents pics. D'autres méthodes ont pu être utilisées comme des méthodes de clustering K-moyennes [26] ou des analyses en composantes principales [27] pour classer les différents composants d'un échantillon et calculer un spectre moyen associé à chacun des composants. Toutes les méthodes utilisées relèvent de l'apprentissage non-supervisé et il n'existe pas de banque de données de spectres CARS ou Raman de référence pour les différents éléments biologiques. L'ensemble de ces travaux a permis de définir une liste des éléments identifiés et

des décalages Raman associés [28]. Il a cependant été constaté que le calcul de la partie imaginaire de $\chi^{(3)}$ provoque un décalage de la position des pics comme dans la figure 1.9 [29], [30].

Pic $\lvert\chi^{(3)}\rvert^2$ (cm^{-1})	Pic $Im[\chi^{(3)}]$ (cm^{-1})	Mode vibrationnel	Élément
3165	3200	O-H s-élong.	Eau
3056	3066	C-H élong. (aromatique)	Protéines
3007	3017	=C-H élong.	Lipides
2975	2953	CH_3 a-élong.	ADN & ARN
2920	∼2930	CH_3 s-élong.	Protéines/Lipides
2882	2902	CH_2 a-élong.	Lipides/Protéines
2844	2854	CH_2 s-élong.	Lipides

TABLEAU 1.1 – Position des pics vibrationnels CARS ainsi que les modes vibrationnels et éléments associés [28].

Le tableau 1.1 montre une partie des pics identifiés qui vont être utiles pour l'analyse des résultats obtenus. Ces pics peuvent être associés à quatre familles de molécules différentes :

- les liaisons associées aux lipides sont responsables des ondes générées à 3007, 2882 et 2844 et dans une moindre importance à 2920 cm^{-1} ;
- les protéines ont leurs liaisons qui vibrent à 3056 et 2920 cm^{-1} et plus légèrement à 2882 cm^{-1} ;
- l'eau émet un signal à 3165 cm^{-1} ;
- les liaisons associées aux acides nucléiques (ADN et ARN) génèrent un signal à 2975 cm^{-1}.

Ces différentes méthodes, combinées avec des marqueurs ou d'autres méthodes d'acquisitions, ont permis de visualiser de multiples constituants et organites parmi lesquels : gouttelettes lipidiques [4], [31], noyau, cytoplasme [26], mitochondries [32], chromosomes et réticulum endoplasmique [7]. Les précédentes études portaient sur des cellules mais des tissus ont aussi pu être observés [33], [34]. La visualisation de ces constituants a permis la comparaison de différentes lignées cellulaires [27], l'observation de différents processus biologiques [35]-[38] ou encore la détection de cellules cancéreuses [4], [39]-[41] et l'observation de leur traitement [42]-[45].

Bien qu'ayant déjà permis de nombreuses applications, produire des images en intégrant sur une plage spectrale limite les capacités de discrimination des organites ou molécules. En effet, les organites étant composés de multiples molécules complexes agencées spatialement, intégrer sur une bande spectrale ne permet pas de visualiser

un seul élément en particulier. S'il devient possible de caractériser les spectres des organites composant l'échantillon observé, la discrimination et la visualisation de ceux-ci peuvent être améliorées. L'absence de base de données référençant ces spectres contraint à l'utilisation de méthodes d'apprentissage non-supervisé.

Parmi les différentes opérations pouvant être appliquées à des données spectrales, la projection en sous-espace consiste en la réduction de dimensionnalité spectrale des données. Cette réduction de dimension est rendue possible par le calcul d'une nouvelle base de représentation permettant de mieux mettre en évidence la dispersion des données. Cette base de plus petite dimension doit être construite de sorte à minimiser l'information perdue. Une fois définie, il est possible d'y projeter les données à analyser et de générer une image par dimension de la base.

Avant de présenter les différentes familles de méthodes de projection en sous-espace, les différents jeux de données CARS sur lesquels seront appliquées les méthodes doivent être présentés.

1.4.2 Présentation des jeux de données étudiés

1.4.2.1 Cartographies de cellules

Toutes les cartographies de cellules étudiées partagent les mêmes paramètres spectraux. Ce sont des cartographies de la zone CH entre 3200 et 2500 cm^{-1} avec une résolution spectrale de 0,8 cm^{-1} donnant 916 échantillons par spectre, la résolution spatiale est la même pour toutes les cellules : $\sim$300 nm pour la résolution latérale et 2 µm pour la résolution axiale [7]. Cependant, le nombre de pixels peut différer entre les cellules.

La lignée cellulaire utilisée pour évaluer les méthodes est la lignée HEK-293. Une lignée cellulaire est une population de cellules, établie à partir de cellules cancéreuses ou modifiées génétiquement, se subdivisant, théoriquement, à l'infini. La lignée HEK-293 correspond à des cellules embryonnaires rénales humaines. De plus, les jeux de données expérimentales utilisés ont été obtenus à partir de cellules dans différents états dont des cellules vivantes et des cellules fixées.

Certaines cellules ont été marquées pour la fluorescence avec le marqueur Hoechst 33342. Le Hoechst 33342 est un marqueur intrinsèquement fluorescent mettant en évidence le noyau grâce à son affinité chimique pour les acides nucléiques. Il possède l'avantage de fonctionner sur les cellules vivantes contrairement au DAPI qui nécessite de fixer la cellule.

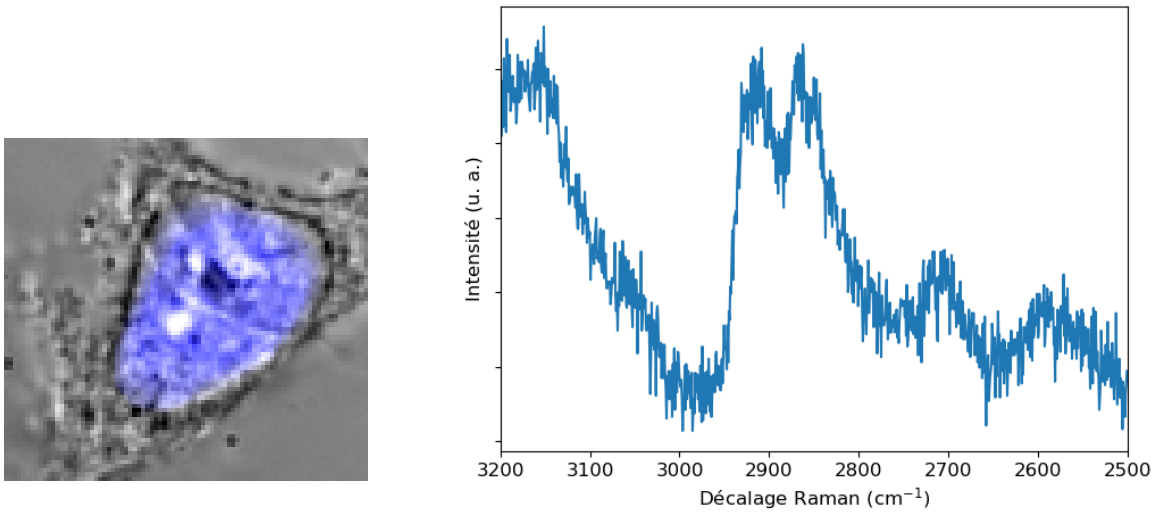
Des jeux de données obtenus à partir de cellules de la lignée HEK-293 sont

génétiquement modifiées pour surexprimer le récepteur kinase B de la tropomyosine, TrkB en anglais, un facteur de croissance (molécule de communication intercellulaire) important dans le système nerveux. TrkB est un récepteur sur lequel se greffe la protéine facteur neurotrophique dérivé du cerveau, BDNF en anglais. Normalement exprimée dans le système nerveux, l'association des deux peut être signe de tumeur lorsqu'elle est retrouvée dans d'autres tissus [46]. L'une des deux cellules modifiées a été traitée par ajout de BDNF dans le milieu de culture alors que l'autre non. L'acquisition a été faite 72h après l'injection du BDNF ou de BDNF recombinant afin de pouvoir étudier les différences physiologiques dues au BDNF.

La cellule HEK-293 fixée en interphase est composée de 80 lignes et 85 colonnes [7]. Cette cellule présentée en figure 1.10 servira de cellule de référence pour l'analyse des résultats. Cette cellule a été marquée avec de la fluorescence DAPI afin de mettre en évidence le noyau. Une acquisition de la lumière transmise de l'échantillon avec la fluorescence en surimpression est montrée en figure 1.10a. Sur cette figure, le noyau apparaît en bleu grâce au marquage DAPI. Autour du noyau, il est possible de voir le cytoplasme qui se « diffuse » légèrement dans l'environnement. Cette « diffusion » est un processus qui peut apparaître lors du marquage de la cellule. La figure 1.10b montre un spectre acquis de cette cellule. Le rapport signal sur bruit est assez faible avec des variations à haute fréquence importantes. Ce bruit est à prendre en compte dans les méthodes d'analyse de données qui doivent être utilisées pour limiter son impact dans les résultats. Il est tout de même possible d'identifier des pics à 3150, 2920 et $\sim$2850 cm^{-1} qui indiquent une présence d'eau, de protéines et de lipides dans ce pixel.

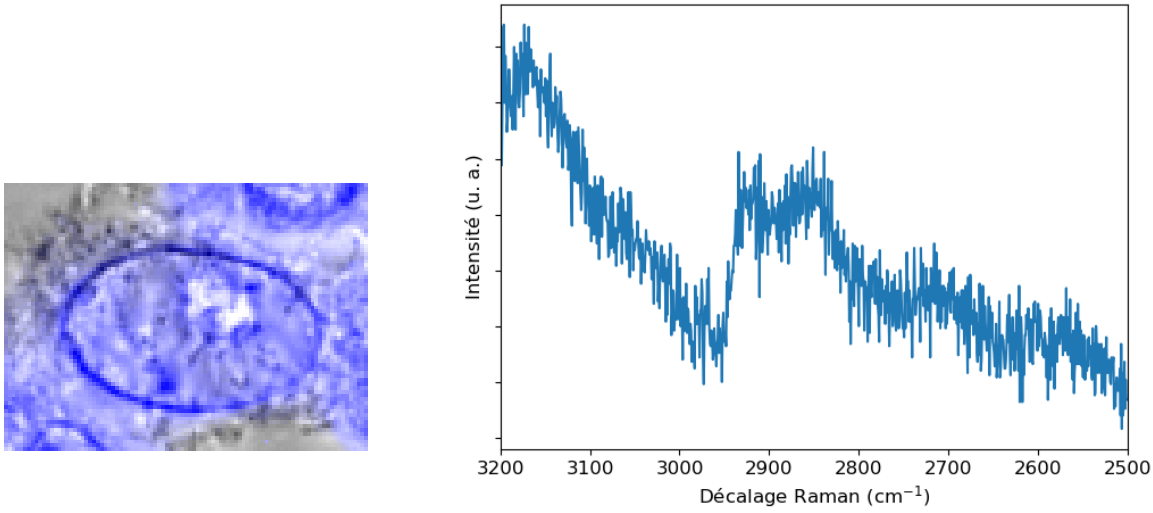
La cellule HEK-293 vivante en interphase est composée 70 lignes et 95 colonnes. Cet échantillon a été marqué pour la fluorescence par Hoechst 33342. Une acquisition de la lumière transmise de l'échantillon avec la fluorescence Hoechst 33342 en surimpression est présentée en figure 1.11a. Bien que l'image soit centrée sur une seule cellule, il est possible d'identifier des parties d'autres cellules sur la partie supérieure droite et inférieure gauche. Le spectre est présenté en figure 1.11b.

Une cellule HEK-293 vivante en interphase surexprimant TrkB composée de 85 lignes et 75 colonnes. Cette cellule a été marquée avec le marqueur Hoechst 33342. Sur la figure 1.12a, il est possible d'identifier le noyau et le cytoplasme de la cellule ainsi que ceux d'une deuxième cellule dans le coin inférieur gauche. Le bruit à haute fréquence du spectre présenté en figure 1.12b est similaire à celui de la figure 1.10b.



(a) **(b)**

FIGURE 1.10 – Cellule HEK-293 fixée en interphase : (a) image en lumière transmise avec la fluorescence DAPI du noyau en surimpression, (b) spectre d'un pixel de la cellule.

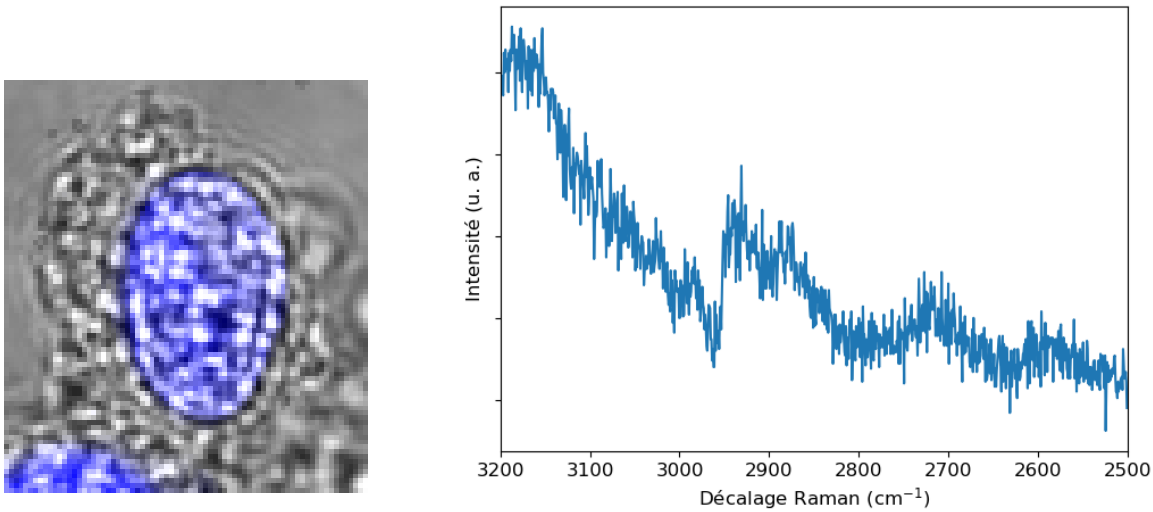


(a) **(b)**

FIGURE 1.11 – Cellule HEK-293 vivante en interphase : (a) image en lumière transmise avec la fluorescence Hoechst 33342 du noyau en surimpression, (b) spectre d'un pixel de la cellule.

Les pics de l'eau, des protéines et des lipides sont tout de même identifiables sur ce spectre bien que le signal soit plus faible que dans la figure 1.10b.

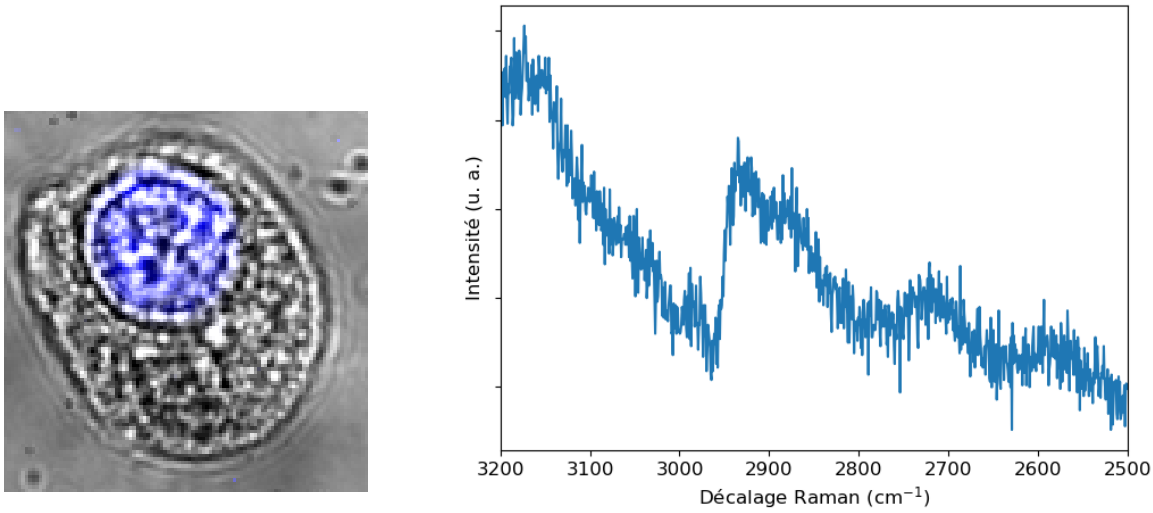
Une cellule HEK-293 vivante en interphase surexprimant TrkB exposée au BDNF composée de 110 lignes et 105 colonnes. Cette cellule a aussi été marquée avec le mar-



(a) **(b)**

FIGURE 1.12 – Cellule HEK-293 modifiée pour exprimer TrkB sans BDNF : (a) image en lumière transmise avec la fluorescence Hoechst 33342 du noyau en surimpression, (b) spectre d'un pixel de la cellule.

queur Hoechst 33342. L'image en lumière transmise avec la fluorescence surimprimée en figure 1.13a permet de très bien distinguer le noyau et le cytoplasme de la cellule. En ce qui concerne le spectre en figure 1.13b, le spectre est similaire à celui de la figure 1.12b à l'exception d'un signal moins important au niveau du pic de lipides à 2850 cm^{-1}.



(a) **(b)**

FIGURE 1.13 – Cellule HEK-293 modifiée pour exprimer TrkB avec BDNF : (a) image en lumière transmise avec la fluorescence Hoechst 33342 du noyau en surimpression, (b) spectre d'un pixel de la cellule.

1.4.2.2 Cartographies de tissus

Une cartographie de tissu adipeux blanc prélevé chez la souris est utilisée pour évaluer les méthodes sur une autre catégorie de données CARS. Le tissu adipeux est un type particulier de tissu conjonctif. Il possède une MEC aqueuse et existe sous trois formes : adipeux blanc, brun ou beige. Le rôle du tissu adipeux brun est de lutter contre l'abaissement de la température corporelle en convertissant les acides gras en chaleur. Il va se trouver principalement au niveau du buste. Le tissu adipeux blanc est le tissu stockant la graisse de l'organisme. Sa localisation varie selon l'espèce et le sexe de l'animal.

Cette acquisition nous a été fournie par H. KANO de l'université de Kyushu, Japon. Elle est composée de 15 tranches de 201 par 201 pixels. Toute la plage vibrationnelle est couverte entre 3900 et 195 cm^{-1} pour un total de 1340 mesures par spectre. Aucune image en lumière transmise n'est disponible mais un exemple de spectre de la cartographie est montré en figure 1.14. La ligne de base influence moins l'allure du spectre que pour les acquisitions de cellule. De plus, le signal de la zone CH est bien plus intense que celui des autres zones, c'est un comportement courant dans les spectres CARS (la zone CH étant la zone avec le signal le plus fort).

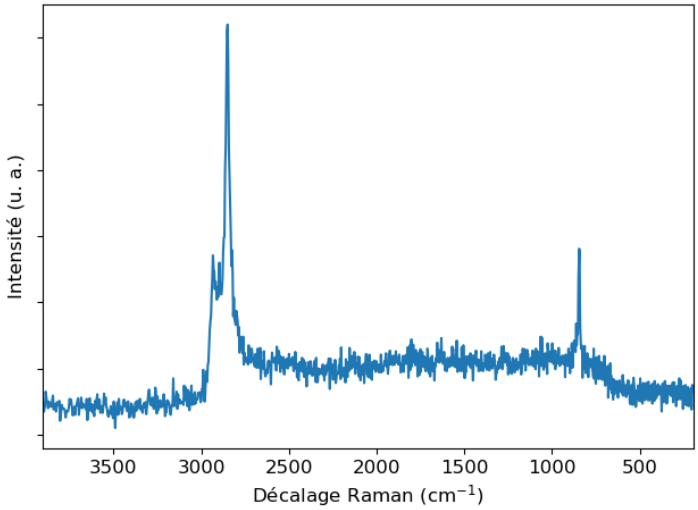


FIGURE 1.14 – Spectre d'un pixel de l'acquisition d'un tissu adipeux blanc.

1.5 Conclusion

Dans ce chapitre, le contexte du projet de thèse a été développé. Les informations sur la biologie des cellules et des tissus nécessaires à l'analyse des résultats ont été données. La composition des cellules et tissus, leur physiologie et leurs rôles au sein de l'organisme ont été présentés. Les méthodes d'imagerie optique usuellement utilisées pour imager des échantillons biologiques ont été exposées avec leurs limites pour discriminer un compartiment d'un échantillon sans l'altérer.

Le fonctionnement de la spectroscopie Raman a été introduit suivi de celui de la spectroscopie CARS. Ces introductions ont été suivies d'une présentation des méthodes de traitement numérique habituellement appliquées aux spectres CARS afin d'extraire l'information vibrationnelle des spectres. Les limites de ces méthodes ont été exposées avec les raisons motivant le choix de ne pas les utiliser en prétraitement des méthodes développées.

Un état de l'art de l'imagerie pour la biologie utilisant des données Raman ou CARS a été présenté ainsi que les données qui seront utilisés pour évaluer l'efficacité des méthodes développées. Cet état de l'art a montré l'intérêt de pouvoir caractériser le spectre des différents compartiments membranaires et des organites des cellules pour améliorer l'imagerie de données biologiques à partir de spectres CARS. L'absence de base de données de référence pousse à se tourner vers des méthodes d'apprentissage non-supervisé. Les méthodes de réduction de dimension permettant de définir une meilleure représentation de la dispersion de données répondent à cette problématique.

2

Projection en sous-espace

Sommaire

LES méthodes de projection en sous-espace construisent un nouvel espace de plus petite dimension que celui d'origine mais offrant une meilleure représentation des données. Dans le contexte de ce projet de thèse, deux objectifs sont identifiés. Le premier est de faire correspondre la base de dimension réduite aux organites ou compartiments membranaires de l'échantillon biologique. Dans la suite du manuscrit, nous nommerons composants les éléments formant la nouvelle base. Le second objectif est de réussir à associer aux différents composants un spectre CARS permettant d'étudier leur composition chimique.

Il existe principalement deux classes de méthodes de projection en sous-espace. La première classe calcule une nouvelle base respectant certaines propriétés statistiques entre les différents axes. La seconde utilise la minimisation d'une erreur de reconstruction pour calculer la nouvelle base. Dans ce chapitre, différentes méthodes de ces deux classes sont introduites et évaluées sur des données CARS. La méthode de résolution en courbes multivariées appartenant à la deuxième famille sera présentée plus en détail. Le choix du nombre de composants et l'impact de la méthode d'initialisation seront discutés. Une discussion sur l'intégration de contraintes spatiales au sein de la résolution de courbes multivariées et plus particulièrement de la segmentation de cellules clôt le chapitre.

2.1 Méthodes de réduction de la dimensionnalité

Les données traitées sont des images multivariées avec $H \times W = M$ pixels qui contiennent chacun N éléments spectraux. L'objectif des méthodes étudiées est de réduire le nombre de dimensions spectrales de N à K composants. Ces K composants permettent ensuite d'obtenir K images et signatures spectrales pouvant être analysées par différents experts.

Dans cette section, les méthodes d'analyse en composantes principales et isomap sont présentées et nous discuterons de leurs limites. Ces deux méthodes cherchent à maximiser la décorrélation des données dans un espace de plus petite dimension.

2.1.1 Analyse en composantes principales

2.1.1.1 Présentation de l'analyse en composantes principales

L'analyse en composantes principales (ACP) est une des plus anciennes méthodes de réduction de dimension. Elle est découverte en 1901 par PEARSON, puis

redécouverte et nommée ainsi en 1933 par HOTELLING. L'ACP repose sur les statistiques pour calculer la base optimale maximisant la décorrélation des données. Cette base optimale est trouvée en appliquant une décomposition en valeurs singulières, *singular value decomposition* (SVD) en anglais, de la matrice variance-covariance Σ des données $D \in \mathbb{R}^{M \times N}$ avec M individus et N variables. La matrice de variance-covariance est une matrice carrée symétrique tel que :

$$\Sigma = D^T \mathfrak{P} D - \bar{d}\bar{d}^T, \tag{2.1}$$

avec $\mathfrak{P}$ la matrice de poids diagonale dont les valeurs sont toutes égales à $1/M$ et $\bar{d}$ la matrice des moyennes de chacune des variables. Chaque élément de la matrice est à interpréter comme la « relation » entre les deux variables.

La nouvelle base correspond aux K vecteurs propres $V^T_{1,..,K}$ de Σ avec les plus grandes valeurs propres $\Lambda_{1,..,K}$ issues de la SVD de Σ :

$$\Sigma = V \Lambda V^{-1}. \tag{2.2}$$

Les données sont ensuite projetées sur la nouvelle base :

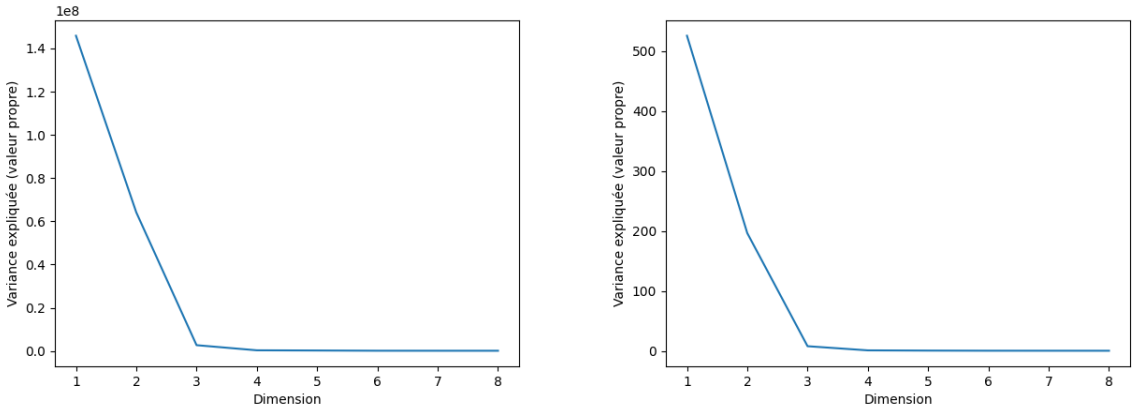
$$Y = DV. \tag{2.3}$$

Dans le cas particulier où $K = N$ et où la base d'origine est orthogonale alors, la nouvelle base correspond à une rotation et translation de celle d'origine.

Le plus souvent, la matrice de données D est centrée, cela signifie que chaque variable est centrée à 0 en soustrayant la valeur moyenne de chacune d'elle. Ce centrage permet d'éviter que la moyenne influe sur la décomposition. Par ailleurs, la matrice peut aussi être réduite, signifiant que chaque variable va être divisée par son écart-type pour normaliser les valeurs entre -1 et 1. Réduire la matrice permet d'éviter qu'une variable avec une échelle de valeurs plus importante que les autres perturbe la décomposition. Lorsque, calculée à partir de données centrées et réduites, la matrice de variance-covariance est bornée entre -1 et 1 et est alors une matrice de corrélation.

2.1.1.2 Application à des données CARS

Pour évaluer l'efficacité de l'ACP sur des données CARS, la méthode a été appliquée sur les spectres de la cellule HEK-293 fixée en interphase présentée en section 1.4.2.1 où $M = 85 * 80 = 6800$ et $N = 916$.



(a) **(b)**

FIGURE 2.1 – Variance expliquée par dimension de l'ACP, (a) ACP appliquée aux données centrées, (b) ACP appliquée aux données centrées et réduites.

Pour choisir la valeur de K, la méthode empirique du « coude » [49] est communément utilisée. Cette méthode consiste en une observation de la courbe dessinée par les valeurs propres et la sélection comme valeur de K le nombre de dimensions correspondant au point d'inflexion de la courbe. En effet, les valeurs propres contiennent la part de variance portée par chaque axe. L'ajout d'axes situés après le point d'inflexion n'apporte pas une grande quantité d'information. La figure 2.1a montre la variance expliquée correspondant à une ACP appliquée à la matrice centrée en figure 2.1a et la figure 2.1b celle correspondant avec matrice centrée et réduite. L'impact de la normalisation des données influence l'ordre de grandeur des des valeurs propres mais très peu l'allure de la courbe et donc le choix de la valeur de K. Ce faible impact s'explique par le fait qu'aucun décalage Raman ne provoque une intensité bien plus importante que les autres et donc toutes les variables sont sur une échelle de grandeur comparable. Nous observons bien sur la figure 2.1 que la courbe des valeurs propres décrochent après la dimension 3, ainsi la valeur de K pour ce jeu de données est 3.

La figure 2.2 présente la projection des données dans l'espace de dimension réduite. La première ligne correspond à l'ACP appliquée aux données centrées et la deuxième aux données centrées et réduites. Comme la courbe des variances expliquées par dimension le laissait supposer, la normalisation des données influe peu sur le résultat. La première dimension met principalement en lumière des éléments présents dans le noyau de la cellule. La seconde illumine toute la cellule avec une plus forte intensité au sein du noyau. La dernière dimension met en avant le cytoplasme de la cellule et une partie de l'environnement.

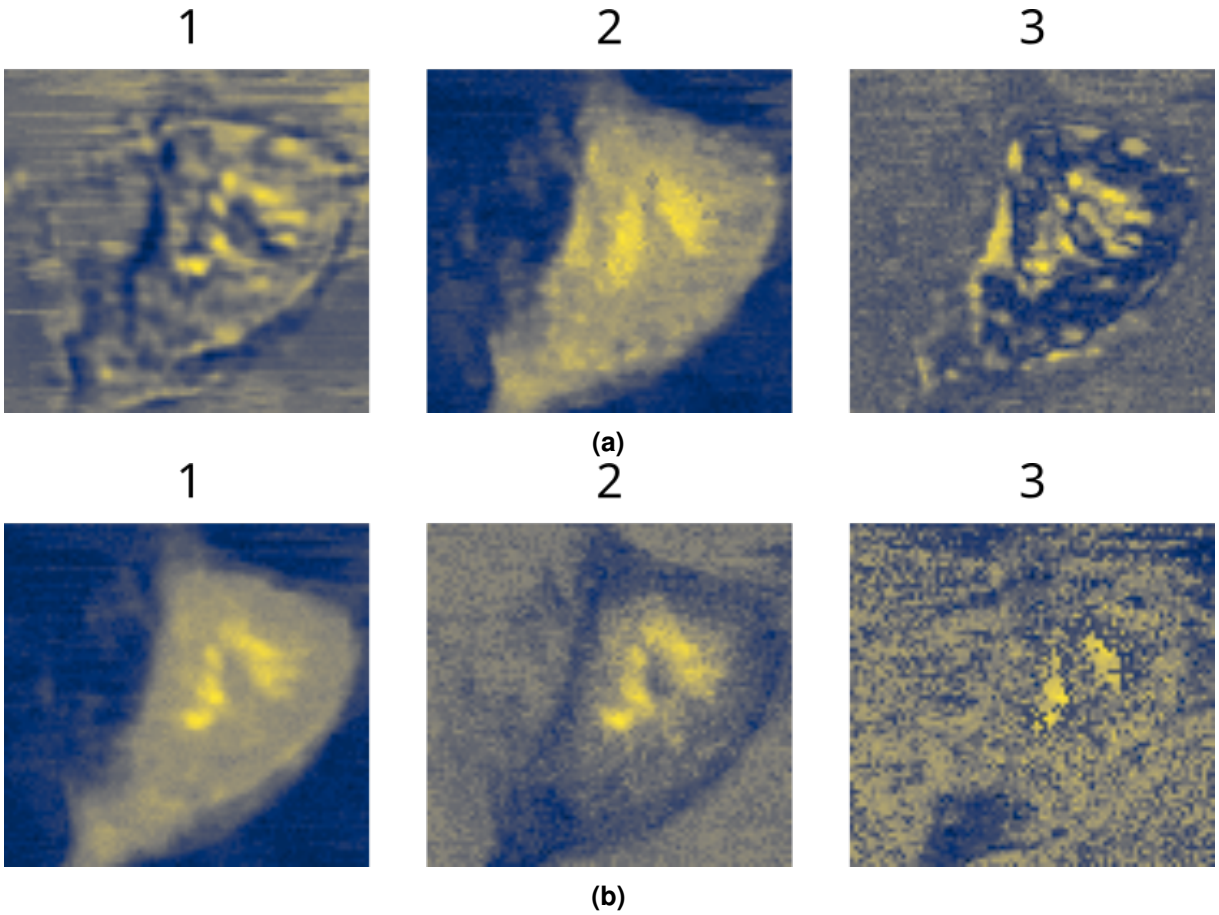


FIGURE 2.2 – Projection des données sur les trois principaux axes calculés par l'ACP, ACP appliquée aux données centrées, (b) ACP appliquée aux données centrées et réduites.

Les axes calculés par l'ACP sont montrés en figure 2.3, la première ligne correspond à l'ACP appliquée aux données centrées et la deuxième aux données centrées et réduites. Contrairement à la variance expliquée et aux projections, les axes de la nouvelle base sont influencés par la normalisation des données. Les bandes rouges correspondent à des zones vibrationnelles de lipides, les bandes vertes à des protéines et la cyan à l'eau. L'effet est principalement visible dans le premier axe où l'allure de la courbe a été fortement altérée. Les pics entre 2500 et 2860 cm^{-1} ainsi que celui à 2920 cm^{-1} sont fortement aplanis. Ce phénomène d'aplanissement des pics est aussi observable dans une moindre mesure sur le deuxième axe mais l'est beaucoup moins dans le troisième.

Les premiers composants expliquant le plus la variance des données, ils sont les plus sensibles aux valeurs des variables et sont plus impactés par la normalisation (division par l'écart-type) des données. Les axes calculés comportent à la fois des valeurs positives et négatives qui mettent en opposition les variables associées. Un

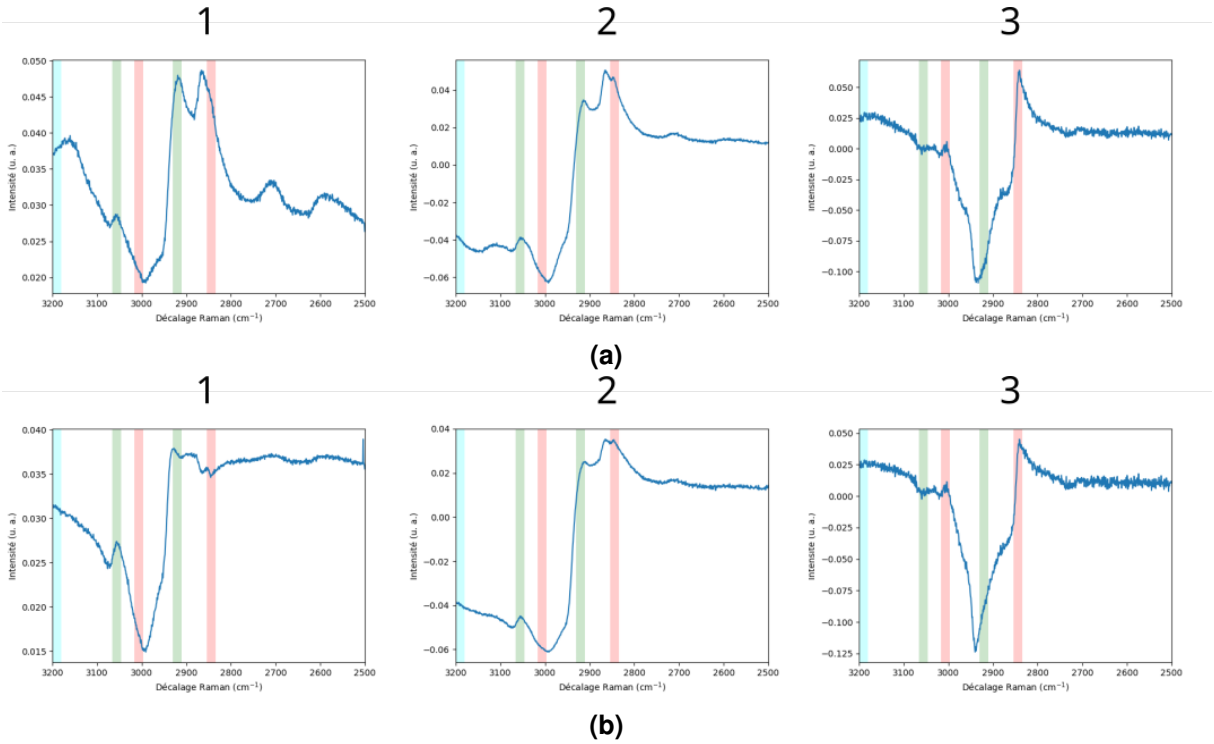


FIGURE 2.3 – Les trois principaux axes calculés par l'ACP, (a) ACP appliquée aux données centrées, (b) ACP appliquée aux données centrées et réduites.

pixel avec une forte intensité positive sur une dimension implique une présence des variables avec des valeurs positives dans le vecteur propre et une absence pour les variables ayant des valeurs négatives. En suivant le même raisonnement, une intensité négative sur ce même axe implique une absence des variables positives et une forte présence pour les variables négatives. Par exemple, les pixels jaunes de la dimension 2 de la figure 2.2 portent de l'information sur les décalages Raman 2910 et 2850 cm^{-1} mais ne portent pas de signal à 3000 cm^{-1}. A l'inverse pour les pixels en bleu sombre qui vont présenter du signal à 3000 cm^{-1} mais pas à 2910 et 2850 cm^{-1}.

2.1.1.3 Limite de l'ACP

Deux principales limites peuvent être associées à la méthode ACP.

Premièrement, la construction de la nouvelle base pour maximiser la décorrélation des données fait apparaître des valeurs négatives à la fois dans la projection et dans les axes de la base de dimension réduite. La finalité des méthodes développées est l'analyse des résultats par des biologistes et physiciens qui ont l'habitude de travailler avec des concentrations de molécules et des spectres CARS qui sont tous deux positifs. Les résultats de l'ACP ne sont donc pas directement exploitables par les experts dans

leur manière de travailler et exploiter les données.

La seconde limite est l'absence de contrainte spatiale pour construire la décomposition. Une cartographie impliquant une cohérence spatiale, il est raisonnable de présumer qu'une méthode prenant en compte l'information spatiale peut obtenir une information plus précise des données et ainsi améliorer la qualité de la calculée ainsi que de la projection des données sur celle-ci.

Afin de calculer une projection s'adaptant mieux aux données, il est possible de s'orienter vers d'autres types d'ACP non linéaires. Parmi elles, la méthode isomap utilise des graphes pour appliquer une SVD à partir d'une corrélation non linéaire.

2.1.2 Isomap

La méthode isomap, créée en 2000 par TENENBAUM *et al.* peut être considérée comme une ACP non linéaire [50]. Elle se base aussi sur une SVD mais, au lieu de l'appliquer à la matrice de variance-covariance des données, la décomposition est calculée à partir d'une matrice de distance sur graphe. Cette spécificité permet d'obtenir une base s'adaptant mieux aux données.

2.1.2.1 Introduction aux graphes

Un graphe $\mathfrak{G}(\mathfrak{V}, \mathfrak{E})$ est un objet mathématique représentant des données à partir de nœuds $\mathfrak{V}$ et d'arêtes $\mathfrak{E}$ les reliant. Les nœuds représentent une information et les arrêtes un lien entre deux informations. Les graphes sont souvent représentés par leur matrice d'adjacence $\mathcal{A} \in \mathbb{B}^{M \times M}, \mathbb{B} = \{0, 1\}$, avec $M = |\mathfrak{V}|$ le nombre de nœuds, le nombre de pixels dans notre contexte applicatif.

Dans certaines applications, la matrice d'adjacence peut être remplacée ou complétée par d'autres matrices comportant de nouvelles informations. C'est le cas dans la méthode isomap qui utilise une matrice de distance qui associe aux différentes arêtes une distance. Plus cette distance est grande, plus les nœuds représentent des informations différentes.

2.1.2.2 Calcul du graphe dans la méthode isomap

Le calcul du graphe se fait deux étapes : d'abord par un calcul du voisinage, puis par une construction du graphe à partir de la distance géodésique.

Le calcul du voisinage peut se faire de deux manières. La première consiste en l'utilisation d'un algorithme type plus proches voisins [51]. La seconde inclut dans

le voisinage tous les éléments inférieurs à une certaine distance. La construction du graphe repose sur un calcul de la distance géodésique entre les différents points. La distance géodésique correspond à une distance qui suit la géométrie de l'espace des données. Par exemple, la distance à la surface terrestre entre deux points, le pied et le sommet d'une colline, est une distance géodésique. Dans le cas de la méthode isomap, cette distance est souvent calculée en utilisant l'algorithme de Dijkstra [52]. Cet algorithme crée la matrice de poids représentant le graphe en créant des arêtes de poids égales au plus court chemin entre un point et un autre en partant du voisinage du point de départ. Le poids entre deux nœuds correspond ainsi à la somme des distances euclidiennes portées par les différentes arêtes composant le plus court chemin. Le choix de la distance euclidienne comme distance du voisinage est discuté en section 2.1.2.4.

2.1.2.3 Application de la SVD

La SVD n'est pas directement appliquée à la matrice des distances. La matrice est d'abord mise au carré puis doublement centrée :

$$\begin{aligned} \mathcal{C} &= -\frac{1}{2}HSH \\ H &= I_M - \frac{1}{M}(1_M 1_M^T) \\ S &= D_{ij}^2, \end{aligned} \tag{2.4}$$

avec M le nombre de données, pixels dans le cas d'images, I_M une matrice identité de tailles $M \times M$ et 1_M un vecteur de taille M et composé seulement de 1. La matrice H est appelée la matrice de centrage qui permet de centrer les distances et ainsi pouvoir appliquer la SVD à C.

Une fois les valeurs et vecteurs propres calculés, ils peuvent être triés par ordre décroissant. Les valeurs dans l'espace de dimension réduite sont alors déterminées en utilisant les valeurs et vecteurs propres :

$$Y = V\sqrt{\Lambda} \tag{2.5}$$

Pour sélectionner le nombre de dimensions dans l'espace réduit, il est toujours possible d'utiliser la méthode du « coude » sur les valeurs propres. La figure 2.4 présente les valeurs propres calculées par la SVD, à l'instar de l'ACP, trois axes est un choix approprié pour ce jeu de données.

Contrairement à l'ACP, il n'est pas possible d'obtenir des composants et donc

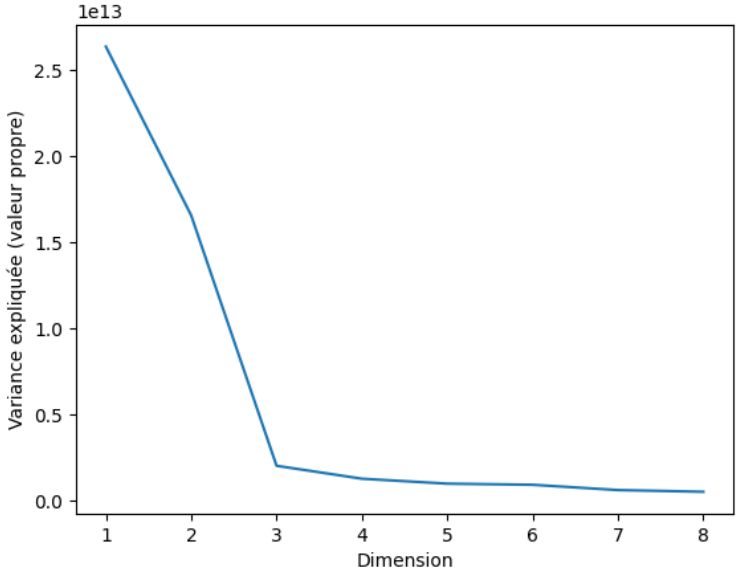


FIGURE 2.4 – Valeurs propres de la SVD

d'associer une dimension aux variables d'origine. En effet, les coordonnées dans la nouvelle base sont obtenues à partir des vecteurs et valeurs propres sans opération de projection. De plus, la SVD étant calculée à partir de la matrice de distance du graphe de taille $M \times M$, toute information sur les variables d'origines, les canaux spectraux dans notre cas, est perdue lors du calcul de distance.

2.1.2.4 Choix de la métrique de distance du calcul de voisinage

Originellement, la distance euclidienne est utilisée comme métrique de distance pour calculer le voisinage. Ce n'est cependant pas une obligation et d'autres types de distance peuvent être utilisés. Deux distances ont été comparées : la distance euclidienne et la distance d'angle spectral, *spectral angle distance* (SAD) en anglais [53]. Cette dernière se définit par l'angle entre 0 et π formé entre deux vecteurs :

$$SAD(x, y) = \text{acos}\left(\frac{x \cdot y}{\|x\|_2 \|y\|_2}\right), \tag{2.6}$$

avec x et y les deux vecteurs comparés et $x \cdot y$ le produit scalaire. Contrairement à la distance euclidienne, la SAD compare l'allure des spectres et est invariante au décalage. L'invariance au décalage signifie que si deux spectres ont exactement la même allure mais avec un décalage d'intensité, alors la SAD est tout de même nulle. Le décalage d'intensité des spectres pour un même composant étant un phénomène courant en

spectroscopie, la SAD est une distance tout indiquée pour comparer des spectres.

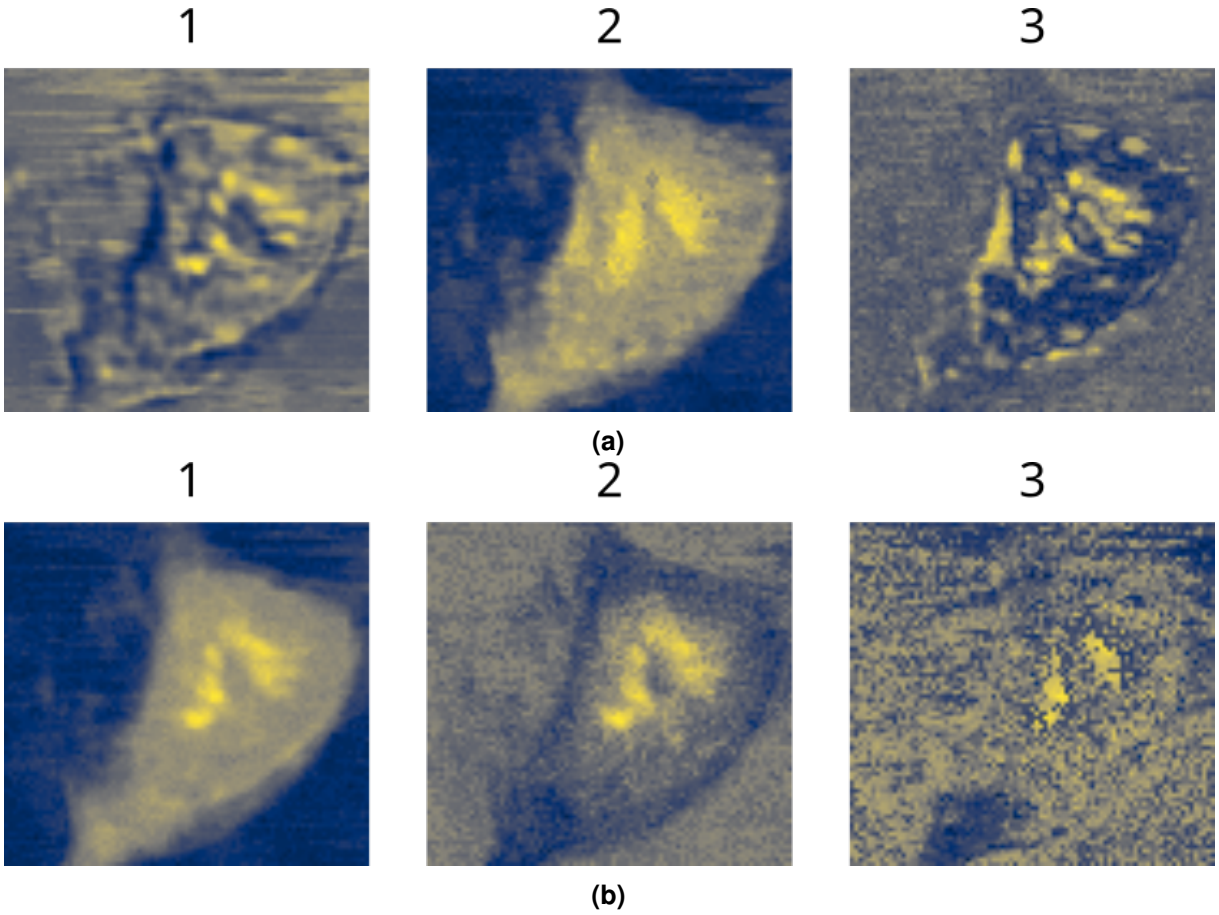


FIGURE 2.5 – Projection des données sur les trois principaux axes calculés par la méthode isomap, (a) la distance utilisée est la distance euclidienne, (b) la distance utilisée est la SAD.

Dans la figure 2.5, les résultats obtenus par l'application de l'algorithme isomap avec la distance euclidienne sont comparés avec ceux obtenus avec la SAD. Le jeu de données utilisé est toujours la cellule de référence de la section 1.4.2.1. Contrairement à l'intuition initiale, l'utilisation de la SAD n'améliore pas les résultats, au contraire, ils se détériorent. En effet, la dimension 2 met en avant une zone qui correspond à l'emplacement des nucléoles mais illumine aussi de l'environnement. Pour la dimension 3, le bruit est encore plus important à l'exception du bas de la cellule qui ne comporte aucune valeur positive. Les plus fortes valeurs de chaque dimension sont situées dans la même zone du noyau. La raison pour laquelle cette zone est autant mise en avant par la méthode n'a pas pu être identifiée.

En ce qui concerne les résultats obtenus avec la distance euclidienne, les deux premiers axes sont très similaires à ceux obtenus avec l'ACP. Le troisième diffère de celui obtenu par l'ACP mais ne montre pas une réalité biologique évidente. En conclusion,

contrairement à l'intuition initiale, la méthode isomap ne permet pas d'améliorer les résultats en ce qui concerne la projection dans un sous-espace pour mettre en avant des molécules, organites ou compartiments membranaires d'une cellule.

2.1.2.5 Limites de la méthode

Trois problèmes ressortent de la méthode. Tout d'abord, l'aspect spatial n'est toujours pas inclus dans la méthode. Cela peut néanmoins être corrigé par l'utilisation d'une métrique de distance utilisant aussi l'information spatiale. Ce type de distance n'a pas été expérimenté sur isomap en raison des autres limites de la méthode.

Deuxièmement, le choix de la métrique de distance influe fortement le résultat. La méthode introduit donc de l'hyperparamétrage pour le choix de la métrique utilisée. Il en est de même pour la sélection du nombre de voisins ou de la taille du noyau d'inclusion du voisinage.

La dernière limite est la plus importante. En raison de la perte de l'information spectrale lors du calcul des distances et des coordonnées finales qui ne sont pas issues d'une projection des données initiales, il n'est pas possible d'avoir le lien entre les variables de l'espace d'origine avec celles de l'espace de dimension réduite. Il n'est alors pas possible d'associer aux images obtenues un spectre et donc de renforcer l'analyse des images par analyse de spectres caractéristiques.

Nous décidons donc de nous tourner vers une seconde classe de méthodes basées sur la minimisation d'une erreur de reconstruction. Dans le domaine de la chimiométrie, l'analyse de données appliquée à la chimie, une famille de méthodes dont le rôle est de trouver les composants principaux d'un jeu de données et quantifier leur présence au sein des données est couramment utilisée. Cette famille de méthodes est appelée résolution de courbes multivariées et permet de contraindre la solution pour qu'elle soit conforme avec le phénomène étudié.

2.2 Résolution de courbes multivariées par moindres carrés alternés

2.2.1 La résolution de courbes multivariées

La résolution de courbes multivariées, *multivariate curve resolution* (MCR) en anglais [54], aussi appelée *self-modeling curve resolution* [55], est une famille venant de la chimiométrie. Cette famille regroupe les méthodes dont l'objectif est de trouver

les composants principaux d'un jeu de données en en déterminant les spectres caractéristiques et leurs concentrations dans chaque spectre d'un jeu de données. Ce jeu de données peut être une image, mais aussi un volume ou encore des élutions d'une réaction chimique. Une famille de méthodes à l'objectif similaire a été développée indépendamment en imagerie hyperspectrale géologique sous le nom de démélange, *unmixing* en anglais [56].

Contrairement aux méthodes précédentes, ces méthodes ne se basent pas sur la construction d'une base de projection maximisant la décorrélation des données, mais plutôt sur la construction d'un dictionnaire à partir d'une minimisation d'erreur de reconstruction et de contraintes appliquées au résultat. Ce dictionnaire permet alors de projeter les données pour obtenir les concentrations des différents composants formant le dictionnaire en chaque spectre acquis.

Le modèle MCR a comme formulation générale :

$$D = \Phi(C, S), \tag{2.7}$$

avec D les données acquises, C les concentrations des composants principaux, S leurs spectres caractéristiques, souvent appelés spectres « purs » dans la littérature, et Φ une fonction qui reconstruit D à partir de C et S. Cependant, le modèle est plus généralement traité sous sa forme linéaire :

$$D = CS^T + \mathcal{E}, \tag{2.8}$$

avec $D \in \mathbb{R}^{M\times N}, C \in \mathbb{R}^{M\times K}, S \in \mathbb{R}^{N\times K}$, et $\mathcal{E} \in \mathbb{R}^{M\times N}$ l'erreur d'approximation.

Dans ce contexte, de nombreuses solutions sont valides mathématiquement bien qu'une seule ne soit physiquement vraie pour un jeu de données. Trois types de transformations permettent de passer d'une solution à une autre :

— l'ambiguïté de permutation désigne les permutations qui peuvent être appliquées aux matrices C et S pour obtenir le même résultat. Cette ambiguïté ne gênant pas l'analyse des résultats, elle est rarement traitée ;

— l'ambiguïté d'intensité correspond à la multiplication par un facteur d'une des matrices qui ne change pas le résultat si son inverse est utilisé comme facteur de la seconde matrice : $D = (kC)\left(\frac{1}{k}S\right)^T + \mathcal{E}$ avec $l \in \mathbb{R}$. Cette ambiguïté est traitée par la normalisation d'une des deux matrices ou par l'utilisation de spectres déjà connus ;

— l'ambiguïté de rotation désigne l'ambiguïté résultant de l'application d'une matrice

de rotation R à C et S : $D = (CR)(R^{-1}S^T) + \mathcal{E}$. Cette ambiguïté est la plus complexe à traiter.

Pour limiter ces ambiguïtés, une contrainte de non-négativité est systématiquement appliquée aux spectres pour pouvoir obtenir des spectres cohérents avec une intensité mesurée. La contrainte de non-négativité est aussi appliquée à la matrice de concentrations associée à une normalisation pour que les concentrations d'un pixel somment à 1. Cette normalisation permet d'obtenir des concentrations qui peuvent être comparées aisément les unes aux autres et quantifier la présence de chaque composant dans chaque pixel.

La méthode la plus utilisée en chimiométrie pour résoudre la MCR est la régression par moindres carrés alternés qui, comme son nom l'indique, applique alternativement une régression par moindres carrés pour calculer C et S.

2.2.2 Résolution par moindres carrés alternés

La résolution de courbes multivariées par moindres carrés alternés, MCR-ALS en anglais [54], détermine C et S à partir d'une estimation initiale d'une des deux matrices et de quatre étapes répétées jusqu'à la convergence :

1. Optimisation par moindres carrés de C ;
2. application des contraintes aux nouvelles valeurs de C ;
3. optimisation par moindres carrés de S ;
4. application des contraintes aux nouvelles valeurs de S.

À la place de l'algorithme des moindres carrés ordinaires, sa variante non-négative, *non-negative least squares* (NNLS) en anglais [57], peut être utilisée pour intégrer la contrainte de non-négativité directement dans la régression. Le fonctionnement de la méthode avec une régression par NNLS et une contrainte de normalisation appliquée à C peut être résumé à l'algorithme 2.1.

Pour fonctionner, la MCR-ALS nécessite d'avoir une première estimation d'une des deux matrices. Le plus souvent, la matrice des spectres est sélectionnée pour l'estimation initiale. Cette première estimation influe fortement les résultats, il est donc primordial de bien choisir la méthode d'initialisation.

2.2.3 Initialisation

Deux méthodes d'initialisation sont comparées pour décider laquelle est la plus à même de fournir une première estimation de la matrice S pour débuter l'algorithme

Algorithme 2.1 : Algorithme de la MCR-ALS avec NNLS et contrainte de normalisation appliquée à C.

Données : $D \in \mathbb{R}^{M \times N}, S_0 \in \mathbb{R}^{N \times K}$
Résultat : $C \in \mathbb{R}^{M \times K}, N \in \mathbb{R}^{N \times K}$
tant que *non Converge*(CS^T) **faire**
  $C \leftarrow$ NNLS(S, D^T);
  $C \leftarrow$ Normalisation(C);
  $S \leftarrow$ NNLS(C, D);
fin

de MCR-ALS : l'analyse de mélange par auto-modélisation interactive simple d'utilisation, *simple-to-use interactive self-modelling mixture analysis* (SIMPLISMA) en anglais [58], et l'analyse des composantes des sommets, *vertex component analysis* (VCA) en anglais [59].

2.2.3.1 SIMPLISMA

L'analyse de mélange par auto-modélisation interactive simple d'utilisation, *simple-to-use interactive self-modelling mixture analysis* (SIMPLISMA) en anglais, fut développée par WINDIG *et al.* en [58] pour trouver les composants principaux d'un jeu de données composé de spectres. Cette méthode est la plus utilisée pour initialiser la MCR-ALS et fait office de méthode de référence en chimiométrie. Elle utilise une approche statistique pour estimer les spectres en sélectionnant ceux avec le plus grand rapport écart-type sur moyenne. Ces spectres sont alors appelés spectres les plus « purs ».

L'algorithme peut être décrit de la manière suivante. Tout d'abord, la « pureté » de chaque spectre est estimé :

$$p_i = \frac{\sigma_i}{\mu_i}, \tag{2.9}$$

avec p l'indice de « pureté », σ l'écart-type et μ la moyenne de chaque spectre. Le spectre le plus pur est celui avec la valeur de p la plus grande. Pour sélectionner les autres spectres, il faut tenir compte de l'information déjà expliquée et sélectionner le spectre le plus indépendant de ceux déjà calculés. Ce problème est résolu en corrigeant les indices de « pureté » à l'aide d'une matrice de corrélation autour de l'origine $\mathcal{O}$:

$$O = \frac{1}{N} \breve{D}^T \breve{D}, \tag{2.10}$$

avec N le nombre de canaux spectraux et $\breve{D} \in \mathbb{R}^{M \times N}$, les données normalisées

telles que :

$$\breve{d}_{i,j} = \frac{\sqrt{\mu_i^2 + (\sigma_j + \varkappa)^2}}{d_{i,j}}, \tag{2.11}$$

où $\varkappa$ est un facteur déterminé par l'utilisateur selon l'intensité de la moyenne des spectres. La matrice O est ensuite utilisée pour calculer des poids $\mathfrak{W}^k$ quantifiant l'impact des spectres déjà sélectionnés :

$$\mathfrak{w}_i^k = \begin{pmatrix} o_{i,i} & o_{i,s_1} & \dots & o_{i,s_{k-1}} \\ o_{s_1,i} & o_{s_1,s_1} & \dots & o_{s_1,s_{k-1}} \\ \dots & \dots & \dots & \dots \\ o_{s_{k-1},i} & o_{s_{k-1},s_1} & \dots & o_{s_{k-1},s_{k-1}} \end{pmatrix}, \tag{2.12}$$

avec s_i le i-ième spectre le plus « pur » et k l'indice du spectre à extraire. Le déterminant de cette matrice est ensuite utilisé pour mettre à jour p :

$$p_i^k = p_i |w_i^k|. \tag{2.13}$$

Le k-ième spectre le plus pur est désormais le spectre avec la plus grande valeur p_i^k.

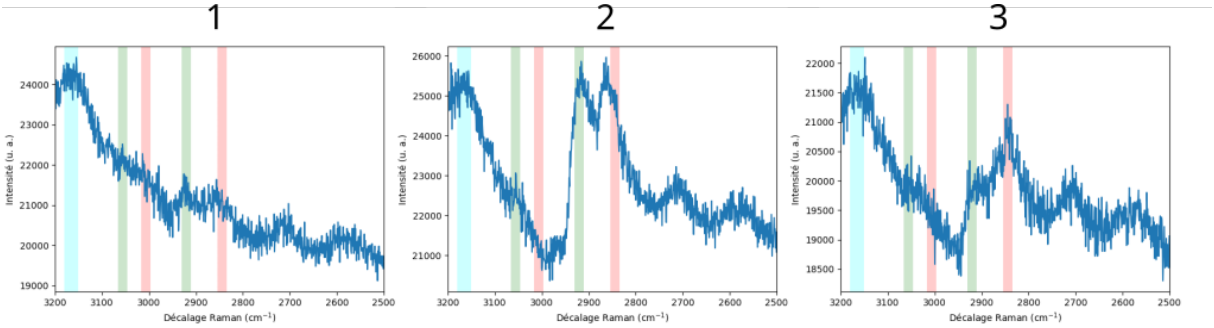


FIGURE 2.6 – Les trois spectres les plus « purs » extraits par la méthode SIMPLISMA sur une cellule HEK-293 fixée en interphase.

Dans la figure 2.6, la méthode SIMPLISMA est appliquée à la cellule de référence de la section 1.4.2.1. Ces spectres présentent bien trois spectres à l'allure différente qui peuvent être de bons candidats pour initialiser la régression. Comme la méthode sélectionne les spectres sans les modifier, ils sont aussi bruités que dans le jeu de données.

2.2.3.2 VCA

L'analyse des composantes des sommets, *vertex component analysis* (VCA) en anglais [59] est une méthode développée en imagerie hyperspectrale, *hyperspectral imaging* (HSI) en anglais, pour trouver les spectres des composants principaux d'un jeu de données. Elle peut aussi être utilisée comme méthode d'initialisation pour des méthodes de MCR. Contrairement à SIMPLISMA, la méthode définit un simplexe, une généralisation du triangle, englobant les données.

La méthode démarre à partir de la formulation vectorielle de la MCR :

$$d = S\gamma c + e, \tag{2.14}$$

où d est un spectre acquis avec D, S la matrice des spectres caractéristiques, γ un coefficient représentant la variabilité d'illumination, c la concentration de chaque composant et e une erreur additive. Deux hypothèses sont émises :

— les concentrations c forment un simplexe et, par extension, Sc en forment un autre ;

— chaque colonne de S est orthogonale aux autres.

Les données peuvent alors être projetées sur un simplexe en utilisant les vecteurs propres de D. S est calculé en projetant les données dans une direction aléatoire mais orthogonale au sous-espace formé par les spectres caractéristiques déjà trouvés. Le k-ième spectre signature est le spectre avec la plus grande valeur une fois projeté. Le premier spectre est trouvé en prenant un vecteur orthogonal à un vecteur pointant dans une seule direction.

Un point important de la VCA est la gestion du bruit. En effet, la méthode démarre par un débruitage des données qui diffère selon le ratio signal sur bruit, *signal-to-noise ratio* (SNR) en anglais, calculé. Si le SNR est inférieur à un certain seuil SNR_{th}, alors les données sont débruitées en utilisant l'ACP dans un nombre de dimensions $K-1$, K étant le nombre de composants recherchés. Si $\text{SNR} > \text{SNR}_{\text{th}}$ est supérieur au seuil, alors une SVD avec K valeurs singulières est appliquée directement aux données.

La figure 2.7 présente trois spectres calculés par la VCA à partir de la cellule de référence de la section 1.4.2.1. Contrairement à SIMPLISMA, les spectres sont assez peu bruités grâce à la SVD appliquée au début de l'algorithme. Il est cependant possible de reconnaître des signatures similaires entre les spectres 1 et 2 de la figure 2.6 et les spectres 1 et 3 de la figure 2.7. Le spectre 3 obtenu avec SIMPLISMA ne peut cependant pas être directement associé au spectre 2 obtenu avec la VCA.

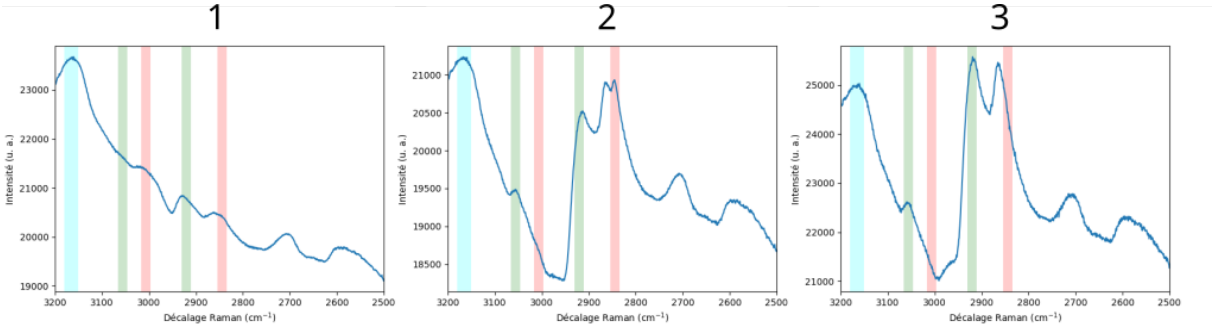


FIGURE 2.7 – Les trois spectres calculés par la méthode VCA sur une cellule HEK-293 fixée en interphase.

La modification des résultats entraînée par le choix de la méthode d'initialisation est discutée en section 2.2.5.5.

En plus du choix de la méthode d'initialisation, la sélection du nombre de composants K recherchés est un point crucial pour appliquer la MCR. Un mauvais choix de valeur de K peut entraîner une perte d'information, voire conduire à une mauvaise analyse des résultats.

2.2.4 Sélection du nombre de composants recherchés

Il existe différentes manières de sélectionner le nombre de composants K recherchés dans l'échantillon. La première est la connaissance *a priori* du jeu de données, ce cas n'est que très peu rencontré dans le contexte d'utilisation des méthodes MCR.

La seconde, la plus souvent utilisée en chimiométrie, est de considérer que la taille du dictionnaire que peut calculer la MCR-ALS est la même que le nombre de composantes principales optimales par l'ACP. L'utilisateur applique alors une ACP au jeu de données et utilise la méthode du « coude » pour sélectionner K. Cependant, cette méthode présente la limite de considérer que la décomposition faite par la MCR-ALS sera similaire à celle de l'ACP. Or, les contraintes appliquées aux matrices impactent la manière dont l'information est séparée entre les différentes dimensions.

Une troisième approche est de définir une erreur de reconstruction et de répéter la méthode de MCR sur plusieurs valeurs de K. La courbe définie par les erreurs de reconstruction peut alors être analysée pour situer où se situe le point d'inflexion. Le choix final de K se fait par l'étude par des experts des solutions trouvées pour les valeurs de K autour du point d'inflexion.

Cette dernière méthode est plus longue puisqu'elle nécessite de répéter la méthode d'analyse, mais elle permet de ne pas supposer un lien entre les dimensions

calculées par une ACP et celle par une méthode de la famille des MCR qui n'existe pas toujours.

Afin d'évaluer la validité de cette approche de sélection de K, une étude a été menée sur les données de cellules présentées en section 1.4.2.1. La méthode MCR-ALS a été appliquée sur ces cellules pour un K allant de 1 à 14. Une fois les matrices C et S calculées, l'erreur $\mathcal{E} = D - CS^T$ est calculée, et la qualité de la reconstruction est évaluée en utilisant le manque de concordance, *lack of fit* (LOF) en anglais :

$$\text{LOF}(\mathcal{E}, D) = \frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}. \tag{2.15}$$

Le LOF moyen et son écart-type par dimension peuvent ensuite être comparés aux statistiques des valeurs propres obtenues par une ACP pour évaluer si le fait de répéter la méthode de MCR peut permettre d'obtenir une meilleure estimation de la valeur optimale de K.

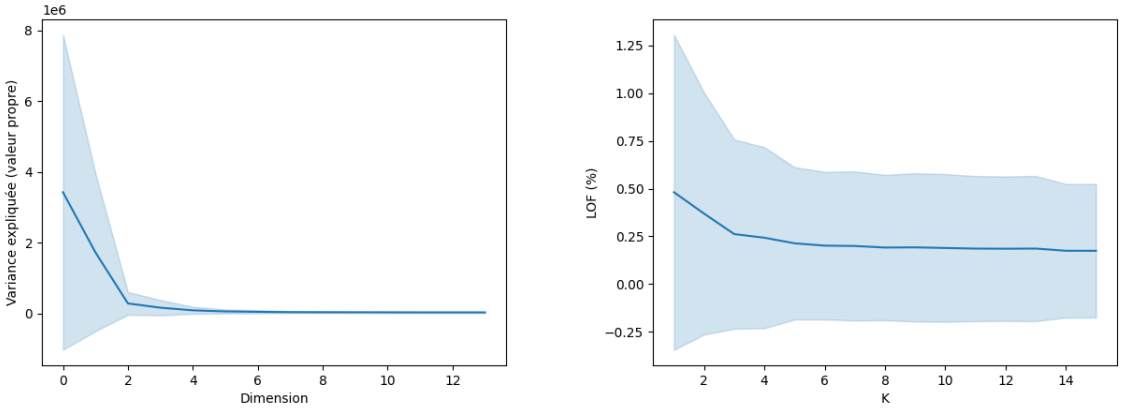


FIGURE 2.8 – Comparaison entre la sélection de K, (a) évolution des valeurs propres de l'ACP, (b) évolution du LOF. La courbe représente la valeur moyenne et l'aire autour de la courbe l'écart-type.

La figure 2.8 compare l'évolution moyenne des valeurs propres obtenues par ACP sur les données de cellules en figure 2.8a avec l'évolution du LOF en figure 2.8b. Les deux courbes suivent une évolution similaire mais un plus grand écart-type est présent sur la courbe du LOF. Cependant, là où la courbe des valeurs singulières converge autour de deux ou trois dimensions, la courbe du LOF descend moins vite et laisse supposer que de l'information utile est présente jusqu'à la dimension 5. Pour vérifier, cette supposition, les résultats obtenus pour une valeur de $K = 5$ vont

être maintenant présentés.

2.2.5 Application à des données CARS

2.2.5.1 Données cellulaires

Pour valider l'utilisation de la méthode MCR-ALS sur des données biologiques CARS, l'algorithme 2.1 a été appliqué à plusieurs cellules dans des états physiologiques différents. Les résultats ont tous été réalisés avec $K = 5$ et S initialisée avec la méthode SIMPLISMA [30]. Les résultats obtenus avec la cellule de référence sont détaillés dans cette section. Afin de mettre en avant le contraste des concentrations et dans l'optique d'une étude qualitative plus que quantitative, les concentrations calculées par la MCR-ALS sont présentées sous forme d'images en fausses couleurs.

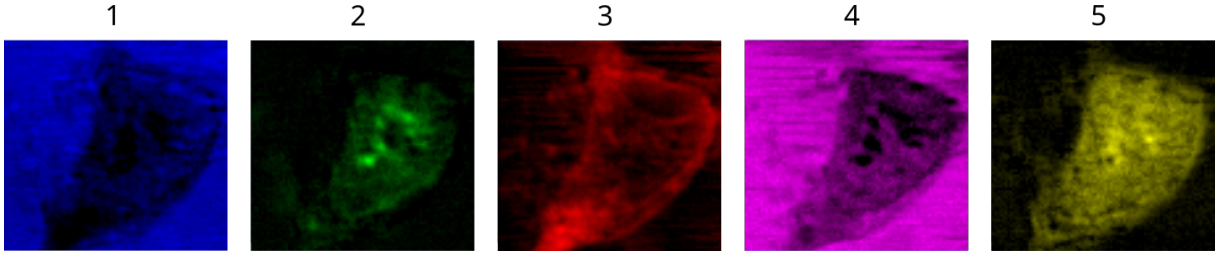


FIGURE 2.9 – Les 5 concentrations calculées par la MCR-ALS sur une cellule HEK-293 fixée en interphase.

Les concentrations obtenues en appliquant la MCR-ALS à une cellule HEK-293 fixée en interphase de la section 1.4.2.1 sont disponibles en figure 2.9 et les spectres en figure 2.10. Le LOF de cette cellule avec $K = 5$ est évalué à $\sim 0.017\%$, une valeur indiquant que la décomposition contient la majorité de l'information contenue dans le signal.

Les différentes intensités au sein des concentrations de la figure 2.9 n'atteignent jamais la valeur maximale possible indiquant que les différents pixels sont composés de plusieurs des composants, nom donné aux dimensions dans le contexte de la MCR, de manière non-négligeable, les axes définis par la matrice S ne décorrèlent pas complètement les données.

Les spectres de la figure 2.10 n'étant pas traités pour extraire $\mathrm{Im}[\chi^{(3)}]$, les pics n'ont pas l'allure d'une fonction lorentzienne et une ligne de base est présente et certains spectres comme celui du composant 5 sont bruités. Bien que rendant l'étude plus complexe, ces difficultés ne la rendent pas impossible.

L'analyse des spectres caractéristiques permet de faire différentes conclusions sur la composition chimique des différentes composants. Les spectres des composants 1 et

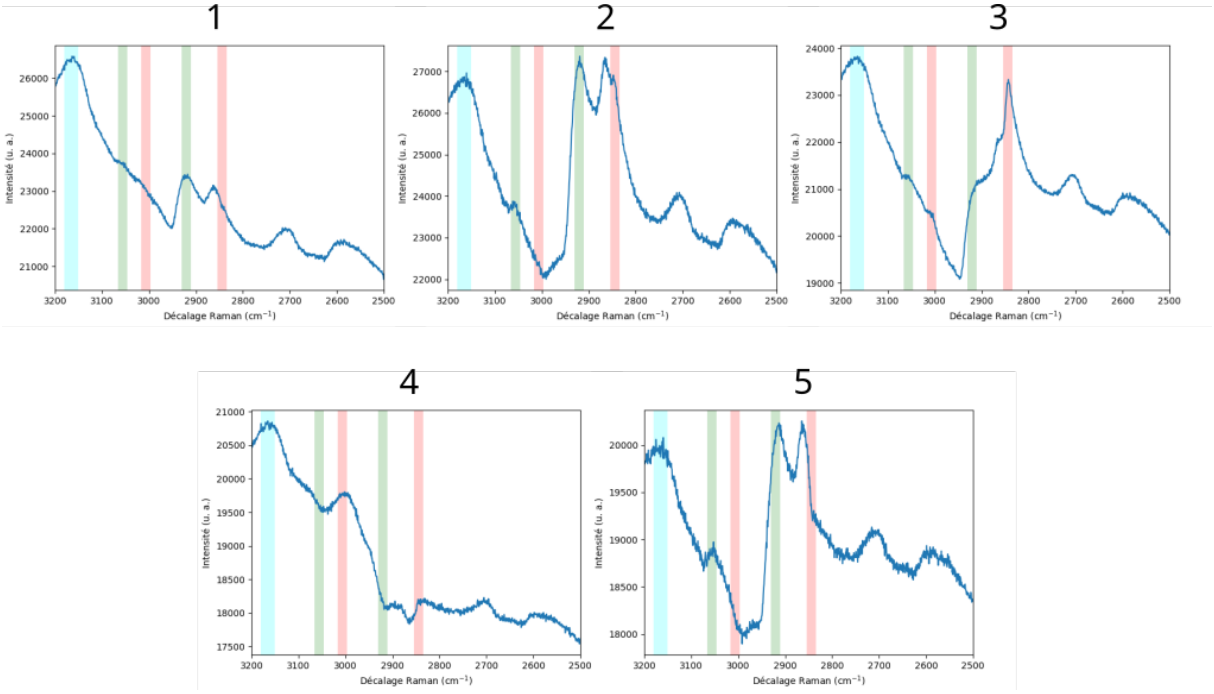


FIGURE 2.10 – Les 5 spectres calculées par la MCR-ALS sur une cellule HEK-293 fixée en interphase.

4 sont caractérisés par un fort signal aqueux couplé à des protéines pour le composant 1 et des lipides pour le composant 4. à l'environnement de la cellule qui est principalement aqueux. Les composants 2 et 5 sont des composants composés majoritairement de protéines, ils diffèrent cependant dans le signal associé aux lipides. Le signal lipidique est faible dans le cas du composant 2 et absent pour le composant 5. Le composant 3 est un composant avec un fort signal lipidique associé à des protéines.

En tenant compte de l'analyse des spectres pour analyser les concentrations obtenues, il est possible d'associer les différents composants à l'environnement, à des organites ou encore à des structures membranaires. Ils diffèrent cependant sur la composition de cet environnement. Le composant 1 met en évidence l'eau associée à des protéines alors que le composant 4 montre l'eau associée aux lipides. Le composant 2 peut être associé au noyau et aux nucléoles en raison de la localisation qui se superpose à la fluorescence DAPI de la figure 1.10a et de la signature spectrale forte en protéines. L'une des principales fonctions des nucléoles est d'être le site des étapes initiales de la biogenèse des ribosomes. Ce rôle implique la présence de protéines liées à la machinerie de transcription et des modifications post-transcriptionnelles par de petites ribonucléoprotéines nucléaires dans le composant fibrillaire dense. L'assemblage des ARNr avec les protéines ribosomiques se produit ensuite dans la composante

granulaire du nucléole [60]. L'association des nucléoles avec un spectre indiquant une forte présence de protéines est donc en accord avec une réalité biologique.

Le composant 3 correspond à la membrane plasmique, ses alentours et le cytoplasme, des zones riches en lipides mais aussi composées de protéines. De par la composition chimique induite par son spectre et la localisation spatiale des concentrations, nous pouvons associer le composant 3 aux phospholipides constituants la membrane cellulaire. Finalement, le composant 5 correspond à un signal fort sur les protéines mais une absence de lipides en adéquation avec les concentrations dans le noyau et les nucléoles. Contrairement aux précédents composants, il est difficile d'associer le composant à un compartiment cellulaire.

2.2.5.2 Données tissulaires

L'extensibilité de la méthode est évaluée en l'appliquant au jeu de données de tissu adipeux blanc de souris présenté en section 1.4.2.2. Ce jeu de données composé de 15 tranches de 201 par 201 pixels avec 1340 mesures par spectres est trop volumineux pour pouvoir être calculé en utilisant tout le jeu de données. En effet, la méthode SIMPLISMA requiert de construire de nombreuses matrices de plus en plus grandes pour pouvoir estimer les spectres initiaux qui demandent une trop importante quantité de mémoire vive (> 1To). L'algorithme est appliqué sur une seule tranche du jeu de données à la fois. Une différence importante dans la méthodologie de sélection de la valeur de K est à noter sur ce jeu de données. En effet, contrairement aux données de cellules où K est déterminé à l'aide de l'évolution de la métrique de reconstruction, elle est sélectionnée de manière exploratoire par des physiciens travaillant sur le phénomène CARS analysant les résultats obtenus pour différentes valeurs de K. Cette méthode bien plus subjective nécessite donc d'être pratiqué avec précaution par l'analyste pour limiter le risque de surinterprétation des résultats. Elle permet cependant d'aller observer les composants trouvés ne portant que peu d'informations pour la reconstruction afin de déterminer s'il ne représente tout de même pas une réalité physique et biologique. Suite à l'exploration des données, une valeur de $K = 15$ a été déterminée nécessaire pour obtenir le maximum d'information. Parmi les 15 composants calculés, seulement 3 ont été déterminées comme représentatifs de réels organites ou constituants organiques [61]. Ces composants sont les numéros 4, 6 et 14. L'ensemble des composants calculés sont disponibles en annexe A.4 mais seulement les trois composants retenus sont discutés dans cette section. Cette difficulté à trouver l'information utile dans le jeu de données met en évidence les limites de la méthode MCR-ALS.

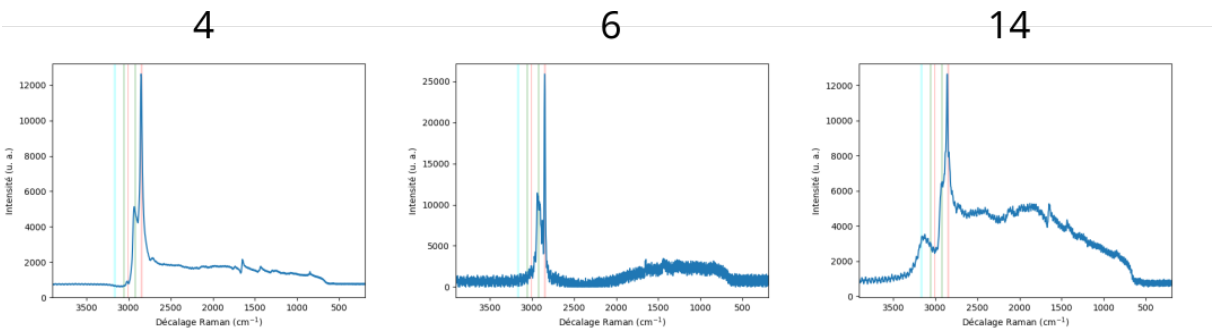


FIGURE 2.11 – Les spectres des composants 4, 6 et 14 calculés par la MCR-ALS à partir d'un tissu adipeux blanc de souris.

La figure 2.11 présente les spectres des trois composants. Ils partagent tous trois un intense signal lipidique à 2882 cm^{-1}. Les composants 4 et 6 comportent un signal lipidique et protéique alors que le composant 14 se distingue par un fort signal aqueux et un signal plus intense en dehors de la zone CH. Le composant 6 possède toutefois un pic de protéines plus important que le 4 laissant à penser qu'il s'agit d'un signal associé au noyau des cellules.

La figure 2.12 présente les concentrations des composants 4, 6 et 14. Les concentrations n'étant pas très bien visibles sur les images de la figure 2.12a, le contraste et la luminosité de celles-ci ont été modifiées pour obtenir la figure 2.12b. Il est alors possible d'associer la localisation des composants avec leur spectre, il est possible d'associer le composant 4 au cytoplasme des cellules, le composant 6 aux noyaux et le 14 à la MEC.

2.2.5.3 Validation de la non-utilisation de méthodes de recouvrement de phase

Pour rappel, l'intensité mesurée en CARS est le module au carré d'un nombre complexe composé d'une information vibrationnelle complexe et d'une information non-vibrationnelle réelle. Il est courant d'appliquer une méthode de recouvrement de phase pour reconstruire le nombre complexe et extraite seulement l'information vibrationnelle. Ces méthodes introduisant des erreurs pouvant faire disparaître de l'information vibrationnelle ou ayant des prérequis qui ne pouvaient pas être satisfait, il a été décidé d'appliquer la MCR-ALS directement aux spectres bruts. Ce choix est maintenant discuté par l'analyse des spectres de la matrice S traités par la MEM puis ajustés de leur lignes de base estimées à l'aide splines cubiques. Les spectres après traitement sont présentés en figure 2.13, les bandes des zones de vibrations de l'eau, des protéines et des lipides ont été décalées de 10 cm^{-1} afin de tenir compte du décalage créé lors du calcul de la partie résonnante de $\chi^{(3)}$. Les spectres corrigés

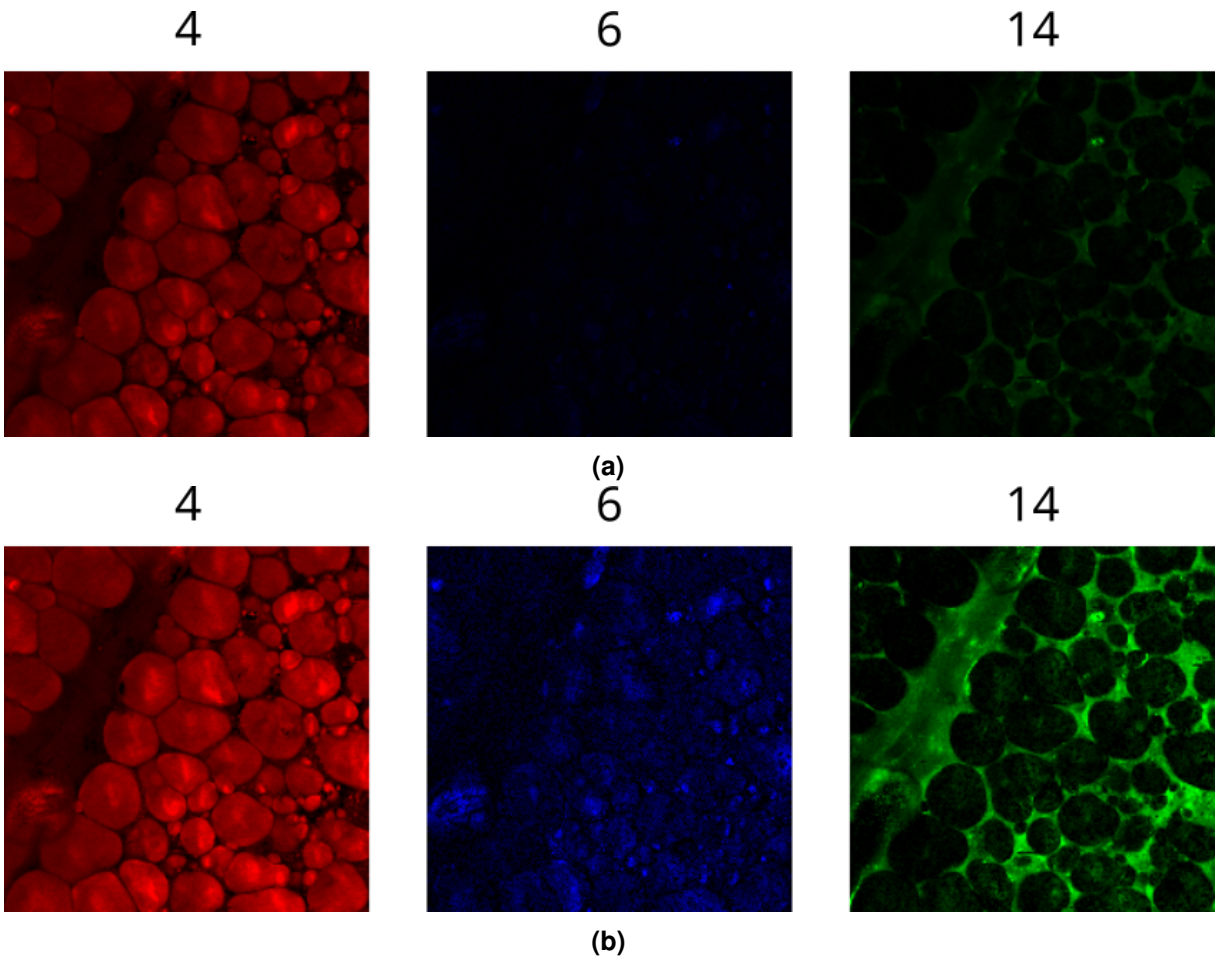


FIGURE 2.12 – Les concentrations des composants 4, 6 et 14 calculés par la MCR-ALS à partir d'un tissu adipeux blanc de souris. (a) Concentrations sans modification du contraste et de la luminosité. (b) Concentrations avec modification du contraste et de la luminosité.

valident l'interprétation des spectres CARS, les pics associés aux lipides sont bien dans les zones vibrationnelles des lipides et il en est de même pour les protéines et l'eau.

2.2.5.4 Utilisation de la matrice de spectres comme base de projection

Dans le but d'étudier l'impact de l'ajout de BDNF dans une cellule surexprimant TrkB, les cellules HEK-293 vivantes surexprimant TrkB présentées en paragraphes 1.4.2.1 et 1.4.2.1 sont analysées. Afin que les cellules soient dans des conditions similaires de développement et que l'effet de l'injection du BDNF soit plus visible, leur acquisition s'est effectuée 72h après le traitement. Dans le contexte d'étudier l'évolution d'une cellule induite par l'ajout de BDNF, nous décidons de ne pas appliquer la MCR-ALS aux deux cellules mais seulement à la cellule qui n'a pas été exposée au BDNF. La matrice S calculée peut alors servir de dictionnaire de spectres sur lequel projeter

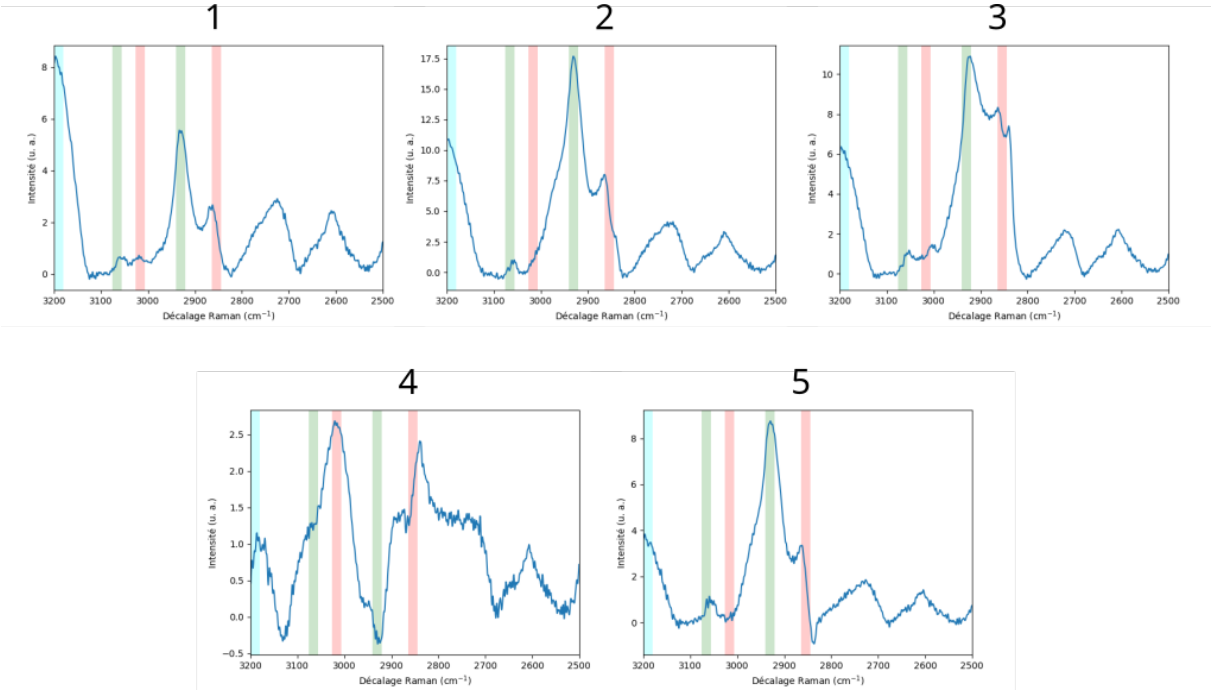


FIGURE 2.13 – Spectres de la figure 2.10 traités par la MEM et ajustés de leur ligne de base.

les données de la cellule avec BDNF pour obtenir des concentrations. Ce choix de projections plutôt que d'appliquer la MCR-ALS aux deux cellules permet d'obtenir les composants d'une cellule « saine » et voir leur évolution spatiale et en concentration sur une cellule avec BDNF. Pour vérifier qu'utiliser la matrice S de la cellule sans BDNF à celle avec BDNF est un choix acceptable, nous pouvons d'abord nous appuyer sur la biologie. Les deux cellules étant de physiologie très proche à l'exception de la présence ou non de BDNF, il est raisonnable de penser que les composants de la cellule sans BDNF soient aussi présents au sein de la cellule avec BDNF. Pour vérifier la validité numérique de l'approche, nous pourrons aussi nous appuyer sur le LOF obtenu par la décomposition en utilisant la matrice S comme dictionnaire de référence pour une projection. Pour effectuer la projection des spectres de la cellule avec BDNF sur la matrice S de la cellule sans BDNF, la régression linéaire NNLS suivie d'une normalisation de somme des pixels à 1 sont utilisées.

Les concentrations obtenues par la MCR-ALS appliquée à la cellule sans BDNF sont disponibles en figure 2.15a et les spectres en figure 2.14. Le résultat de la projection est disponible en figure 2.15b. Le premier composant correspond à un signal principalement composé d'eau avec des lipides. Contrairement à la cellule HEK-293 fixée, le signal est aussi présent dans le noyau, à l'exception des nucléoles, de la cellule et non seulement dans l'environnement extracellulaire. Le second composant présente

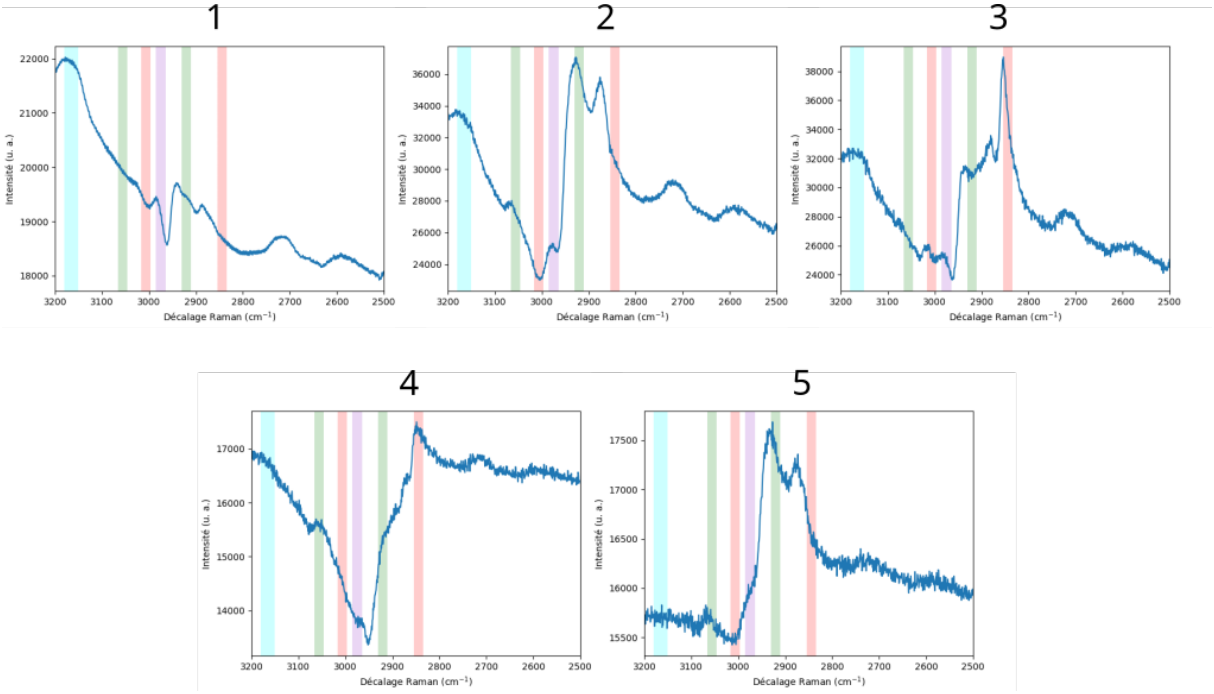


FIGURE 2.14 – Les 5 spectres calculés par la MCR-ALS sur une cellule HEK-293 vivante en interphase surexprimant TrkB.

la particularité d'émettre un signal à ~ 2970 cm^{-1} qui est attribué à l'ADN en plus d'un signal sur les protéines. Il illumine très précisément les nucléoles. En raison de son spectre portant un signal principalement lipidique et de sa localisation au sein du cytoplasme, le composant 3 est attribué aux gouttelettes lipidiques. Le composant 4 est associé à un signal lipidique présent à l'intérieur de toute la cellule, même au sein du noyau. Finalement, le composant 5 contient des protéines et montre une absence de signal aqueux. Il se localise dans le noyau et le cytoplasme des cellules avec un point intense au sein du noyau dans les deux cellules et un aspect réticulaire dans le cas de la cellule non traitée par TrkB.

Les différences entre les composants 1 et 2 obtenus avec la cellule HEK-293 fixée en interphase avec celle vivante modifiée pour surexprimer TrkB sont aussi observables sur une cellule HEK-293 vivante en interphase présentée en annexe A.3 et sont attribuables au processus de fixation. En effet, la fixation d'une cellule est susceptible de provoquer une perte de contenu cellulaire et induire une contraction de celle-ci. D'autres phénomènes d'altération possibles sont la réticulation des protéines ou des protéines-ADN [62] ou encore l'altération des propriétés mécaniques de la cellule [63]. La réticulation est le processus chimique induit dans les cellules par le paraformaldéhyde utilisé lors de la fixation. L'augmentation du nombre de gouttelettes

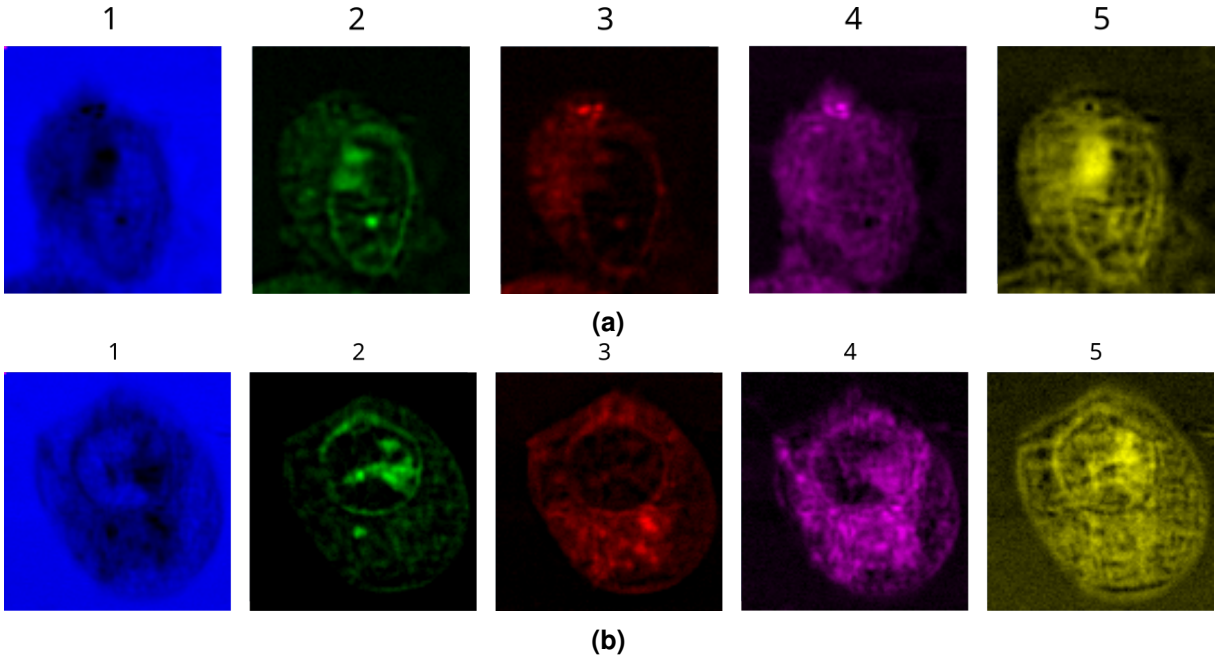


FIGURE 2.15 – Comparaison des concentrations obtenues sans et avec injection de BDNF, (a) concentrations calculées par la MCR-ALS sur une cellule HEK-293 vivante en interphase surexprimant TrkB, (b) projection sur une cellule HEK-293 vivante en interphase surexprimant TrkB avec ajout de BDNF.

lipidiques visible au sein de cellule traitée par BDNF par rapport à celle n'ayant pas reçu d'injection est en accord avec l'analyse effectuée par GUERENNE-DEL BEN *et al.* qui attribuent cette dissimilitude à une modification du métabolisme lipidique [4]. Il en est de même pour le composant 4 qui présente des concentrations de plus fortes intensités dans le cas de la cellule ayant reçu l'injection de BDNF. Pour finir, le composant 5 est aussi présent au niveau de la membrane plasmique pour la cellule avec BDNF injecté alors que le spectre est principalement protéique. Cette particularité est attribuée à l'exposition au facteur de croissance ainsi qu'à l'activation de TrkB par le BDNF. Une plus grande investigation serait nécessaire pour interpréter le contenu intracellulaire de ce composant.

2.2.5.5 Influence de la méthode d'initialisation

Pour présenter l'impact de l'initialisation sur les résultats obtenus, la méthode MCR-ALS a été appliquée sur la cellule de référence en remplaçant l'initialisation par SIMPLISMA par celle obtenue avec la méthode VCA.

Les concentrations obtenues sont présentées en figure 2.16 et les spectres en figure 2.17. Contrairement aux figures 2.9 et 2.10, un seul composant illustre l'envi-

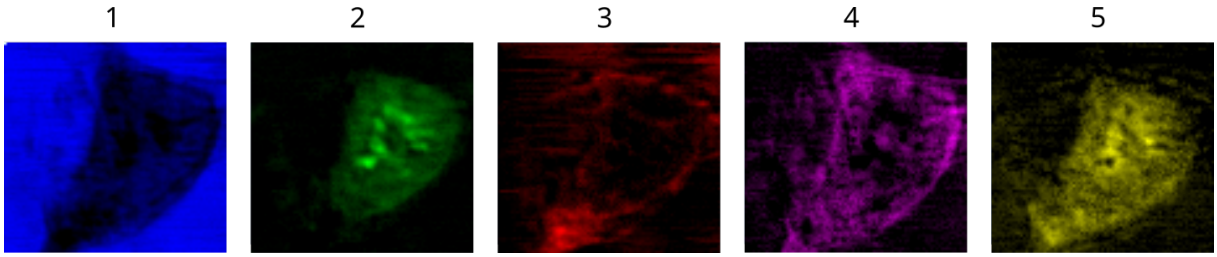


FIGURE 2.16 – Les 5 concentrations calculées par la MCR-ALS sur une cellule HEK-293 fixée en interphase en utilisant la méthode VCA en initialisation.

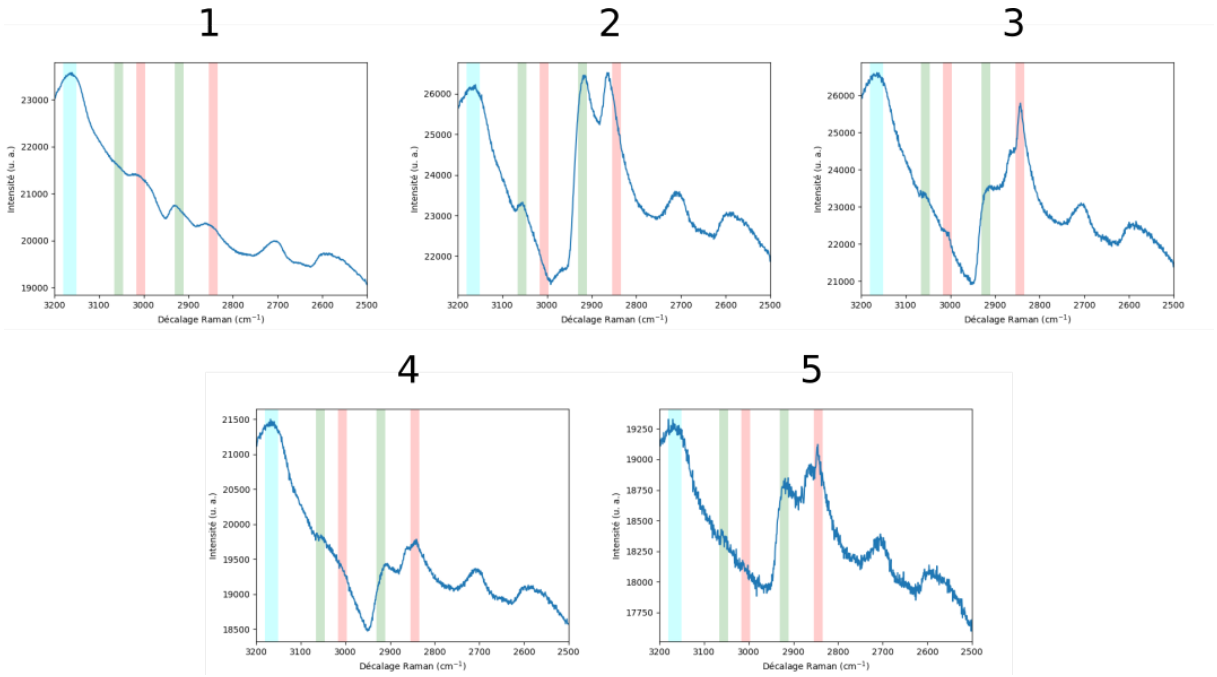


FIGURE 2.17 – Les 5 spectres calculées par la MCR-ALS sur une cellule HEK-293 fixée en interphase en utilisant la méthode VCA en initialisation.

ronnement. La signature spectrale de ce composant contient à la fois des protéines et des lipides. Les composants 2 et 3 sont assez similaires dans les deux résultats. Les composants 4 et 5 par contre diffèrent fortement. Le composant 4 contient principalement un signal lipidique mais possède aussi des protéines vibrant à 2844 cm^{-1}. Ce signal est principalement localisé dans le cytoplasme de la cellule. Le composant 5 présente un signal lipidique très intense rappelant le composant 5 des cellules modifiées pour surexprimer TrkB à la différence qu'un signal aqueux est aussi présent sur la cellule fixée en interphase. Le composant est principalement localisé en bas de la cellule dans son cytoplasme.

Le LOF obtenu est le même que celui de la décomposition en utilisant SIMPLISMA : $\sim 0.017\%$. La décomposition obtenue avec la VCA est donc tout aussi valable

numériquement dans sa capacité de reconstruction que celle avec SIMPLISMA.

Deux avantages sont à noter pour l'initialisation avec VCA. Premièrement, les composants étant initialisés orthogonalement, les composants calculés par la MCR-ALS sont bien plus décorrélés qu'avec l'initialisation par SIMPLISMA. Cette décorrélation permet d'obtenir des concentrations bien plus contrastées. Le second avantage est l'étape de débruitage qui permet d'obtenir des spectres plus lisses que ceux obtenus avec SIMPLISMA bien que la ligne de base et l'information non-résonnante soient toujours présentes. La méthode VCA semble donc une méthode d'initialisation plus efficace que la méthode SIMPLISMA.

La méthode MCR-ALS permet avec succès d'obtenir les composants principaux d'une cartographie de cellule. Cependant, cette méthode ne tient pas compte de l'information spatiale. De plus, une opération qui peut être effectuée lors de l'étude de cellules est sa segmentation afin de l'isoler de son environnement et mieux visualiser l'information en son sein. Intégrer la segmentation au sein de la MCR-ALS peut permettre de combiner les deux opérations et d'obtenir un résultat sur la cellule segmentée directement.

2.3 Segmentation d'image

La segmentation d'image au sens région regroupe les méthodes qui ont pour objectif de calculer des régions similaires au sein d'une image. Elle peut être interprétée comme une forme de classification d'une image où chaque pixel serait associé à une classe. La segmentation est le plus souvent binaire et consiste en l'extraction de zones d'intérêt du fond de l'image. Elle peut cependant aussi être faite avec plusieurs classes dans le contexte de la segmentation sémantique qui définit une classe par région et permet ainsi de segmenter des images sur plus que deux classes.

De nombreuses techniques de segmentation existent comme les méthodes de seuillage [64], K-moyennes [65], méthodes à contours actifs [66] et plus récemment les réseaux de neurones.

Dans cette section, l'intégration de la segmentation au sein de la MCR-ALS est discutée avec deux approches. Une segmentation appliquée sur les spectres CARS bruts à l'aide d'un réseau de neurones et une segmentation appliquée à la matrice de concentrations avec la méthode de segmentation de Chan-Sandberg-Vese.

2.3.1 Segmentation par réseau de neurones

Les approches par réseau de neurones sont devenues les approches de référence dans beaucoup d'opérations de l'analyse d'image, la segmentation ne fait pas exception.

2.3.1.1 Introduction aux réseaux de neurones

Un réseau de neurones est une structure composée de multiples unités appelées neurones. Un neurone est composé de paramètres qui sont combinés linéairement aux données d'entrées, combinaison à laquelle peut être ajouté un biais :

$$y = \sum_i x_i w_i + b, \tag{2.16}$$

avec y la sortie, x le vecteur d'entrée, w les paramètres du neurone et b le biais optionnel. Cette sortie est ensuite utilisée comme entrée à diverses fonctions non linéaires. La mise en parallèle de plusieurs neurones appelée couche dense et l'enchaînement de couches forment alors le réseau de neurones.

Une extension aux réseaux de neurones a été développée avec les réseaux de neurones convolutifs. Ces réseaux de neurones utilisent des couches de neurones particulières appelées couches convolutives. L'objectif de ces couches est le calcul de descripteurs des données. À la place d'une combinaison linéaire, les couches convolutives appliquent une convolution entre les données et leur paramètre, un biais optionnel est toujours possible à l'issue de la convolution.

$$y = \sum_i w_i \star x_i + b, \tag{2.17}$$

avec $\star$ l'opérateur de corrélation croisée, w_i est un filtre et x_i un signal sous forme de vecteur. Les descripteurs calculés vont, en général, du plus général sur les premières couches au plus précis sur les dernières couches. Dans les architectures de classification, ces descripteurs sont ensuite linéarisés sous forme de vecteurs pour utiliser des couches denses pour opérer la classification.

Les paramètres des différents neurones constituent des valeurs qui doivent être apprises. Pour l'apprentissage, une fonction à optimiser appelée fonction de coût est définie. Dans le cas de la classification et de la segmentation, la méthode faisant

référence est l'entropie croisée $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\sum_{i=0}^{P-1} v_i \log(y_i), \tag{2.18}$$

avec P classes, v un vecteur composé de 0 et d'un 1 pour la classe désirée et y le vecteur de sortie du réseau de neurones sous forme de probabilités. La sortie du réseau correspond donc aux probabilités que la donnée x appartienne à chaque classe. Ces probabilités sont en général calculées en utilisant la fonction *softmax* :

$$\textit{Softmax}(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}, \tag{2.19}$$

avec i l'indice de la classe normalisée.

Pour optimiser les paramètres du réseau par rapport à la fonction de coût, la méthode la plus courante est la descente de gradient par lots. Un des optimiseurs les plus utilisés est l'optimiseur ADAM [67] qui fait varier le taux d'apprentissage (pas de discrétisation de la descente de gradient) pour chaque paramètre en fonction de l'évolution de la fonction de coût.

Deux structures neuronales sont évaluées pour segmenter une image composée de spectres bruts CARS. La première est une approche pixel à pixel classifiant les spectres à partir d'un réseau de neurones dense. La deuxième est un réseau de neurones convolutif 1D.

2.3.1.2 Réseau de neurones dense

Le réseau de neurones dense implémenté est composé de 4 couches. Son architecture est décrite en tableau 2.1. Le réseau est composé de 4 couches denses et 635 913 paramètres à entraîner. Chaque couche est suivie par une normalisation par lots qui normalise les données par la moyenne et l'écart-type du lot. Cette opération a pour effet de lisser la fonction de coût et de limiter le risque que l'apprentissage bloque dans un minimum local [68]. Les couches intermédiaires, c'est-à-dire les couches autres que la dernière, sont toutes suivies d'une couche d'abandon. Ces couches servent à désactiver des neurones aléatoirement selon une probabilité $\rho = 0.5$, elles ne sont actives que durant l'entraînement et servent à limiter le risque de surapprentissage. Le surapprentissage correspond au risque que les paramètres des neurones s'ajustent trop par rapport aux données d'apprentissage et échouent à traiter des données inconnues. Les fonctions d'activation utilisées sont la fonction unité linéaire rectifiée, *rectified*

linear unit (ReLU) en anglais, pour les couches intermédiaires et *softmax* pour la dernière. La fonction ReLU est une fonction linéaire dans les positifs et qui renvoie toujours 0 dans les négatifs.

Couche	Norm. lots	Abandon	Activation
512	✓	0.5	ReLU
256	✓	0.5	ReLU
128	✓	0.5	ReLU
P	✓	0.0	*Softmax*

TABLEAU 2.1 – Architecture du réseau de neurones dense. Norm. lots signifie normalisation par lots, abandon indique la probabilité qu'un neurone soit désactivé.

2.3.1.3 Réseau de neurones convolutif 1D

Le réseau de neurones convolutif 1D implémenté est décrit en tableau 2.2. Il est composé de 2 couches convolutives et 3 couches denses pour un total de 1 364 422 paramètres à entraîner. Les couches convolutives utilisent toutes deux un filtre de taille 3 mais la première couche effectue un saut de 3 éléments dans le signal entre chaque convolution. Ce saut de 3 éléments est effectué pour réduire le nombre de paramètres dans la partie dense du réseau. Une normalisation par lots est appliquée après chaque couche. Les couches convolutives sont toutes suivies par une couche de sous-échantillonnage par la valeur maximale divisant par deux la taille du signal. Cette opération permet de réduire le risque de surapprentissage. Des couches d'abandon avec une probabilité de 0,5 sont mises après chaque couche dense. Tout comme le réseau de neurones dense, toutes les couches intermédiaires utilisent la fonction d'activation ReLU et la dernière utilise la fonction *softmax*.

La segmentation par réseau de neurones repose principalement sur un apprentissage supervisé. Cela signifie que le réseau est entraîné sur un jeu de données dont les classes des pixels sont connues. Il peut ensuite être utilisé sur des jeux de données inconnus. Il est est donc nécessaire de construire cette base d'apprentissage.

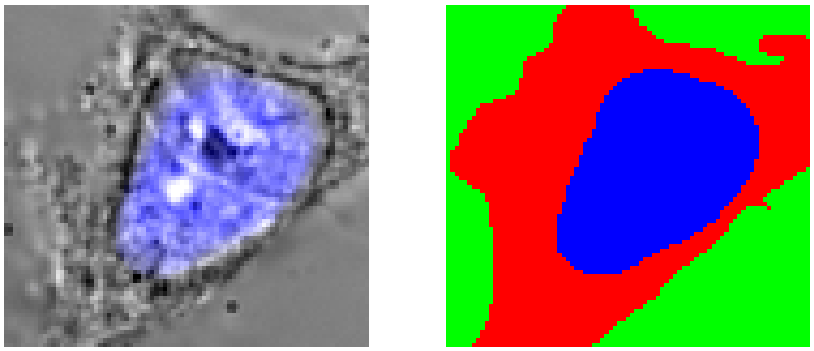
2.3.1.4 Construction de la base d'apprentissage

Pour construire la base de données, un expert biologiste a utilisé les images en lumière transmise et en fluorescence, quand disponibles, pour associer à chaque pixel une classe parmi trois : environnement, cytoplasme et noyau. Le résultat de la segmentation sur la cellule de référence est disponible en figure 2.18. La figure 2.18a montre les

Couche	Norm. lots	Abandon	Activation.
Conv(filtre : 3, saut : 3, descr. : 16)	✓	0.0	ReLU
MaxPool(2)			
Conv(filtre : 3, saut : 1, descr. : 32)	✓	0.0	ReLU
MaxPool(2)			
Lin(descr : 2400)			
Dense(descr. : 512)	✓	0.5	ReLU
Dense(descr : 256)	✓	0.5	ReLU
Dense(descr : P)	✓	0.0	*Softmax*

TABLEAU 2.2 – Architecture du réseau de neurones convolutif. Conv signifie convolutif, lin équivaut à linéarise, descr. descripteur, norm. lots normalisation par lots, abandon indique la probabilité qu'un neurone soit désactivé et MaxPool. sous-échantillonnage par valeur maximale.

images qui ont servi à la segmentation et la figure 2.18b le résultat de la segmentation.



(a) **(b)**

FIGURE 2.18 – Segmentation d'une cellule HEK-293 fixée en interphase, (a) image en lumière transmise avec la fluorescence DAPI du noyau en surimpression, (b) segmentation de la cellule. En vert l'environnement, en rouge le cytoplasme et en bleu le noyau.

Bien que construite par un expert, la base d'apprentissage comporte des erreurs liées aux moyens à disposition pour la construire. En effet, les images en lumière transmise ne capturent pas une tranche de la cellule comme la microscopie CARS mais l'intégralité de la cellule sur sa hauteur. Cela induit des différences au niveau de la morphologie cellulaire entre les deux images à la source d'erreur lors de la segmentation faite par l'expert. Dans le cas des cellules vivantes, les erreurs sont amplifiées par la différence temporelle entre le moment de l'acquisition CARS et l'acquisition de la lumière transmise.

2.3.1.5 Résultats

Les deux modèles ont été entraînés sur 50 itérations du jeu de données et des lots de 32 spectres. Le nombre de spectres pour chaque classe au sein du jeu de données est le même. Les spectres sont normalisés par la moyenne et l'écart-type de chaque décalage Raman. Le jeu de données est artificiellement augmenté avec une translation aléatoire de l'intensité. Le jeu de données d'entraînement est composé de 8374 spectres et le jeu de validation de 2791 spectres. L'optimiseur utilisé est ADAM avec la fonction de coût entropie croisée.

À l'issue de l'entraînement, le Dice score $\mathfrak{D}$ [69], [70], ou F1 score, a été calculé sur la segmentation de la cellule de référence. Le Dice score est une métrique d'évaluation d'un modèle de classification reposant sur la précision $\mathcal{P}$ et le rappel $\mathcal{R}$. La précision d'une classe i correspond au nombre de pixels bien classé dans cette classe par rapport au nombre de pixels qui ont été associés dans la classe :

$$\mathcal{P}_i = \frac{VP_i}{FP_i}, \tag{2.20}$$

avec VP vrai positif et FP faux positif. Elle permet de mesurer si le modèle surestime la classe. Le rappel d'une classe i correspond au nombre de documents bien classé par rapport au nombre réel de pixels appartenant à cette classe :

$$\mathcal{R}_i = \frac{VP_i}{VP_i + FN_i}, \tag{2.21}$$

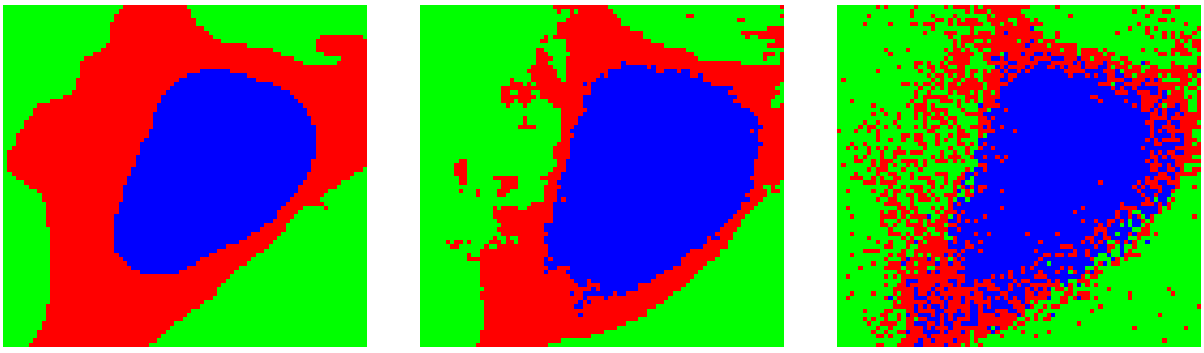
avec VP vrai positif et FN faux négatif. Cette valeur permet d'estimer si le modèle sous-estime la classe ou non. Le Dice score est alors estimé :

$$\mathfrak{D} = 2\frac{\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}}. \tag{2.22}$$

Une valeur proche de 1 indique une bonne capacité de classification du modèle et, à l'opposé, une valeur proche de 0 indique un modèle peu fiable. Appliqué sur la cellule de référence, le modèle dense obtient un score de ~ 0.71 et le modèle convolutif un score de ~ 0.74. Ce sont des scores corrects bien que présentant de nombreuses erreurs.

Pour pouvoir identifier où se situent ces erreurs, les segmentations obtenues sont présentées en figure 2.19. La figure 2.19a montre la segmentation faite par l'expert, la figure 2.19b le résultat obtenu avec le réseau de neurones dense et la figure 2.19c celui obtenu avec le réseau de neurones convolutif. Bien que le Dice score était légèrement meilleur sur le réseau convolutif, il donne un aspect plus « bruité » à la segmentation

qui la rend inutilisable dans un contexte de contrainte appliquée à la MCR-ALS qui va être étudié visuellement par un expert pour ses analyses. Le problème est bien moins présent avec le réseau dense mais ils sont toujours présents avec des pixels ou petits groupes de pixels isolés. Le problème est principalement présent dans les zones du cytoplasme qui sont difficiles à identifier visuellement ou qui pourraient être erronées dans l'ensemble d'apprentissage. Le phénomène provoquant de moins bons résultats avec le modèle convolutif est probablement un surapprentissage où le modèle a associé du bruit à un descripteur perturbant la classification.



(a) **(b)** **(c)**

FIGURE 2.19 – Segmentation d'une cellule HEK-293 fixée en interphase, (a) image en lumière transmise avec la fluorescence DAPI du noyau en surimpression, (b) segmentation de la cellule par réseau un réseau de neurones dense, (c) segmentation de la cellule par réseau un réseau de neurones convolutif. En vert l'environnement, en rouge le cytoplasme et en bleu le noyau.

À la lumière de ces résultats et des problèmes induits par le bruit des spectres et la difficulté de construire une base d'apprentissage, il est décidé que la méthode de segmentation est à appliquer sur la matrice de concentrations et doit être non-supervisée. Parmi les différentes méthodes existantes, les méthodes paramétriques présentent l'avantage de calculer des régions contiguës qui évite d'obtenir une segmentation avec des pixels isolés, phénomène qui ne devrait pas être présent dans de la segmentation de cellules. Pour cette raison, la méthode de Chan-Sandberg-Vese est retenue pour implémenter la segmentation dans la MCR-ALS.

2.3.2 Méthode de Chan-Sandberg-Vese

2.3.2.1 Le cas monovalué : Chan-Vese

La méthode de Chan-Vese (CV) [66] fait partie de la famille des méthodes à contours actifs. Développé en 1999 par CHAN *et al.*, la méthode repose sur le positionnement de contours $\mathfrak{C}$ qui représentent les frontières entre deux régions de l'image

correspondant aux deux classes de la segmentation. Cette méthode a déjà pu être utilisée avec succès dans la segmentation de cellules [71]-[73]. Pour une image en niveau de gris, la position optimale des contours peut être obtenue par la minimisation de la fonction suivante :

$$\begin{aligned} &\underset{c1,c2,\mathfrak{C}}{\arg\min}\, F(\mathfrak{c}_1, \mathfrak{c}_2, \mathfrak{C}) \\ F(c_1, c_2, \mathfrak{C}) &= \varrho_1 \int_{x,y \in interieur(\mathfrak{C})} |\mathcal{I}_{x,y} - \mathfrak{c}_1|^2 dxdy \\ &+ \varrho_2 \int_{x,y \in exterieur(\mathfrak{C})} |\mathcal{I}_{x,y} - \mathfrak{c}_2|^2 dxdy \\ &+ \upsilon \text{Longueur}(\mathfrak{C}) + \varpi \text{Aire}(interieur(\mathfrak{C})), \end{aligned} \tag{2.23}$$

avec $\mathcal{I}$ l'image à segmenter, $\mathfrak{c}_1$ et $\mathfrak{c}_2$ les intensités moyennes de chaque classe, $\varrho_1, \varrho_2, \upsilon$ et ϖ sont des paramètres à définir pour donner plus d'importance aux différentes parties de l'équation. Les deux premières parties de l'équation correspondent à l'homogénéisation de l'énergie au sein de chaque classe et les deux dernières à des contraintes de régularisation sur la longueur la courbe $\mathfrak{C}$ et sur l'aire à l'intérieur de celle-ci. Les coefficients de pondération $\varrho_1, \varrho_2, \upsilon$ et ϖ peuvent être réglés en suivant les règles suivantes :

— ϱ_1 et ϱ_2 sont à régler l'un par rapport à l'autre pour privilégier que l'intensité dans une classe soit plus homogène que dans l'autre. Si $\varrho_1 = \varrho_2 = 1$, aucune classe n'est privilégiée ;

— υ est la pondération de la contrainte de régularisation sur la longueur de $\mathfrak{C}$. Sa valeur varie généralement entre 0 et 1 et est le paramètre le plus important à régler [74]. Une faible valeur favorise une segmentation avec une très grande précision, pouvant entraîner une sur-segmentation, et une valeur importante favorise des limites assez lisses pouvant être à l'origine d'une sous-segmentation.

— ϖ pondère la contrainte appliquée à l'aire à l'intérieur de $\mathfrak{C}$. Ce paramètre peut être négatif ou positif selon l'*a priori* sur le contenu à segmenter. Une valeur négative va encourager une aire importante dans la segmentation et une valeur positive va la décourager.

Cette équation est résolue à l'aide d'une méthode de ligne de niveau [75] et d'une descente de gradient semi-implicite. $\mathfrak{C}$ est représentée par les points où la fonction de ligne de niveau est égale à 0. La résolution de l'équation introduit alors deux nouveaux paramètres :

— l'initialisation de la fonction de ligne de niveau, c'est-à-dire la position de la courbe

au début de l'algorithme ;

— le paramètre Δt utilisé lors de la discrétisation de la descente de gradient. Une faible valeur donne de meilleurs résultats mais entraîne une augmentation du temps de calcul et une valeur importante peut entraîner des problèmes de convergence.

2.3.2.2 Le cas multivalué : Chan-Sandberg-Vese

En 2000, CHAN *et al.* étendent la méthode aux images multivaluées. La méthode est alors appelée méthode de CSV. Cette extension se fait assez directement en remplaçant la distance de Manhattan par la distance euclidienne dans l'équation 2.23 :

$$\begin{aligned} &\arg\min_{c1,c2,\mathfrak{C}} F(c_1, c_2, \mathfrak{C}) \qquad (2.24)\\ F(c_1, c_2, \mathfrak{C}) &= \varrho_1 \int_{x,y \in inside(\mathfrak{C})} \|\mathcal{I}_{x,y} - \mathfrak{c}_1\|^2 dxdy \\ &+ \varrho_2 \int_{x,y \in outside(\mathfrak{C})} \|\mathcal{I}_{x,y} - \mathfrak{c}_2\|^2 dxdy \\ &+ \upsilon \text{Longueur}(\mathfrak{C}) + \varpi \text{Aire}(inside(\mathfrak{C})). \end{aligned}$$

Modifiée ainsi, la méthode de CSV peut être intégrée dans la MCR-ALS en tant que contrainte appliquée à la matrice de concentrations.

2.3.3 Intégration de la segmentation au sein de la MCR

Afin de réduire le temps de calcul et limiter l'impact du bruit présent dans les spectres du jeu de données, la méthode de segmentation est intégrée dans la méthode MCR-ALS en tant que contrainte appliquée à la matrice de concentration C. L'ajout de contraintes spatiales en tant que contrainte a déjà pu être abordé pour l'intégration de contrainte de régularisation par variation totale [77] ou de manière plus générale avec différents opérateurs dont la segmentation par seuillage appliquée à un composant déterminé préalablement [78]. L'approche choisie pour segmenter la cellule de l'environnement est d'utiliser une segmentation opérant sur tous les composants afin d'isoler la cellule de son environnement.

La segmentation est recalculée à chaque itération de la MCR-ALS et n'est pas appliquée lors de l'extraction des spectres signatures. Pour pouvoir l'appliquer, la matrice C est remodelée sous forme d'image afin de récupérer l'information spatiale. L'algorithme

MCR-ALS avec les contraintes appliquées est présenté dans l'algorithme 2.2.

Algorithme 2.2 : Algorithme de la MCR-ALS avec les contraintes de non-négativité, normalisation et de segmentation.

Données : $D \in \mathbb{R}^{M \times N}, S_0 \in \mathbb{R}^{N \times K}$
Résultat : $C \in \mathbb{R}^{M \times K}, N \in \mathbb{R}^{N \times K}$
tant que *non Converge*(CS^T) **faire**
  $C \leftarrow$ NNLS(S, D^T);
  $\mathcal{I} \leftarrow$ Remodèle$(C, \mathbb{R}^{H \times W \times K})$; // $H \times W = M$, C sous forme d'image.
  $\Theta \leftarrow$ Segmentation$(\mathcal{I})$; // Calcule un masque de segmentation.
  $\mathcal{I}_s \leftarrow \mathcal{I} \odot \Theta$; // Applique le masque. $\odot$ est le produit de Hadamard.
  $C \leftarrow$ Remodèle$(\mathcal{I}_s, \mathbb{R}^{M \times K})$; // Linéarisation de l'image segmentée.
  $C \leftarrow$ Normalisation(C);
  $S \leftarrow$ NNLS(C, D);
fin

Cet algorithme peut être utilisé avec n'importe quelle méthode de segmentation pouvant opérer sur des images avec plusieurs canaux. Il présente toutefois la limite de ne produire qu'une segmentation binaire. Cette contrainte a été implémentée avec la méthode de CSV qui demande de déterminer certains paramètres pour être efficace.

2.3.3.1 Paramétrisation

Comme présenté en section 2.3.2, la méthode CSV requiert de paramétriser 5 paramètres : ϱ_1, ϱ_2, υ, ϖ et Δt. Il est cependant possible de réduire le nombre de paramètres à partir des règles qui ont été évoquées en section 2.3.2.1. Ainsi, ϱ_1 et ϱ_2 sont fixés à 1 pour ne privilégier aucune classe, ϖ est paramétré à 0 puisqu'aucun *a priori* n'est effectué sur l'aire de la segmentation et $\Delta t = 0.5$ qui est la valeur par défaut [74]. Il reste à estimer υ. Son paramétrage a été effectué par évaluation d'un expert biologiste et d'un expert physicien à partir des résultats obtenus avec plusieurs valeurs entre 0 et 1 disponibles en annexe A.5. À l'issue de leur étude, la valeur de 0.35 a été sélectionnée comme donnant la meilleure segmentation.

Les figures 2.20 et 2.22 présentent respectivement les concentrations et les spectres obtenus en appliquant la MCR-ALS avec la contrainte de segmentation et une initialisation par la méthode VCA. En comparant les concentrations de la figure 2.16 avec celles de la figure 2.20, le résultat attendu est obtenu : la cellule est bien segmentée permettant une augmentation du contraste sur l'information à l'intérieur de la cellule.

Afin de pouvoir quantifier la qualité de la segmentation, la base de données présentée en section a été corrigée 2.3.1.4 et utilisée pour le calculer le $\mathfrak{D}$ du masque

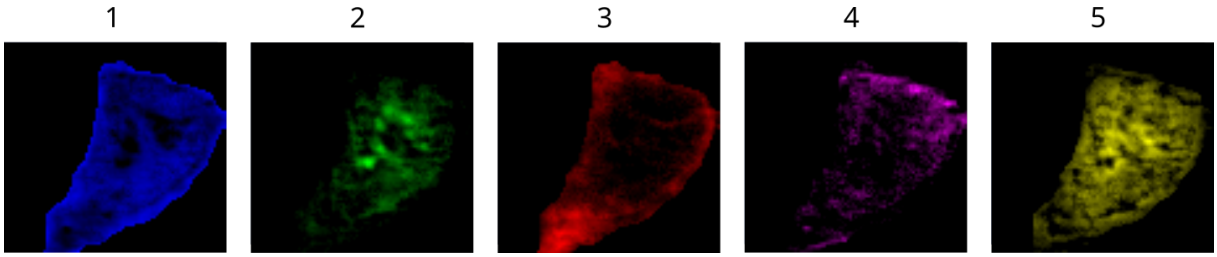


FIGURE 2.20 – Les 5 concentrations calculées par la MCR-ALS avec contrainte CSV sur une cellule HEK-293 fixée en interphase.

de segmentation. Le masque corrigé de la cellule de référence est présentée en figure 2.21, la figure 2.21a correspond à la segmentation initiale et la figure 2.21b à la segmentation corrigée. Cette correction consiste en une réduction du nombre de pixel appartenant au cytoplasme. Le $\mathfrak{D}$ obtenu par le masque calculé par la contrainte CSV est 0.77 surpassant les résultats obtenus avec la base de données initiale utilisée pour entraîner les modèles neuronaux. Les erreurs proviennent principalement de pixels du cytoplasme non-segmentés et peu à des pixels de l'environnement associés à la cellule.

FIGURE 2.21 – Correction de la segmentation d'une cellule HEK-293 fixée en interphase, (a) segmentation initiale de la cellule, (b) segmentation corrigée de la cellule. Dans la segmentation initiale : en vert l'environnement, en rouge le cytoplasme et en bleu le noyau. Dans la segmentation corrigée : en noir l'environnement, en blanc la cellule.

En confrontant les figures 2.17 et 2.22, il est possible de constater que les spectres ont été très peu impactés par l'ajout de la contrainte. Cela s'explique par l'initialisation qui n'est pas modifiée, les spectres initiaux prennent toujours en compte l'environnement. En ce qui concerne l'évolution du LOF, il passe de $\sim 0.017\%$ sans segmentation avec à $\sim 53,00\%$ avec la contrainte. Ce chiffre est cependant à relativiser puisqu'il prend en compte aussi l'environnement. Au sein de la partie segmentée, le LOF est estimé à $\sim 0.019\%$. La contrainte permet donc d'obtenir une bonne décorrélation de l'intérieur de la cellule.

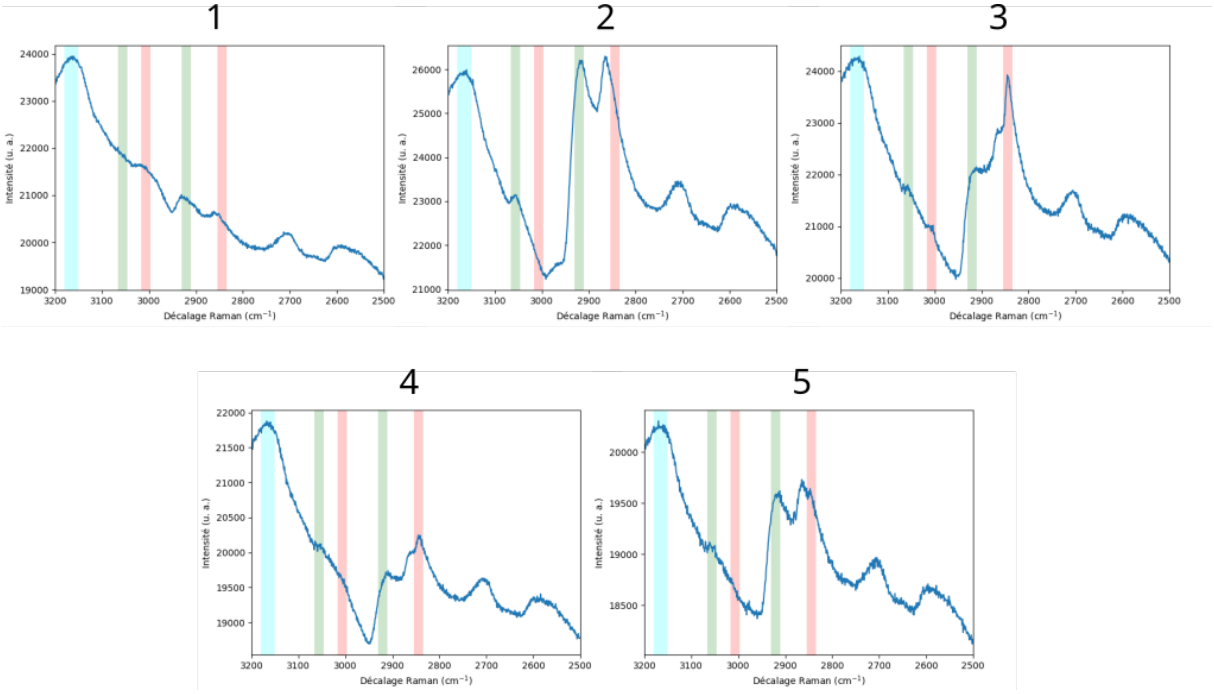


FIGURE 2.22 – Les 5 spectres calculées par la MCR-ALS avec contrainte CSV sur une cellule HEK-293 fixée en interphase.

Pour conclure, l'ajout de la contrainte de segmentation au sein de la MCR permet d'obtenir une seconde visualisation pour aider à l'analyse des résultats au sein d'une cellule. Elle permet aussi d'obtenir un masque de segmentation à partir des données en dimension réduites qui peut être utilisé dans d'autres opérations.

2.4 Conclusion

Dans ce chapitre, la réduction de dimension a été introduite avec les méthodes ACP et isomap. Elles permettent d'obtenir une nouvelle base maximisant la décorrélation linéaire, ou non, des données. Les limites de ces méthodes dans leur capacité à fournir des résultats en adéquation avec les contraintes numériques associées au phénomène CARS ont été montrées.

La famille des méthodes MCR a été introduite. Cette famille de méthodes permet de calculer à la fois des spectres et des concentrations des composants principaux d'un jeu de données. Cette famille introduit aussi l'utilisation de contraintes pour restreindre les solutions au problème posé à des solutions en accord avec la réalité du phénomène physico-chimique étudié. Un algorithme nommé MCR-ALS qui permet de résoudre de manière linéaire le problème a été introduit en détail. Les méthodes d'initialisation

SIMPLISMA et VCA pour l'algorithme ont été présentés. Une étude de la sélection du nombre de composants recherchés a été menée en comparant la sélection en utilisant une SVD à la mesure de la qualité de reconstruction. La méthode MCR-ALS a été appliquée à des données cellulaires et tissulaires et a permis de faire une analyse de l'état physiologique en accord avec l'état de l'art en biologie [30], [61]. La non-utilisation de méthode de recouvrement de phase a pu être validée en appliquant la MEM aux spectres calculés par la MCR-ALS. L'utilisation de la matrice de spectres S comme base de projection pour un jeu de données similaire a été introduite pour étudier l'évolution des composants entre deux jeux de données. Finalement, l'impact de la méthode d'initialisation a été étudié par la comparaison des résultats obtenus avec une initialisation par SIMPLISMA avec ceux obtenus avec l'initialisation par VCA. Ce test a permis de montrer la très forte dépendance à l'initialisation de la MCR-ALS. La méthode VCA permettant d'obtenir une plus grande dissimilarité entre les différents composants calculés, elle est à privilégier à la méthode SIMPLISMA.

Une contrainte de segmentation pour la MCR-ALS a été introduite [30]. Les limites de l'utilisation de réseaux de neurones pour implémenter cette contrainte sur les spectres CARS bruts, en raison du bruit des spectres et de la difficulté de construire une base d'apprentissage fiable, ont été mises en évidence. Une contrainte de segmentation appliquée aux concentrations avec la méthode de CSV a été implémentée et appliquée à des données de cellules. Cette nouvelle contrainte apporte une modification de la visualisation restreinte au contenu intracellulaire avec augmentation du contraste des concentrations dans la zone segmentée.

Bien qu'ayant obtenu des premiers résultats prometteurs, la méthode MCR-ALS présente toutefois des limites. Sur de gros volumes de données, la méthode requiert une quantité importante de mémoire vive. De plus, les régressions effectuées minimisent la norme L_2 qui n'est pas la plus adaptée pour évaluer des spectres. Ces limites conduisent à des échecs pour extraire dans les premiers composants une information pourtant présente.

Parmi les différentes architectures de réseaux de neurones, les auto-encodeurs présentent la particularité de pouvoir être entraînés de manière non-supervisés, levant la contrainte de construction d'un jeu d'apprentissage. Ils permettent aussi d'intégrer des opérations non linéaires et spatiales ou encore d'optimiser selon d'autres métriques que la norme L_2 par lots ne nécessitant pas de charger toutes les données dans la mémoire vive. Ces différentes propriétés ouvrent alors des possibilités nouvelles pour la MCR.

3

Résolution de courbes mutivariées par auto-encodeurs

Sommaire

Un auto-encodeur (AE) est une architecture de réseau de neurones formée de deux blocs appelés encodeur et décodeur. L'encodeur transforme les données dans un nouvel espace, souvent de plus petite dimension, puis le décodeur transforme les données encodées pour reconstruire celles d'origine. Un AE peut être entraîné en évaluant sa capacité à reconstruire les données d'entrée permettant ainsi de faire de l'apprentissage non-supervisé. En ajoutant certaines contraintes au modèle, il est alors possible d'utiliser un AE résoudre la MCR. L'utilisation des AE pour la MCR permet d'utiliser des opérations non linéaires et de tenir compte de l'aspect spatial du jeu de données tout en ayant le choix parmi différentes métriques de reconstruction.

Dans ce chapitre, les AE et leur utilisation pour la MCR sont introduits avec leur utilisation dans le cadre des données HSI. Deux méthodes de l'état de l'art sont étudiées sur un jeu de données HSI pour analyser les limites des méthodes existantes. Finalement, différentes propositions sur l'utilisation d'AE pour appliquer la MCR à des données CARS sont effectuées. Ces propositions permettent de discuter l'impact du prétraitement des données, l'initialisation du modèle, l'intégration de la non-négativité des spectres ainsi que l'intégration de la contrainte spatiale à partir de l'analyse des résultats obtenus sur un jeu de données artificiel.

3.1 Introduction aux auto-encodeurs

3.1.1 Présentation des auto-encodeurs

Les AE sont composés de deux sous-réseaux, ou blocs, appelés encodeur $\mathcal{E}$ et décodeur $\mathcal{D}$. Ces blocs représentent chacun une fonction projetant les données dans un nouvel espace. L'encodeur projette les données x dans un espace appelé espace latent $z = \mathcal{E}(x)$, z est alors appelé données encodées. Le décodeur est ensuite appliqué aux données projetées pour transformer une nouvelle fois les données $\hat{x} = \mathcal{D}(z)$, avec $\hat{x}$ les données décodées. Le plus souvent, l'espace latent est de dimension bien inférieure à l'espace d'origine et l'espace des données en sortie est de dimension égale à celui des données. De cette manière, il est possible d'apprendre au modèle à compresser des données et les décompresser en basant l'apprentissage sur la capacité du modèle à reconstruire les données d'origine à l'aide d'une fonction de coût $\mathcal{L}(x, \hat{x})$. La figure 3.1 présente le processus d'apprentissage d'un AE, les données x sont transformées par l'encodeur $\mathcal{E}$ en z et celles-ci sont retransformées en $\hat{x}$ par le décodeur $\mathcal{D}$. La fonction de coût peut alors être calculée entre x et $\hat{x}$ afin de pouvoir modifier les paramètres du modèle par descente de gradient. Lorsque $\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|$ et que l'encodeur et

le décodeur n'ont tous les deux qu'une couche sans fonction d'activation, la solution optimale est celle obtenue par ACP [79]. Cependant, si l'entraînement par descente de gradient n'applique pas de contrainte particulière, celle-ci n'est pas toujours atteinte.

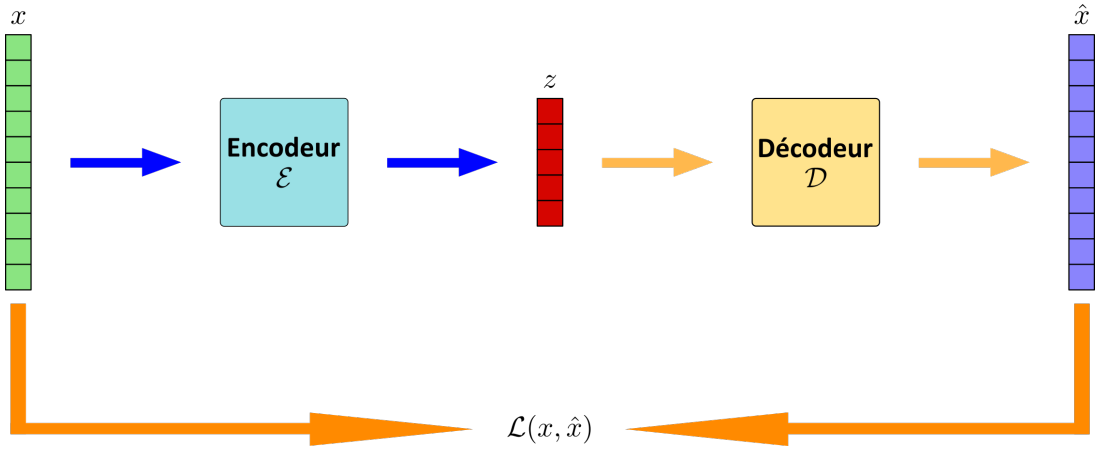


FIGURE 3.1 – Processus d'apprentissage non-supervisé standard d'un AE.

Il existe aussi différentes variations du modèle AE qui adaptent l'architecture au contexte d'application.

Les AE épars [80] utilisent un espace latent pouvant avoir une plus grande dimensionnalité que l'espace de départ mais impose une contrainte de parcimonie sur z. Ainsi, l'entraînement encourage le modèle à activer seulement quelques neurones à l'issue de l'encodage pour obtenir des descripteurs avec une bonne capacité de discrimination.

Les AE débruiteurs [81] sont conçus de manière semblable aux AE traditionnels mais sont entraînés sur des données bruitées en entrée avec comme sortie désirée les données débruitées :

$$\mathcal{L}(x, \mathcal{D}(\mathcal{E}(\tilde{x}))), \tag{3.1}$$

avec $\tilde{x} = x + \mathcal{N}(\mu, \sigma)$ les données bruitées par un bruit gaussien $\mathcal{N}$. Ces AE permettent d'intégrer une capacité de débruitage au réseau pour qu'ils puissent mieux gérer des données bruitées par la suite.

Les AE contractifs [82] sont similaires aux AE débruiteurs à la nuance qu'ils sont entraînés pour être robustes aux petites variations dans les données d'entrées plutôt que par un bruit additif gaussien. Cette robustesse est faite par une contrainte sur

la norme de la matrice jacobienne de z. La matrice jacobienne J_z correspond à la matrice des dérivées partielles de chacun des éléments de z en fonction de x. Une faible valeur pour la norme de Frobenius $\|J_z\|_F$ implique une faible variation de z pour une faible variation de x.

Les AEV sont un type d'AE [83] très différents des autres. Ces modèles sont principalement utilisés pour des tâches de génération. Les auto-encodeurs variationnels (AEV) définissent des modèles gaussiens qui servent à générer un nouvel échantillon à partir d'une donnée d'entrée. L'encodeur calcule les moyennes et écarts-types des différents modèles à partir de la donnée d'entrée. Ces différents modèles sont ensuite utilisés avec un échantillonnage aléatoire suivant une loi normale pour produire les données de l'espace latent qui génère une nouvelle donnée similaire à celle d'entrée. Les AEV sont principalement utilisés pour obtenir un modèle génératif de manière non-supervisée.

Grâce à leur capacité à projeter des données dans des espaces de plus petites dimensions ainsi que leur flexibilité sur le choix de l'architecture et de la fonction de coût, les AE font de bons candidats pour améliorer les résultats de la MCR. Ces dernières années, ils sont activement utilisés au sein de la communauté de l'HSI et ont su dépasser les méthodes de références.

3.1.2 Utilisation des auto-encodeurs dans l'imagerie hyperspectrale

3.1.2.1 Adaptation des AE pour la MCR

Le premier article utilisant des AE pour du démélange, équivalent de la MCR développée indépendamment au sein de la communauté HSI, remonte à 2015 où des auto-encodeurs débruiteurs et épars sont utilisés [84]. La résolution de la MCR est intégrée dans l'AE en utilisant l'espace latent comme concentration $Z = C$ avec Z l'ensemble des spectres encodés.

La détermination des signatures spectrales S varie selon la structure de l'AE utilisée mais est toujours extraite du décodeur. Dans le cas où ce dernier n'est composé que d'une seule couche dense, S correspond aux poids appris par cette couche [84].

Lorsque le décodeur est composé d'une couche convolutive, S est obtenu en faisant la somme des poids des filtres appris [85]. Cette approche est rendue possible

en redéfinissant le modèle MCR :

$$d_i = Sc_i + \sum_{k \in \mathfrak{N}(c_i)} Sc_k + e, \tag{3.2}$$

$\mathcal{N}(c_i)$ étant le voisinage du pixel c_i. Si un décodeur est composé d'une couche convolutive, un pixel $\hat{x}$ est reconstruit en faisant la somme des produits entre les paramètres du filtre W de taille $U \times U$ et les concentrations c du pixel et de son voisinage :

$$\hat{x}_{i,j} = \sum_{k=i-U/2}^{i+U/2} \sum_{l=j-U/2}^{j+U/2} w_{k,l} c_{k,l}. \tag{3.3}$$

Cette équation peut être reformulée de la manière suivante :

$$\hat{x}_{i,j} = (\sum_{k=i-U/2}^{i+U/2} \sum_{l=j-U/2}^{j+U/2} w_{k,l}) c_{i,j} + \sum_{k=i-U/2, k \neq i}^{i+U/2} \sum_{l=j-U/2, k \neq l}^{j+U/2} w_{k,l}(c_{k,l} - c_{i,j}). \tag{3.4}$$

On retrouve alors $S = \sum_{k=i-U/2}^{i+U/2} \sum_{l=j-U/2}^{j+U/2} w_{k,l}$, la somme des poids du filtre appris.

Si le décodeur est composé de plusieurs couches, l'obtention de S dépend des contraintes appliquées au décodeur. WANG *et al.* posent l'hypothèse d'un modèle de décodeur multi-couches denses avec S contenue dans les poids de la première couche du décodeur, les couches suivantes servent alors à modéliser le bruit contenu dans les données initiales [86]. Ce modèle est implémenté avec une première couche de décodeur qui passe de l'espace de dimension réduite à K dimensions à celui de départ à N dimensions. Les couches suivantes sont des couches denses avec des fonctions d'activation non linéaires à N dimensions. Une autre approche est l'utilisation de AEV comme générateur des spectres de la matrice S une fois entraîné. Les concentrations peuvent ensuite être déterminées à l'aide d'une autre méthode ou d'un encodeur séparé ou partagé avec celui du AEV [87], [88].

Dans l'apprentissage des réseaux de neurones, le choix de la fonction de coût est crucial, d'autant plus dans l'apprentissage de l'AE puisque ce choix correspond au choix de la métrique de reconstruction qui est sélectionnée pour être optimisée. La fonction de coût influe donc sur les critères qui vont être utilisés pour construire les matrices C et S. Elle peut être composée de différentes parties en fonction des contraintes qui sont appliquées au problème mais possède toujours une métrique de reconstruction pour évaluer la capacité du modèle, et ainsi des matrices calculées, à reconstruire les données d'origine.

3.1.2.2 Revue de l'existant

Deux métriques de reconstruction sont principalement utilisées comme fonction de coût : la norme L_2 [84], [88]-[93] et la SAD [85], [94]-[96]. La norme L_2 est la métrique de reconstruction la plus utilisée pour l'entraînement des AE [84], [88]-[93]. Cependant, comme il l'a été évoqué en section 2.1.2.4, cette métrique n'est pas la meilleure pour évaluer la proximité des spectres. En 2018, PALSSON *et al.* montrent que la métrique SAD permet d'obtenir une meilleure reconstruction et de meilleurs spectres signatures [94].

Pour coller au modèle de la MCR, il faut pouvoir appliquer des contraintes aux matrices C et S qui sont extraites du modèle. Les deux contraintes les plus usuelles étant celles de non-négativité et de normalisation, différentes propositions ont pu être faites dans les articles constituant l'état de l'art.

L'intégration de la contrainte de non-négativité aux concentrations. L'une des implémentations les plus communes est l'application d'une fonction de normalisation aux données encodées avec, par exemple, la fonction *softmax* [85], [87], [93], [95]. D'autres fonctions comme une fonction logistique [84], l'absolue [86] ou la ReLU [90], [92], [94], [96], [97] ont pu être utilisées. La fonction *softmax* comme dernière fonction d'activation de l'encodeur présente l'avantage de fusionner la contrainte de normalisation et de non-négativité. D'autres modèles utilisent l'initialisation par *fully constrained least squares* (FCLS) [98] avec un apprentissage qui ne change pas le signe du résultat [89]. La méthode FCLS est une méthode de résolution de moindres-carrés non-négatifs intégrant la contrainte de normalisation. La dernière approche est l'utilisation d'une fonction caractéristique n'acceptant que les solutions respectant la contrainte de non-négativité [88], [91] . L'approche par fonction d'activation *softmax* permettant d'assurer à la fois le respect de la contrainte de non-négativité et de normalisation dans la structure du réseau avant même l'entraînement fait de cette implémentation la plus intéressante pour la MCR par AE.

En plus des concentrations, les spectres sont traditionnellement contraints à la positivité.

La non-négativité des spectres. Cette opération n'est pas triviale puisqu'elle demande d'ajouter une contrainte aux paramètres du décodeur. Les modèles AEV utilisent des fonctions d'activation non-négatives en fin de modèle pour contraindre les spectres générés à être non-négatifs [87], [88]. Cette approche est efficace mais comme le

décodeur l'AEV sert de générateur de matrice de spectres S, il est ensuite nécessaire d'obtenir les concentrations d'une autre manière et de les combiner ensuite avec une méthode calculant les concentrations qui doit lui aussi être entraîné. D'autres modèles intègrent la non-négativité lors de la correction des poids [84], [89], [90] ce qui empêche l'utilisation d'architectures multi-couches et non linéaires.

La dernière contrainte usuelle est la normalisation des concentrations pour permettre une comparaison quantitative de la composition des différents pixels.

La normalisation des concentrations est principalement effectuée de deux manières distinctes. La première est l'utilisation d'une fonction de normalisation comme la fonction *softmax* [85], [87], [93], [95], [97] ou une division par la norme L_1 [86], [92], [94], [96] ou encore la norme $L_{2,1}$ [88]. La deuxième est d'intégrer la normalisation dans la régression [89], [90]. En définissant $\bar{S}$ et $\bar{D}$, les versions « augmentées » de S et D :

$$\bar{S} = \begin{bmatrix} S \\ \zeta 1_K^T \end{bmatrix}, \bar{D} = \begin{bmatrix} D \\ \zeta 1_M^T \end{bmatrix}, \tag{3.5}$$

K étant le nombre de composants, M le nombre de pixels et ζ un hyperparamètre à définir contrôlant l'importance de la contrainte de normalisation. L'augmentation des matrices permet alors à l'algorithme de régression calculant C de construire des concentrations dont la somme approche 1 pour chaque pixel [99]. Il existe aussi une troisième approche, moins courante, consistant en l'utilisation d'une fonction caractéristique à l'instar de la non-négativité des concentrations [88], [91]. Puisqu'elle permet d'implémenter deux contraintes en une seule opération en les intégrant dans la structure du réseau, la fonction *softmax* est l'approche la plus intéressante.

L'initialisation aléatoire des paramètres d'un réseau de neurones entraîne bien souvent une solution différente à chaque entraînement. Les paramètres du décodeur correspondant le plus souvent à la matrice S, il est attendu que les paramètres convergent vers des résultats proches, à des permutations des lignes près. Pour diminuer la variation dans les paramètres appris du modèle, une solution souvent adoptée est l'initialisation de certaines couches de neurones ou des concentrations.

L'initialisation par VCA . L'initialisation la plus courante est l'initialisation du décodeur à partir de la VCA [85], [89]-[92], [94]-[96]. Cependant, certains modèles initialisent aussi la matrice de concentration avec la méthode FCLS lorsqu'ils utilisent une régression sur-mesure [89]-[91]. Très peu de modèles n'utilisent pas d'initialisation [85], [87], [88].

Bien que permettant de faciliter fortement l'entraînement en initialisant le modèle dans des paramètres proches de la solution optimale, cette approche rend les différents modèles très dépendant de la méthode d'initialisation choisie et limite le décodeur à être formé que d'une seule couche dense définissant la matrice S.

L'une des principales caractéristiques des réseaux de neurones est de pouvoir être non linéaire. Cette non linéarité a pu être utilisée à plusieurs reprises pour la MCR.

L'intégration de la non linéarité. Cette intégration se fait principalement pour la résolution de C par l'utilisation d'un encodeur multicouche utilisant des fonctions d'activation non linéaires [86], [93]-[97]. Ces couches peuvent être denses [86], [94], [95] ou convolutives [96], [97]. WANG *et al.* proposent un décodeur non linéaire dont les spectres signatures sont représentés par la première couche de décodage alors que les couches suivantes ont pour objectif de contenir le bruit présent dans les données [86] et le réintégrer lors de la reconstruction. Les modèles par AEV reposent sur des modèles non linéaires pour générer les spectres [87], [88].

Les AE permettent aussi d'intégrer une contrainte spatiale dans la MCR à l'aide de convolutions ou encore de régularisation dans la fonction de coût.

L'ajout de la contrainte spatiale. Elle peut être abordée de différentes manières. La première est l'utilisation de contraintes de régularisation spatiales dans la fonction de coût [88], [95]. Une autre solution est l'utilisation de couches convolutives qui permettent de tenir compte du voisinage dans le calcul des concentrations [85], [97] ou des spectres [85]. Une troisième approche, proposée par QI *et al.*, utilise un algorithme de super-pixels [100] pour intégrer une contrainte de régularisation spatiale sur des pixels similaires spectralement et proches spatialement [96]. Cette régularisation spatiale est implémentée par deux AE partageant le même décodeur. Le premier AE opère sur les super-pixels alors que le deuxième traite tous les pixels indépendamment. Une fonction de coût composée des erreurs des deux modèles et une contrainte minimisant la différence entre les concentrations calculées par les deux modèles permettent l'apprentissage des deux modèles.

À ce stade, plus précisément afin d'identifier les limites des méthodes par AE pour la MCR, une étude sur deux modèles de l'état de l'art a été effectuée. Nous la détaillerons dans la section suivante.

3.2 Etude de modèles existants

Deux modèles ont été sélectionnés à partir d'une revue comparative proposée par PALSSON *et al.* [101] : EndNet [92] et CNNAEU [85]. Le modèle EndNet implémenté n'est pas le modèle original [92] mais une variation proposée par PALSSON *et al.* qui améliore le calcul des concentrations [101]. Ces deux modèles ont été choisis pour leurs bons résultats sur les données hyperspectrales et la variabilité des spectres calculée qui est moindre que celle des autres modèles évalués. La robustesse des résultats et la sensibilité de l'hyperparamétrage sont évaluées à l'aide d'un jeu de données HSI : Jasper Ridge. Le choix de d'abord évaluer les modèles sur des données HSI est fait afin de pouvoir valider l'implémentation ainsi que pour pouvoir profiter de données expertisées dont le nombre, les spectres et les concentrations des composants sont connus.

3.2.1 Le jeu de données Jasper Ridge

Pour étudier les modèles de référence sélectionnés, le jeu de données HSI « Japser Ridge », présenté en figure 3.2, est utilisé. Ce jeu de données comprend 100×100 pixels. À l'origine, les spectres sont composés de 224 canaux allant de 380 à 2500 nm. Cependant, des canaux perturbés par la vapeur d'eau et les effets atmosphériques sont supprimés, réduisant le nombre de mesures par spectre à 198.

FIGURE 3.2 – Jeu de données Jasper Ridge en rouge vert bleu.

4 composants sont présents au sein du jeu de données : les arbres, l'eau, la terre et la route. Leurs spectres caractéristiques et concentrations sont disponibles en figures 3.3 et 3.4. Le composant le plus complexe à calculer avec les algorithmes de

MCR est celui de la route en raison de sa faible présence spatiale et de la proximité de son spectre avec celui du sol.

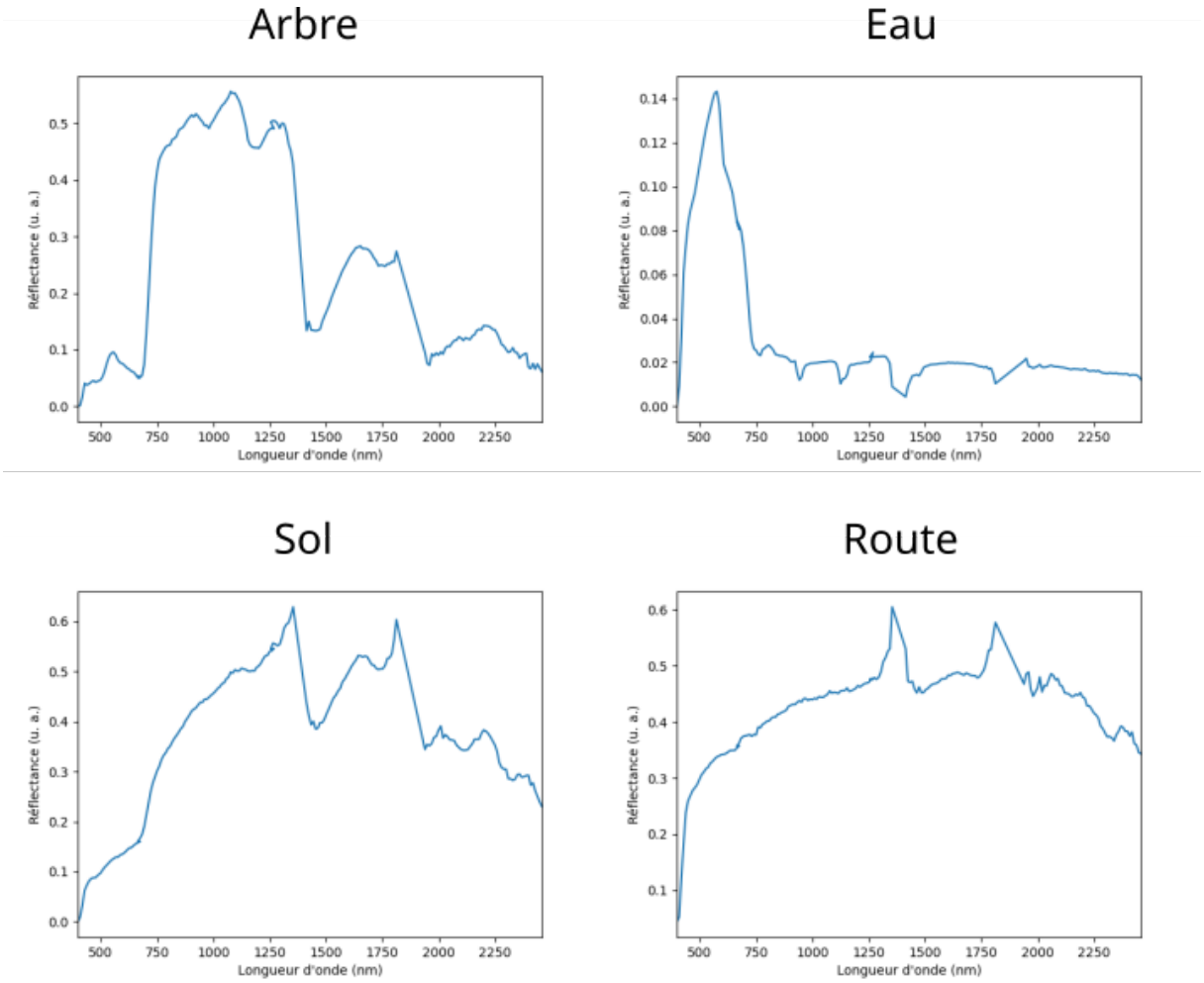


FIGURE 3.3 – Les spectres des 4 composants de Jasper Ridge.

3.2.2 EndNet

3.2.2.1 Présentation du modèle

Le modèle EndNet [92], [101] est résumé en tableau 3.1. Ce modèle présente la particularité d'utiliser une couche dense personnalisée pour l'encodeur. En effet, au lieu de faire un produit matriciel, la SAD entre les spectres en entrée et les paramètres de l'encodeur est calculée pour obtenir les concentrations qui sont ensuite normalisées par la fonction *softmax*. Les spectres S sont obtenus à partir des poids du décodeur appris. À la fois l'encodeur et le décodeur sont initialisés en utilisant la méthode VCA.

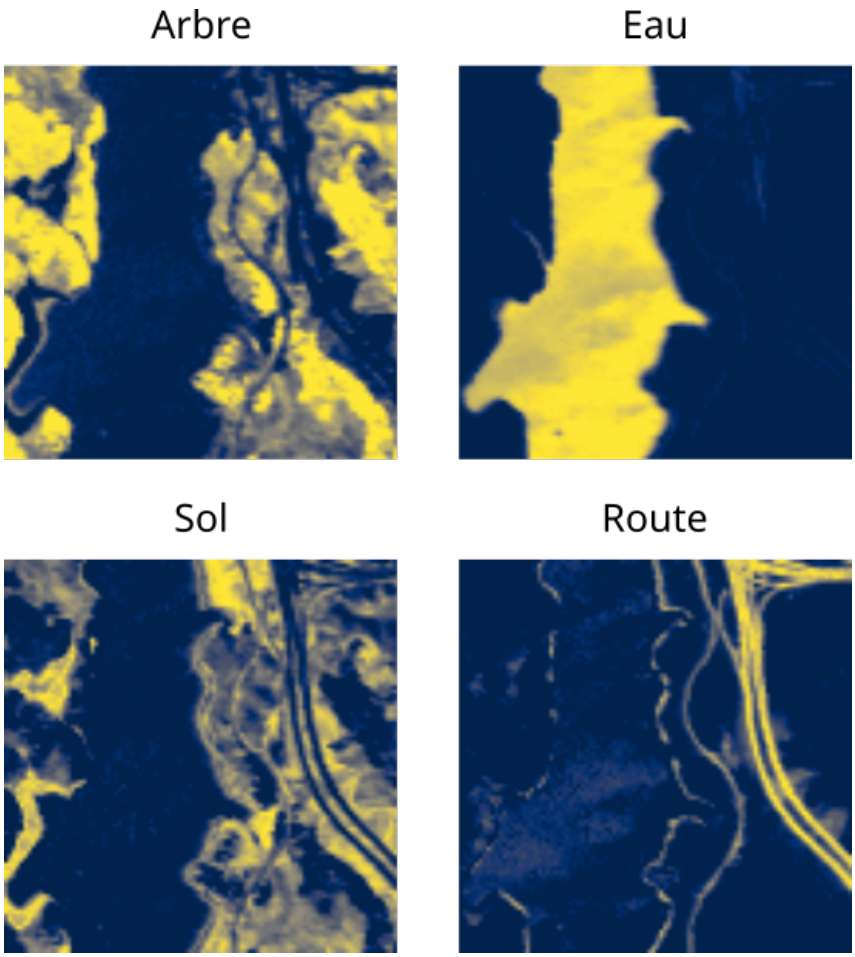


FIGURE 3.4 – Les concentrations des 4 composants de Jasper Ridge.

Bloc	Couche	Descr.	Activation.
Encodeur	DenseSAD	K	*Softmax*
Décodeur	Dense	N	

TABLEAU 3.1 – Architecture du modèle EndNet. N correspond au nombre de canaux spectraux, K est le nombre de composants recherché, descr. signifie descripteur. DenseSAD correspond à une couche dense utilisant la SAD plutôt que le produit matriciel.

La fonction de coût utilise à la fois l'erreur quadratique moyenne (EQM), et la SAD. De plus, des contraintes de régularisation sont appliquées avec une contrainte de norme L_1 sur les concentrations ainsi qu'une contrainte de norme L_2 sur les poids du décodeur :

$$\mathcal{L}_{EndNet}(X,Y) = \lambda_1 \frac{\sum_{i,j}(x_{ij}-y_{ij})^2}{BN} + \lambda_2 \frac{\sum_i SAD(x_i,y_i)}{B} + \lambda_3 \frac{\sum_i |c_i|}{B} + \lambda_4 \frac{\sum_i \|s_i\|_2}{B}, \quad (3.6)$$

avec λ_1, λ_2, λ_3 et λ_4 des hyperparamètres pondérant les différents éléments de la

fonction de coût et B le nombre d'éléments dans le lot.

3.2.2.2 Influence de l'initialisation

Pour observer l'influence de l'initialisation sur les capacités du modèle, trois types d'initialisation ont été évalués :

— initialisation de l'encodeur et du décodeur avec la VCA,

— initialisation de l'encodeur aléatoire et du décodeur avec la VCA,

— initialisation de l'encodeur et du décodeur aléatoire.

L'optimiseur ADAM est utilisé avec des lots de taille 64 comme proposé par OZKAN *et al.* [92]. Les hyperparamètres λ_1 et λ_2 sont fixés empiriquement à 5×10^{-3} et 10 et les paramètres λ_3 et λ_4 en utilisant les valeurs proposées par OZKAN *et al.* : 0,1 et 1×10^{-4} [92]. 25 entraînements de 250 itérations sont répétés sur le jeu de données « Jasper Ridge ». Les spectres obtenus sont ensuite classés dans un des composants de la vérité terrain à l'aide de l'algorithme du plus proche voisin avec la SAD afin de pouvoir étudier la variabilité des résultats.

L'initialisation de l'encodeur et du décodeur par VCA. Elle permet d'obtenir les résultats présentés en figures 3.5 et 3.6. Les spectres ne varient pas entre les différents entraînements montrant une grande stabilité du décodeur en raison de l'initialisation. Les spectres calculés pour les arbres et l'eau sont visuellement similaires à ceux de la vérité terrain. Il en est de même pour les concentrations, bien que trop contrastées par rapport à la vérité terrain. Le modèle a plus de difficultés avec le sol et la route qui sont mélangés dans les images des concentrations. Ce mélange est aussi visible sur les spectres obtenus.

Les résultats obtenus en initialisant aléatoirement l'encodeur. Ils sont présentés en figures 3.7 et 3.8. Sur la figure 3.7 présentant les spectres, il est possible de constater que des variations sont apparues sur tous les composants. Le composant des arbres est le moins impacté et varie assez peu, le composant de l'eau l'est davantage avec des pics liés à l'eau présents dans certains spectres et des valeurs négatives. Le sol présente des variations principalement d'échelles d'intensités, même si certains spectres ont aussi leur allure qui varient. Pour finir, la route présente peu d'allures différentes parmi les différents spectres calculés. Cependant, parmi ces allures, l'une d'elles diffère fortement des autres et mélange à la fois l'eau et la route. Au regard des concentrations en figure

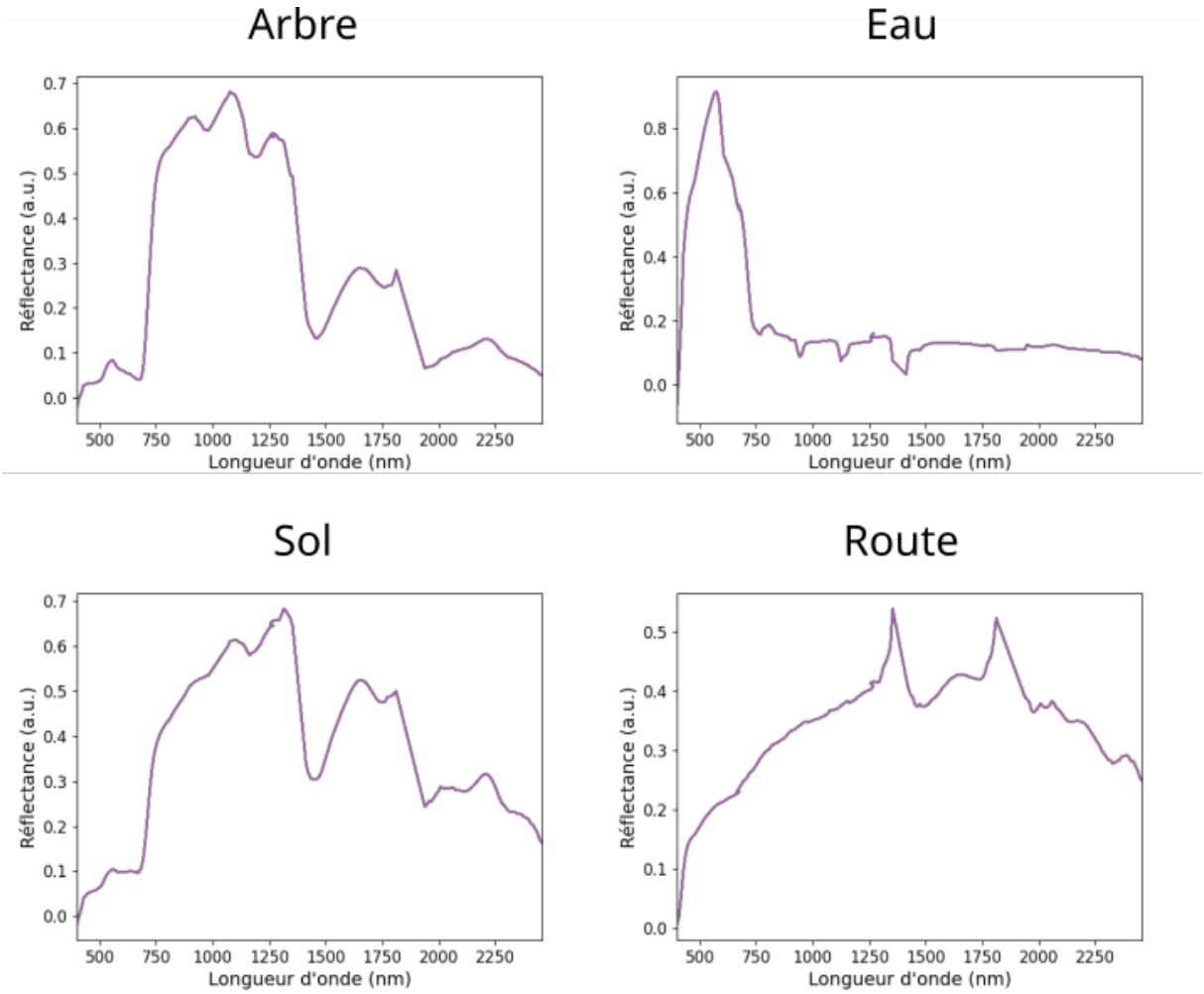


FIGURE 3.5 – Les spectres calculés sur 25 entraînements par le modèle EndNet initialisé par la VCA.

3.8, à l'exception des arbres, ces dernières sont beaucoup moins contrastées que dans la vérité terrain. La route se retrouve aussi dans les concentrations de l'eau et du sol et, réciproquement, l'eau et le sol se retrouvent dans les concentrations de la route.

Initialisation aléatoire. Les résultats obtenus en initialisant aléatoirement le modèle sont présentés en figures 3.9 et 3.10. Le phénomène déjà constaté en initialisant aléatoirement l'encodeur est amplifié. Les variations de spectres présentés dans la figure 3.9 sont plus importantes et impactent tous les composants. Les arbres et l'eau ont des spectres avec des valeurs négatives. Les erreurs de concentrations de la figure 3.10 sont aussi amplifiées : l'eau, le sol et la route sont plus mélangés qu'en initialisant aléatoirement seulement le décodeur.

Le tableau 3.2 présente les SAD et EQM moyennes ainsi que leurs écarts-types

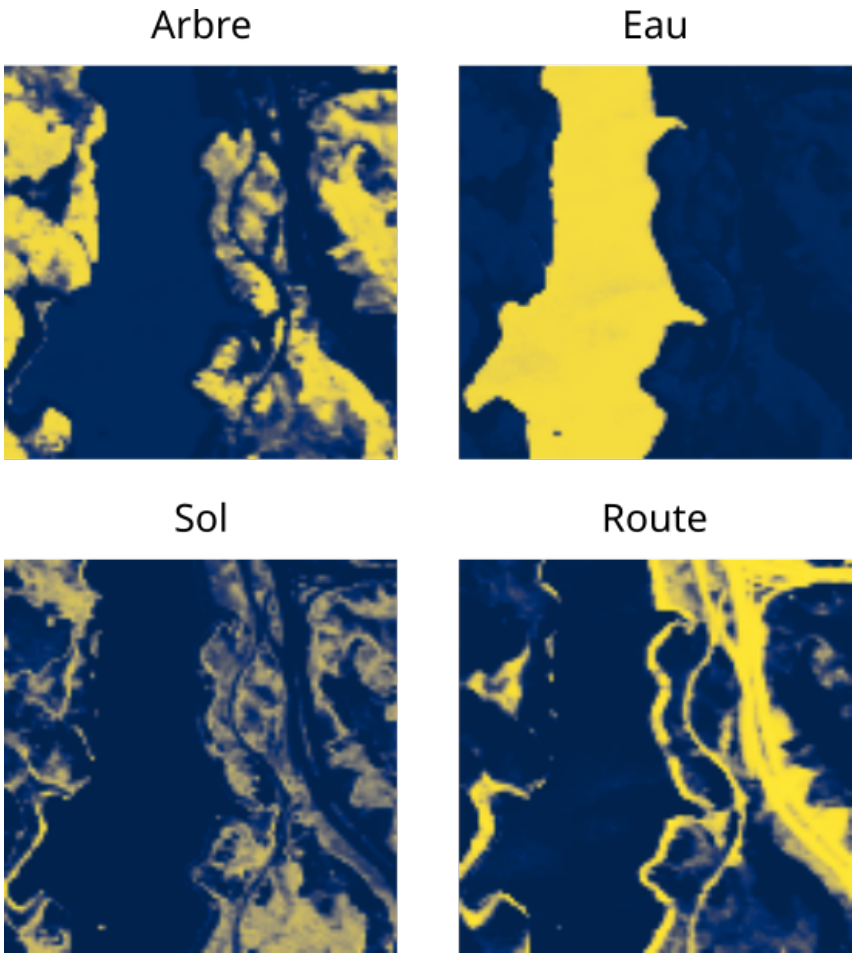


FIGURE 3.6 – Concentrations moyennes sur 25 entraînements calculés par le modèle EndNet initialisé par la VCA.

pour chaque modèle sur les 25 entraînements effectués. La SAD est calculée entre les spectres calculés et les spectres de référence alors que la EQM est calculée sur les concentrations obtenues et celles de référence. Pour le modèle initialisé par VCA, EndNetA, les écarts-types nuls sur les spectres confirment la stabilité de l'apprentissage du décodeur. Avec un angle de 0,175 et 0,192 radian respectivement, les SAD des spectres du sol et de la route sont bien plus importantes que celles des arbres et de l'eau comme l'analyse visuelle l'indiquait. Contrairement aux spectres, les écarts-types des concentrations ne sont pas tous nuls. Ce comportement s'explique par le non-déterminisme de certains algorithmes utilisés pour l'optimisation ainsi que par l'in-utilisation de contrainte sur les poids de l'encodeur. Lorsque l'initialisation est aléatoire, la plupart des métriques augmentent en moyenne ou voient leur écart-type augmenter. Cette variation montre une forte sensibilité du modèle à la méthode d'initialisation qui pourrait donner des résultats bien différents selon la méthode d'initialisation sur des

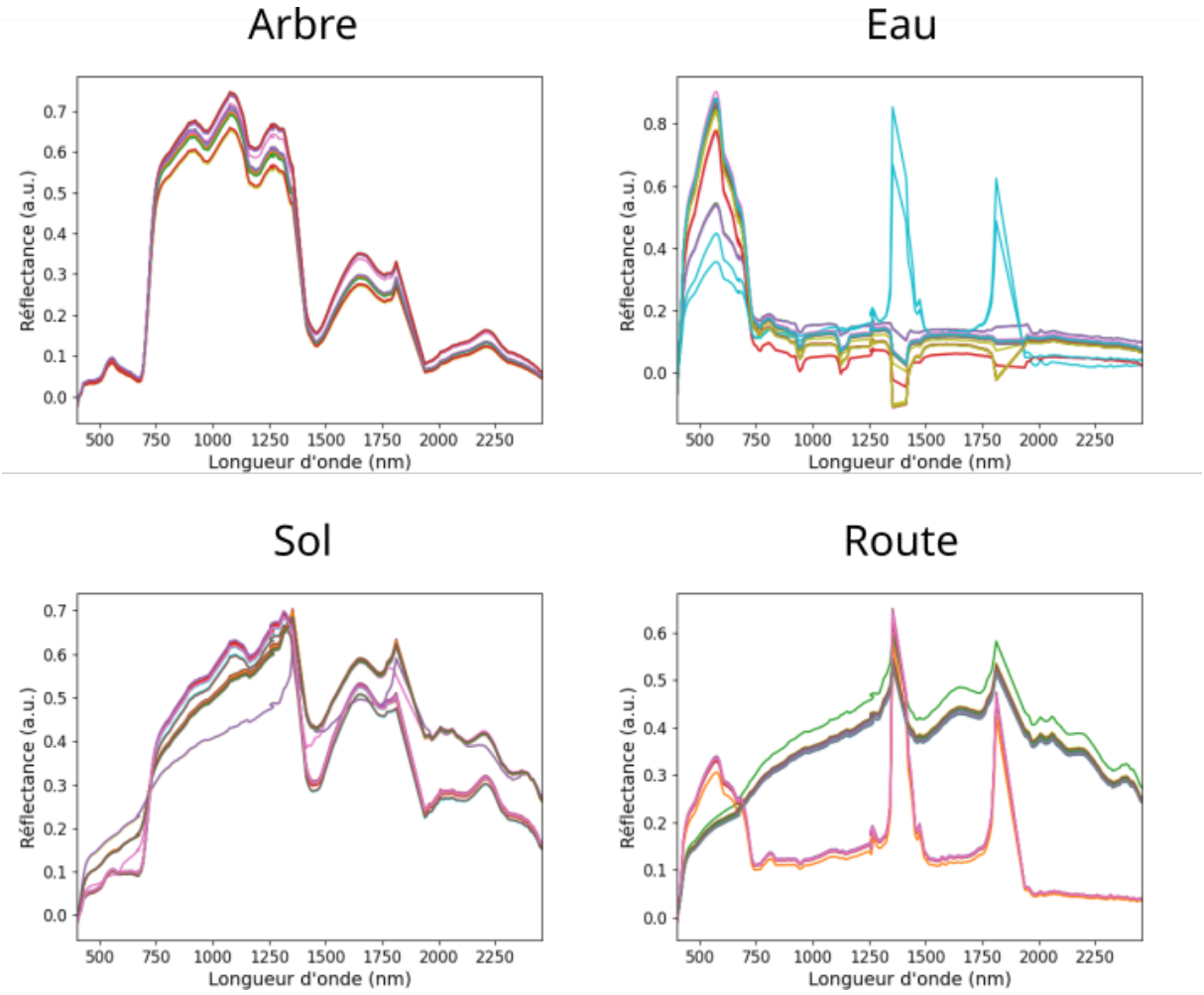


FIGURE 3.7 – Les spectres calculés sur 25 entraînements par le modèle EndNet avec l'encodeur initialisé aléatoirement.

données plus complexes à décomposer.

Une dernière information permettant d'évaluer la qualité du résultat est le nombre de spectres classés dans chacun des composants. Si le modèle donne des résultats stables et réussi à retrouver les différents composants, 25 spectres doivent être associés à chaque composant, sinon de la redondance est présente dans la matrice S calculée et certains composants absents. Lorsque le modèle est initialisé avec la VCA, 25 spectres sont associés à chaque composant. En initialisant l'encodeur aléatoirement, les routes sont associées à 25 spectres, l'eau à 30, le sol à 27 et la route à 18. Finalement, en initialisant aléatoirement, 26 spectres sont associés aux arbres, 36 à l'eau, 25 au sol et 13 à la route. Cette évolution dans la classification des spectres renforce la conclusion de la forte dépendance du modèle à l'initialisation avec l'ajout de l'aléatoire qui fait apparaître des composants en doublon et d'autres manquants dans certains entraînements.

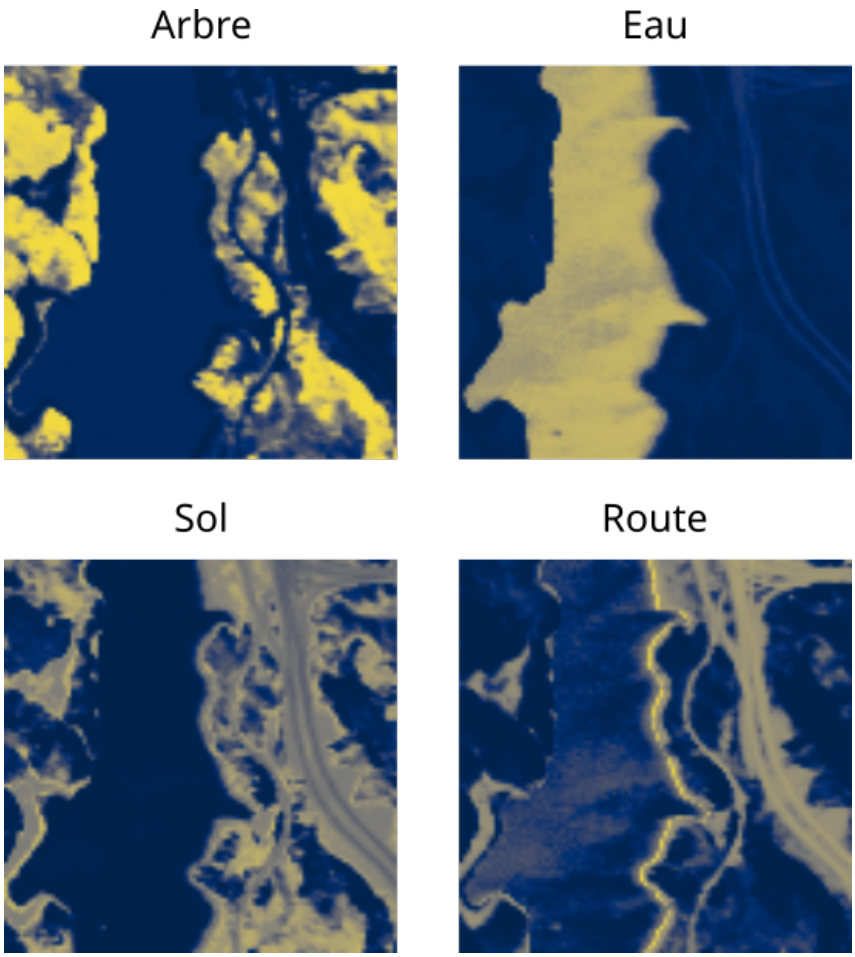


FIGURE 3.8 – Concentrations moyennes sur 25 entraînements calculés par le modèle EndNet avec l'encodeur initialisé aléatoirement.

3.2.3 CNNAEU

3.2.3.1 Présentation du modèle

Le modèle CNNAEU [85] est résumé en tableau 3.3. C'est un modèle d'AE dont à la fois l'encodeur et le décodeur sont convolutifs. L'encodeur est composé de deux couches convolutives utilisant la fonction LeakyReLU. La fonction LeakyReLU est une fonction ReLU modifiée pour ne pas être valoir 0 lorsque l'entrée est négative mais des valeurs négatives formant une pente de faible intensité :

$$LeakyReLU(x) = \begin{cases} x & \text{si } x \geq 0, \\ \alpha x & \text{sinon,} \end{cases} \tag{3.7}$$

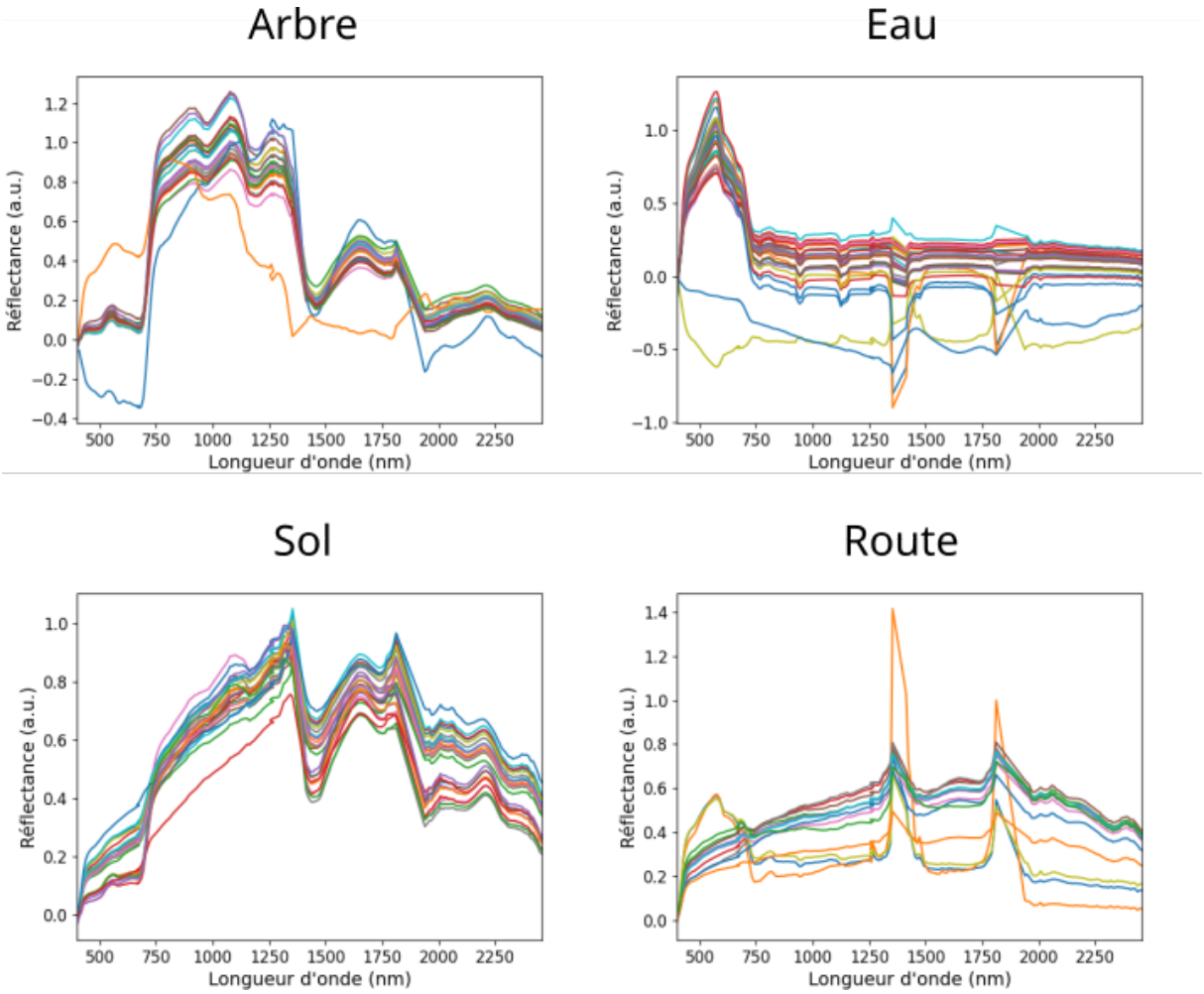


FIGURE 3.9 – Les spectres calculés sur 25 entraînements par le modèle EndNet initialisé aléatoirement.

$0 < \alpha < 1$ étant un hyperparamètre à régler. L'objectif de cette fonction est de limiter l'impact du gradient nul dans les négatifs de la fonction ReLU. Pour calculer les concentrations, les descripteurs calculés par les couches convolutives sont multipliés par un scalaire ϵ puis normalisés par la fonction *softmax*. Le paramètre de mise à l'échelle ϵ a pour rôle de favoriser des concentrations éparses [101].

Le décodeur est composé d'une seule couche convolutive permettant d'extraire les spectres à partir des filtres appris :

$$s_i = \sum_{j=1}^{F_X} \sum_{k=1}^{F_Y} w_{i,j,k}, \tag{3.8}$$

avec w_i le filtre du $i$ème descripteur, U_X et U_Y les tailles de filtre en horizontal et vertical. Le modèle CNNAEU présente la particularité d'initialiser aléatoirement le décodeur.

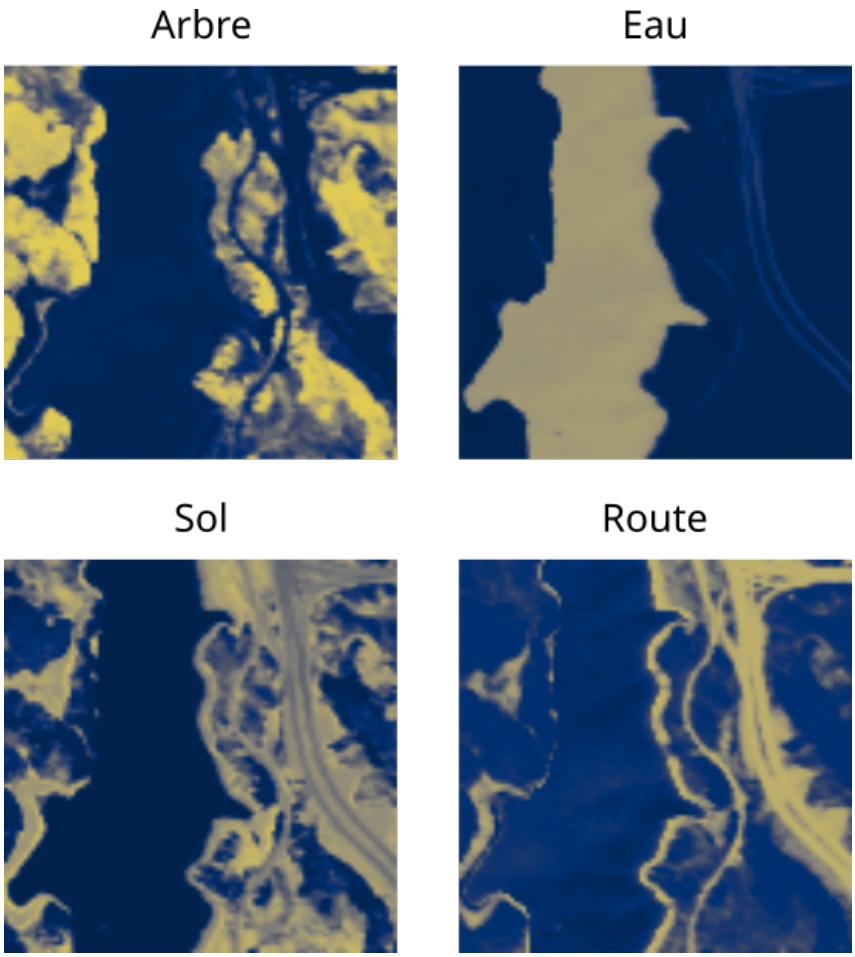


FIGURE 3.10 – Concentrations moyennes sur 25 entraînements calculés par le modèle EndNet initialisé aléatoirement.

Afin d'évaluer la sensibilité de l'hyperparamétrage du modèle, deux valeurs d'ϵ, le paramètre de mise à l'échelle des concentrations avant normalisation, sont évaluées. Les deux valeurs de ϵ sont 3,5, la valeur proposée par PALSSON *et al.* [85], et 1 pour analyser les résultats sans mise à l'échelle.

3.2.3.2 Influence du paramètre de mise à l'échelle

Pour évaluer le modèle, 25 entraînements de 320 itérations sont répétés. Les différents choix d'hyperparamétrage sont ceux proposés par PALSSON *et al.* [85]. Des patchs de 40×40 pixels sont créés à partir du jeu de données Jasper Ridge pour générer les données d'entraînements. Afin d'augmenter la taille du jeu de données et introduire de l'invariance à la translation dans l'apprentissage des filtres, l'espacement entre chaque patch est de 20 pixels. L'optimiseur utilisé est RMSprop [102] basé sur une moyenne glissante des gradients pour éviter les minimums locaux, avec un taux

Modèle	Métrique	Arbre	Eau	Sol	Route
EndNetA	SAD	0.073 ± 0.0	0.034 ± 0.0	0.175 ± 0.0	0.192 ± 0.0
	EQM	0.317 ± 0.006	0.458 ± 0.038	0.148 ± 0.0	0.192 ± 0.0
EndNetB	SAD	0.063 ± 0.014	0.130 ± 0.151	0.106 ± 0.078	0.298 ± 0.245
	EQM	0.315 ± 0.009	0.391 ± 0.072	0.179 ± 0.034	0.147 ± 0.056
EndNetC	SAD	0.112 ± 0.115	0.312 ± 0.508	0.078 ± 0.057	0.176 ± 0.191
	EQM	0.293 ± 0.032	0.370 ± 0.077	0.179 ± 0.037	0.170 ± 0.038

TABLEAU 3.2 – SAD et EQM des concentrations et spectres calculés par le différentes configurations de EndNet. EndNetA correspond au modèle complètement initialisé avec la VCA, EndNetB correspond au modèle dont l'encodeur est initialisé aléatoirement et EndNetC au modèle complètement initialisé aléatoirement.

Bloc	Couche	Descr.	Norm. lots	Abandon	Activation
Encodeur	Conv(3×3)	48	✓	0.2	LeakyReLU
	Conv(3×3)	K	✓	0.2	LeakyReLU
	Mise à l'échelle ϵ	K			*Softmax*
Décodeur	Conv(11×11)	N			

TABLEAU 3.3 – Architecture du modèle CNNAEU. Conv signifie convolutif, descr. descripteur, norm. lots normalisation par lots et abandon indique la probabilité qu'un neurone soit désactivé.

d'apprentissage de 0,0003. La taille des lots est de 15 patchs. La fonction de coût utilisée est la SAD multipliée par 10 pour amplifier l'erreur et augmenter son gradient. Aucune contrainte n'est appliquée sur les concentrations et les spectres.

Résultats obtenus avec $\epsilon = 3.5$. Les résultats sont présentés en figures 3.11 et 3.12. À l'analyse des spectres de la figure 3.11, assez peu de variation apparaît entre les différents spectres, à l'exception des spectres du sol. En effet, une partie des spectres associés au sol sont aussi similaires à ceux de la route. Les concentrations moyennes, présentées en figure 3.12, montrent les arbres et l'eau bien localisés spatialement mais avec un contraste bien trop important. Le modèle a cependant des difficultés à séparer le sol et la route, le sol ayant ses concentrations les plus élevées aux mêmes pixels que l'eau mais de manière moins intense. Ses valeurs s'expliquent par les spectres du sol qui sont proches de ceux de l'eau.

Résultats obtenus en n'utilisant pas la mise à l'échelle. Ils sont présentés en figures 3.13 et 3.14. Les spectres sont peu impactés par la mise à l'échelle avec une variation plus légèrement plus importante sur les différents composants. Les concentrations moyennes de la figure 3.12 sont plus impactées. Les concentrations

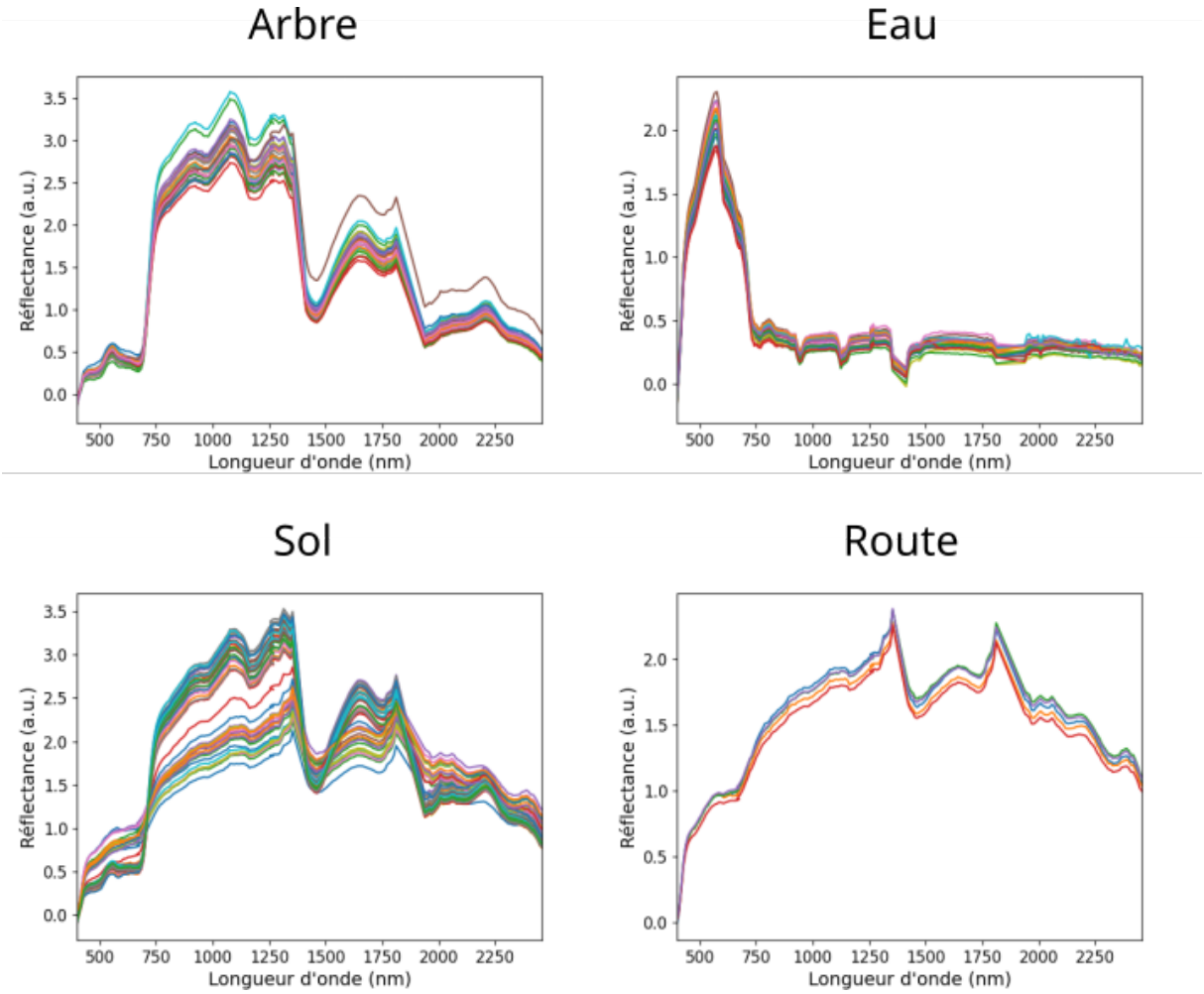


FIGURE 3.11 – Les spectres calculés sur 25 entraînements par le modèle CNNEAU avec $\epsilon = 3.5$.

des arbres sont moins intenses, de même pour les sols et la route. L'eau cependant n'est pas impactée par la mise à l'échelle. Le sol et la route sont toujours mélangés mais les concentrations de la route la mettent plus en avant qu'avec la mise à l'échelle. Il est aussi possible de constater un effet de flou sur les concentrations. Ce flou est imputable à la convolution spatiale pour calculer les concentrations.

Le tableau 3.4 présente les SAD et EQM moyennes ainsi que leurs écarts-types pour chaque modèle sur les 25 entraînements effectués. Les différentes métriques d'évaluation des spectres évoluent peu mais une augmentation de l'écart-type est présente. En ce qui concerne les concentrations, celles-ci s'améliorent en retirant la mise à l'échelle. L'amélioration des concentrations s'explique par la parcimonie des concentrations générées par la mise à l'échelle qui est trop importante avec $\epsilon = 3.5$. L'obtention de concentrations trop contrastées par rapport à la vérité terrain est un phénomène déjà constaté lors du développement du modèle [85].

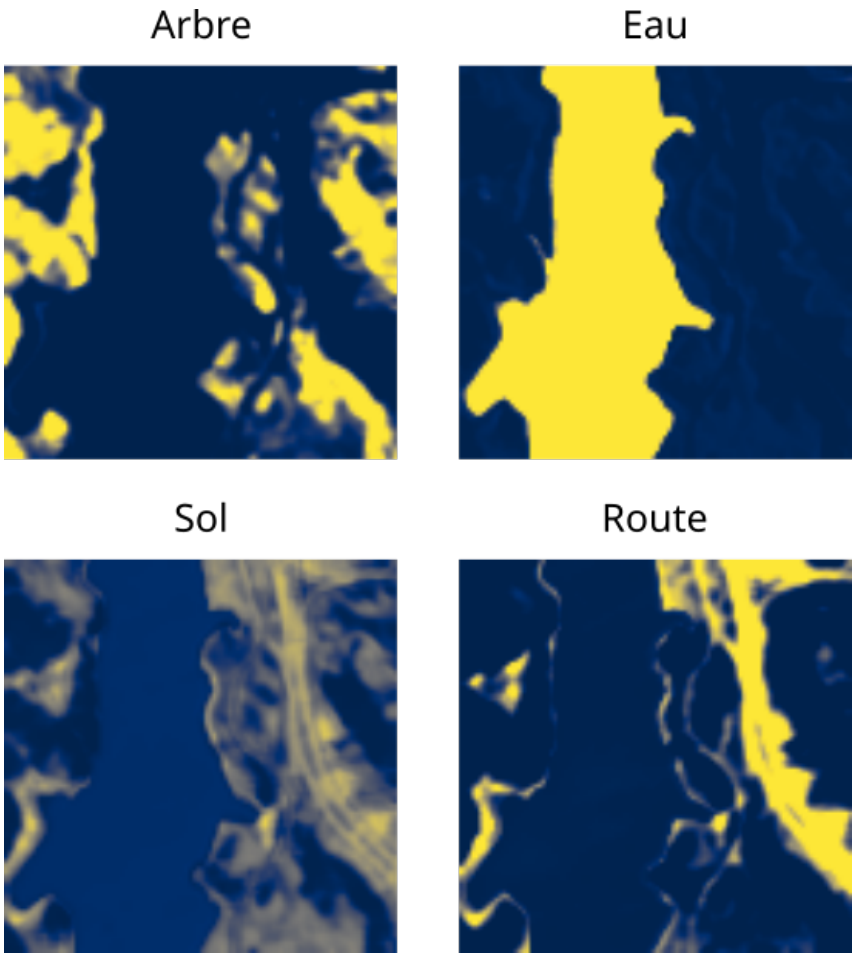


FIGURE 3.12 – Concentrations moyennes sur 25 entraînements calculés par le modèle CNNEAU avec $\epsilon = 3.5$.

Le classement des spectres calculés du modèle avec mise à l'échelle groupe 28 spectres pour les arbres, 25 pour l'eau, 43 pour le sol et 4 pour l'eau. En enlevant la mise à l'échelle, le classement associe 27 spectres aux arbres, 25 à l'eau, 43 au sol et 5 à l'eau. Le modèle a donc d'importantes difficultés à calculer des spectres similaires à celui de l'eau. Ce constat est en accord avec l'analyse des spectres qui montre que des spectres associés au sol et à l'eau très proches les un des autres.

Un dernier test a été effectué sur le modèle avec mise à l'échelle en l'entraînant sur des données CARS afin d'évaluer la capacité du modèle à traiter des spectres d'autres origines. Il a été choisi de seulement tester le modèle CNNAEU en raison de son indépendance à toute autre méthode puisqu'il ne nécessite pas d'initialisation pour être efficace contrairement à EndNet.

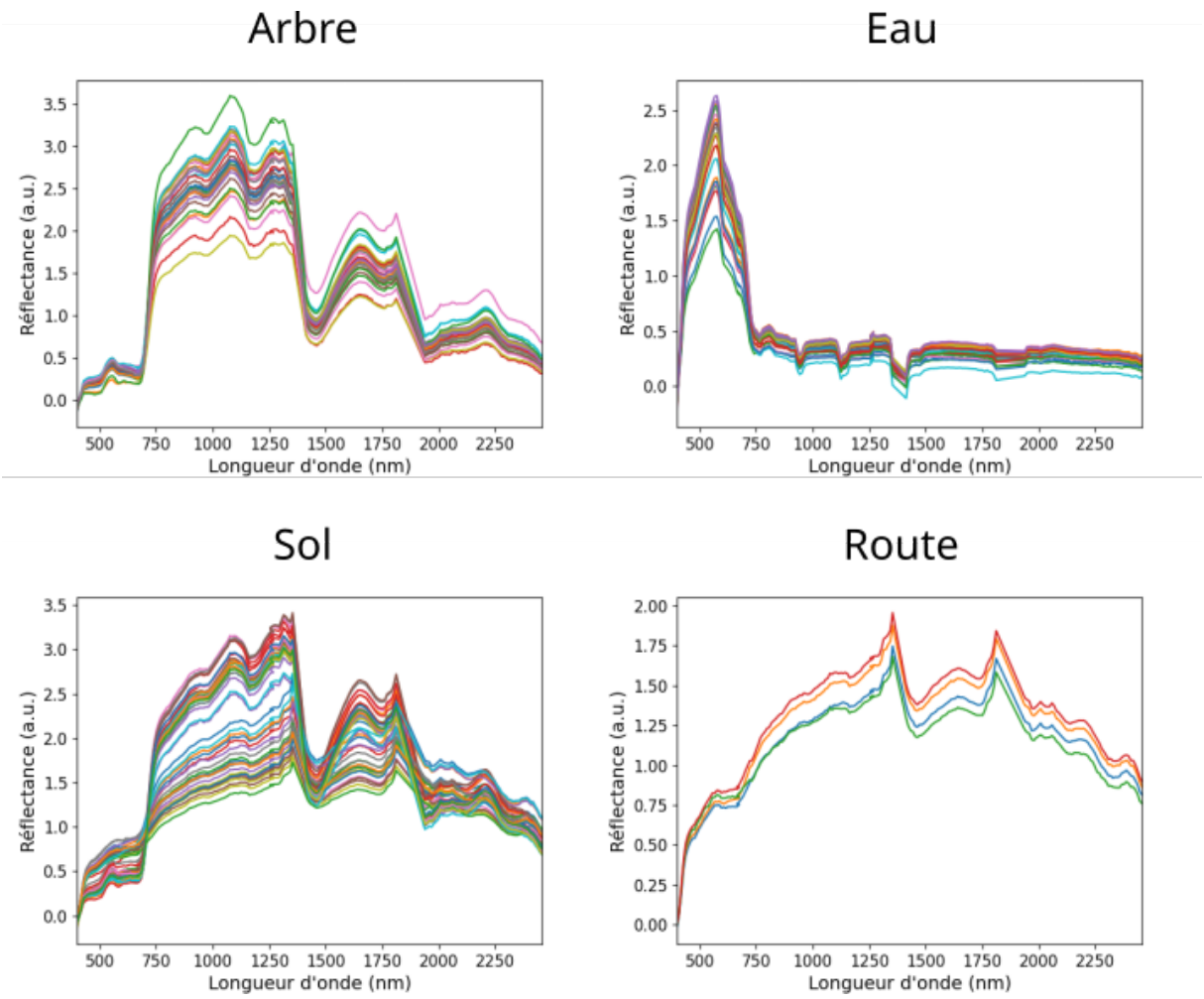


FIGURE 3.13 – Les spectres calculés sur 25 entraînements par le modèle CNNEAU avec $\epsilon = 1$.

3.2.3.3 CNNAEU appliqué aux données CARS

Le modèle CNNAEU est évalué sur la cellule de référence de la section 1.4.2.1. L'hyperparamétrage utilisé est le même que pour les données HSI de Jasper Ridge présenté en section 3.2.3.2. Comme aucune vérité terrain n'est disponible pour classer les spectres calculés, l'algorithme K-moyennes avec $K = 4$ est utilisé pour définir 4 groupes de spectres.

Les spectres calculés sont disponibles en figures 3.15 et 3.16. Les spectres des différents composants de la figure 3.15 sont particulièrement bruités et similaires. Il est difficile de les associer à des éléments biologiques. Les concentrations de la figure 3.16 n'aident pas à l'analyse. Le composant 1 semble illuminer tout sauf les membranes, le composant 2 illumine le cytoplasme et un peu plus fortement les nucléoles. Le composant 3 complète le composant 1 et donc semble pouvoir être associé aux

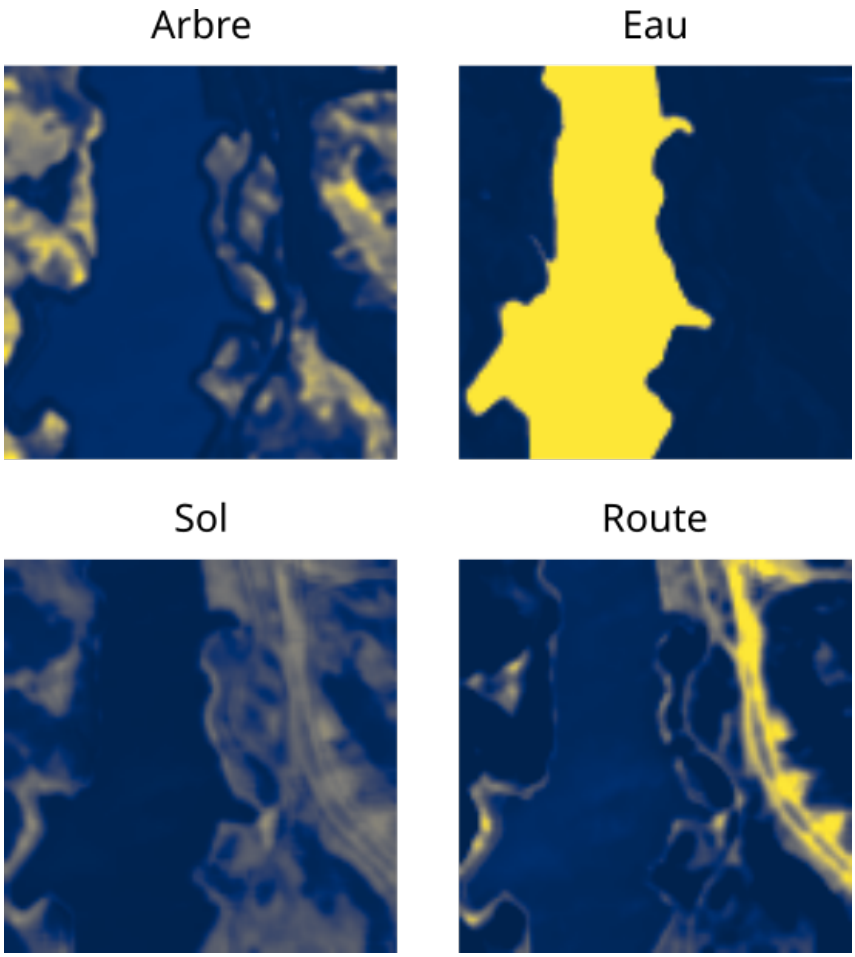


FIGURE 3.14 – Concentrations moyennes sur 25 entraînements calculés par le modèle CNNEAU avec $\epsilon = 1$.

Modèle	Métrique	Arbre	Eau	Sol	Route
CNNAEUA	SAD	0.088 ± 0.031	0.060 ± 0.019	0.139 ± 0.051	0.111 ± 0.006
	EQM	0.389 ± 0.073	1.666 ± 0.760	0.268 ± 0.307	0.212 ± 0.063
CNNAEUB	SAD	0.090 ± 0.033	0.061 ± 0.029	0.139 ± 0.051	0.102 ± 0.008
	EQM	0.272 ± 0.109	1.250 ± 0.566	0.142 ± 0.024	0.104 ± 0.017

TABLEAU 3.4 – SAD et EQM des concentrations et spectres calculés par le différentes configurations de CNNAEU. EndNetA correspond au modèle complètement initialisé avec la VCA, EndNetB correspond au modèle dont l'encodeur est initialisé aléatoirement et EndNetC au modèle complètement initialisé aléatoirement.

membranes. Le composant 4 met en avant principalement le cytoplasme. En l'absence de spectres analysables, il n'est pas possible de statuer sur la véracité des hypothèses émises sur les concentrations.

À la suite de l'analyse des résultats obtenus, nous pouvons conclure que le modèle CNNAEU échoue à analyser les spectres CARS de la cellule de référence. Pour

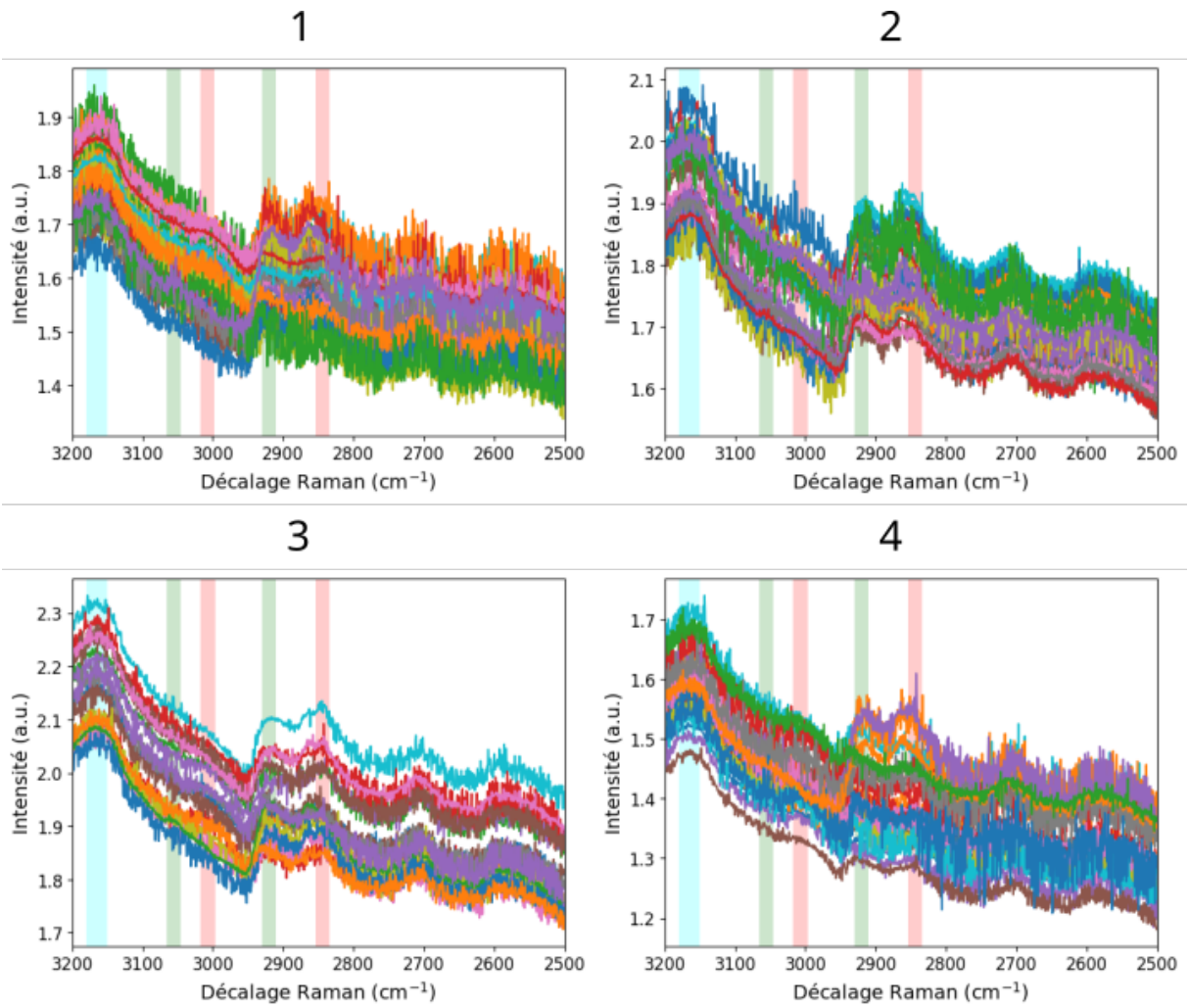


FIGURE 3.15 – Les spectres calculés sur 25 entraînements par le modèle CNNEAU appliqué à des données CARS.

mieux comprendre où se situe les limites des AE pour analyser des données CARS, différents tests ont été effectués sur la cellule de référence.

3.3 Etude du paramétrage des auto-encodeurs pour la résolution de courbes multivariées

Différentes catégories de tests ont été effectués pour évaluer l'architecture à mettre en place pour utiliser un AE comme MCR. Dans un premier temps, les prétraitements à appliquer aux données et l'implémentation de la contrainte de non-négativité aux poids du décodeur sont discutés. Ensuite, l'impact de l'initialisation, le choix de la

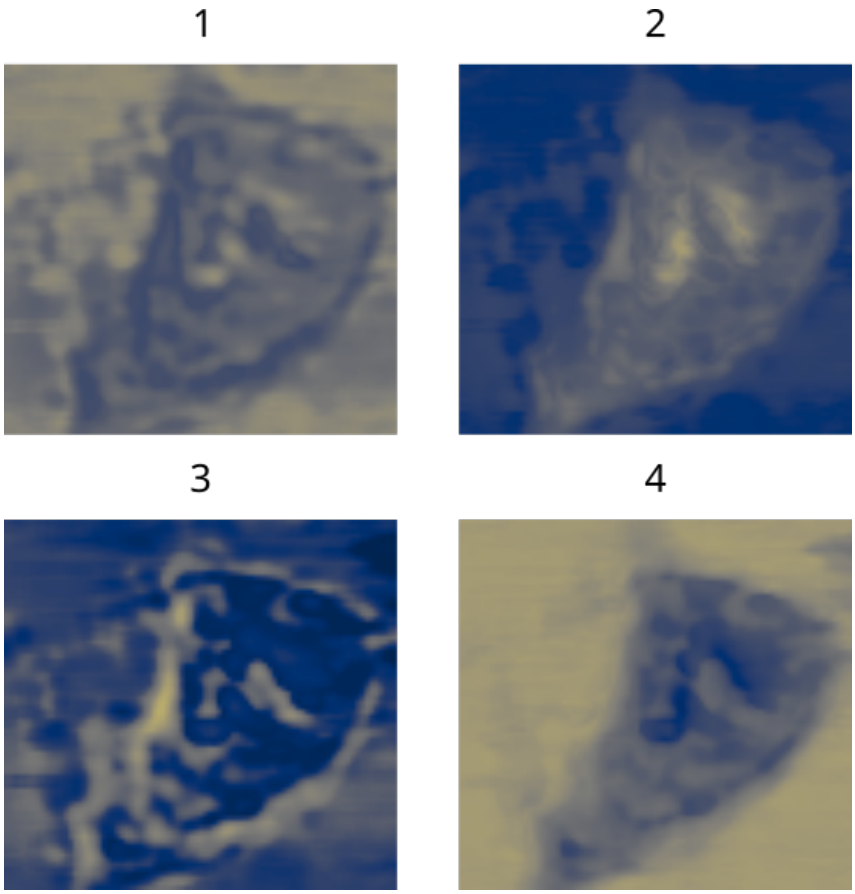


FIGURE 3.16 – Concentrations moyennes sur 25 entraînements calculés par le modèle CNNEAU appliqué à des données CARS.

fonction de coût et l'application de contraintes aux résultats sont examinés.

Bloc	Couche	Descr.	Norm. lots	Abandon	Activation.
Encodeur	Dense	16	✓	0.2	ReLU
	Dense	K			*Softmax*
Décodeur	Dense			N	

TABLEAU 3.5 – Architecture du modèle utilisé pour évaluer le paramétrage des AE pour la MCR.

Pour mener ces études, un modèle commun présenté en tableau 3.5 est défini. Il se compose de deux couches denses pour l'encodeur et une couche dense pour le décodeur. La couche intermédiaire calcule 16 descripteurs et est suivie d'une normalisation du lot, d'une couche d'abandon et utilise la fonction d'activation ReLU. La normalisation du lot a pour rôle de lisser la fonction de coût et d'éviter certains minima locaux [68]. La couche d'abandon a pour rôle de limiter le surapprentissage qui entraînerait des composants trop dépendants de descripteurs intermédiaires spécifiques. Le choix de 16 descripteurs pour la couche intermédiaire est fait empiriquement afin

de réduire le nombre de paramètres à apprendre dans le réseau. La couche finale de l'encodeur utilise la fonction d'activation *softmax* pour implémenter les contraintes de non-négativité et de normalisation.

Les modèles sont entraînés 10 fois sur 200 itérations avec l'optimiseur ADAM en utilisant des lots de taille 64. La fonction de coût utilisée est la SAD. Afin de permettre une étude quantitative des résultats, un jeu de données artificiel est utilisé pour ces expériences. Ce jeu de données permet l'obtention d'une vérité terrain et de classer les différents composants calculés par les différents modèles. Le classement des composants se fait en associant aux spectres trouvés le composant au spectre réel le plus proche en terme de SAD. Ainsi, pour évaluer la qualité des résultats, les moyennes et écart-types des SAD entre les spectres calculés et les spectres réels des composants sont calculées. Il en est de même pour l'EQM des concentrations calculées avec celles du jeu de données. De plus, les figures de concentrations montrées utiliseront la carte de couleur *cividis* avec une échelle de couleur allant de 0 à 1 afin de pouvoir identifier si les différentes méthodes permettent de caractériser complètement certains pixels.

3.3.1 Jeu de données artificiel

Afin de pouvoir faire une évaluation quantitative de la qualité des méthodes étudiées, un jeu de données artificiel est utilisé. Ce dernier est construit de sorte à représenter, de manière simplifiée, une cartographie de cellule. Pour ce faire, les spectres des composants calculés par la MCR-ALS sur la cellule de référence de la section 1.4.2.1. servent de modèles pour construire les spectres des composants artificiels. Dans un soucis de simplification de la modélisation, les pics présents dans chaque composant sont supprimés et la réalité biologique est simplifiée. Les spectres CARS sont générés à l'aide de l'équation 1.4. Pour générer $\chi^{(3)}_{NR}$, une combinaison de deux fonctions sigmoïdes est utilisée [23], [24] :

$$\chi^{(3)}_{NR}(\omega) = \frac{1}{1+e^{-(\omega-q_1)l_1}} * \frac{1}{1+e^{-(\omega-q_2)l_2}}, \tag{3.9}$$

avec q_1 et q_2 les coefficients d'inclinaison des sigmoïdes et l_1 et l_2 leurs points d'inflection. $\chi^{(3)}_R$ est généré à partir de l'équation 1.7 en définissant les caractéristiques des zones vibrationnelles des différents composants. Les valeurs des paramètres $q_1 = 0.01, l_1 = 2640, q_2 = 0.001$ et $l_2 = 2990$ sont choisis empiriquement. Le spectre de $\chi^{(3)}_{NR}$ obtenu avec ces paramètres est présenté en figure 3.17.

Le jeu de données utilisé est composé 90×90 pixels et 916 canaux spectraux allant

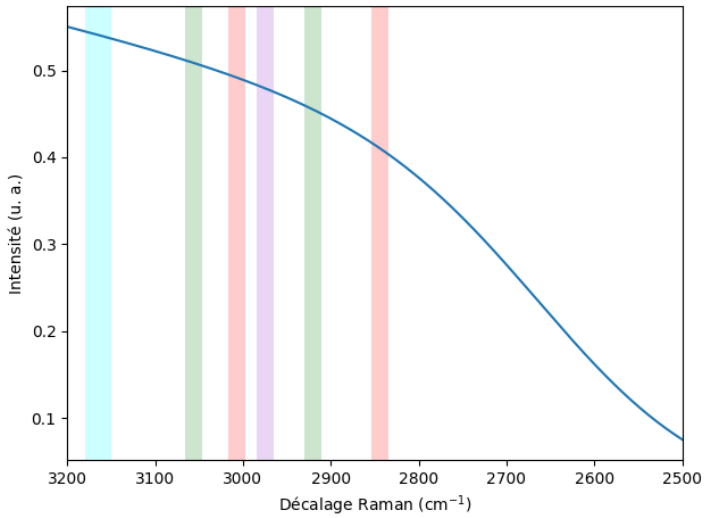


FIGURE 3.17 – $\chi^{(3)}_{NR}$ généré avec $q_1 = 0.01, l_1 = 2640, q_2 = 0.001$ et $l_2 = 2990$.

de 3200 à 2500 cm^{-1}. Quatre composants forment le jeu de données : l'environnement, le cytoplasme, le noyau et les membranes. Chaque composant est caractérisé par un spectre pour former la matrice S.

3.3.1.1 Spectres des composants

L'environnement est composée d'un seul pic vibrationnel. Ce dernier est un pic aqueux situé à 3165 cm^{-1}, fait une largeur de 20 décalages Raman et a une intensité de 1. Le spectre CARS ainsi que le spectre défini par la partie réelle et le module de $\chi^{(3)}_R$ sont présentés en figure 3.18.

Le cytoplasme mélange eau, protéines et lipides. Le cytoplasme est le composant le plus simplifié par rapport à la réalité. En effet, il est considéré comme uniforme dans le jeu de données or de nombreux organites sont présents en son sein dans une vraie cellule empêchant toute uniformité spectrale. Le spectre associé au composant est présenté en figure 3.19. Il contient un pic aqueux à 3165 cm^{-1} d'une largeur de 20 décalages Raman et une intensité de 0.6, un pic de protéines d'une largeur de 10 décalages Raman et d'intensité 0.12 à 3056 cm^{-1} et un deuxième pic protéique à 2920 cm^{-1} de largeur de 25 décalages Raman et d'intensité 1. Des lipides sont aussi présents avec un pic d'intensité 0.15 et de largeur de 10 décalages à 3007 cm^{-1} et un autre à 2844 cm^{-1} d'intensité 0.15 et de largeur 15.

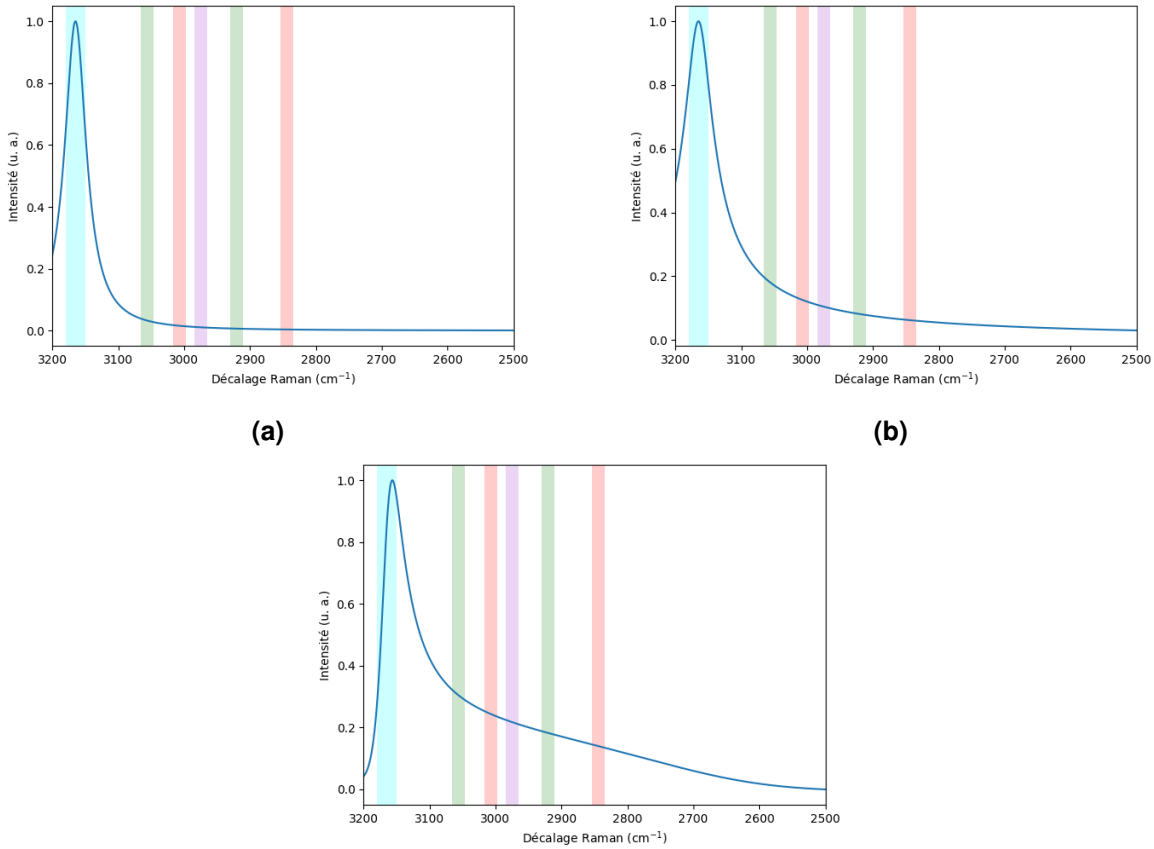


FIGURE 3.18 – Spectres du composant eau, (a) partie imaginaire de $\chi_R^{(3)}$, (b) module de $\chi_R^{(3)}$, (c) spectre CARS.

Le noyau contient de l'eau, des protéines et de l'ADN. L'eau est représenté par un pic à 3165 cm^{-1} de largeur 20 et d'intensité 0.3. Les protéines sont représentées par un pic d'une largeur de 10 décalages Raman et d'intensité 0.12 à 3056 et d'un second à 2920 cm^{-1} de largeur 25 et d'intensité 1. L'ADN est symbolisée par un pic d'une largeur de 10 décalages Raman et d'intensité 0.05 à 2875 cm^{-1}. Le spectre correspondant au composant est montré en figure 3.20.

Les membranes contiennent des lipides et des protéines. Les lipides sont représentés par deux pics de largeurs de 10 décalages Raman et d'intensités 0.1 à 3007 et 2844 cm^{-1}. Les protéines sont présentées par un pic d'une largeur de 10 décalages Raman et d'intensité 0.12 à 3056 et d'un deuxième d'intensité 1 et de largeur de bande de 25 décalages Raman à 2920 cm^{-1}. Le spectre du composant est disponible en figure 3.21.

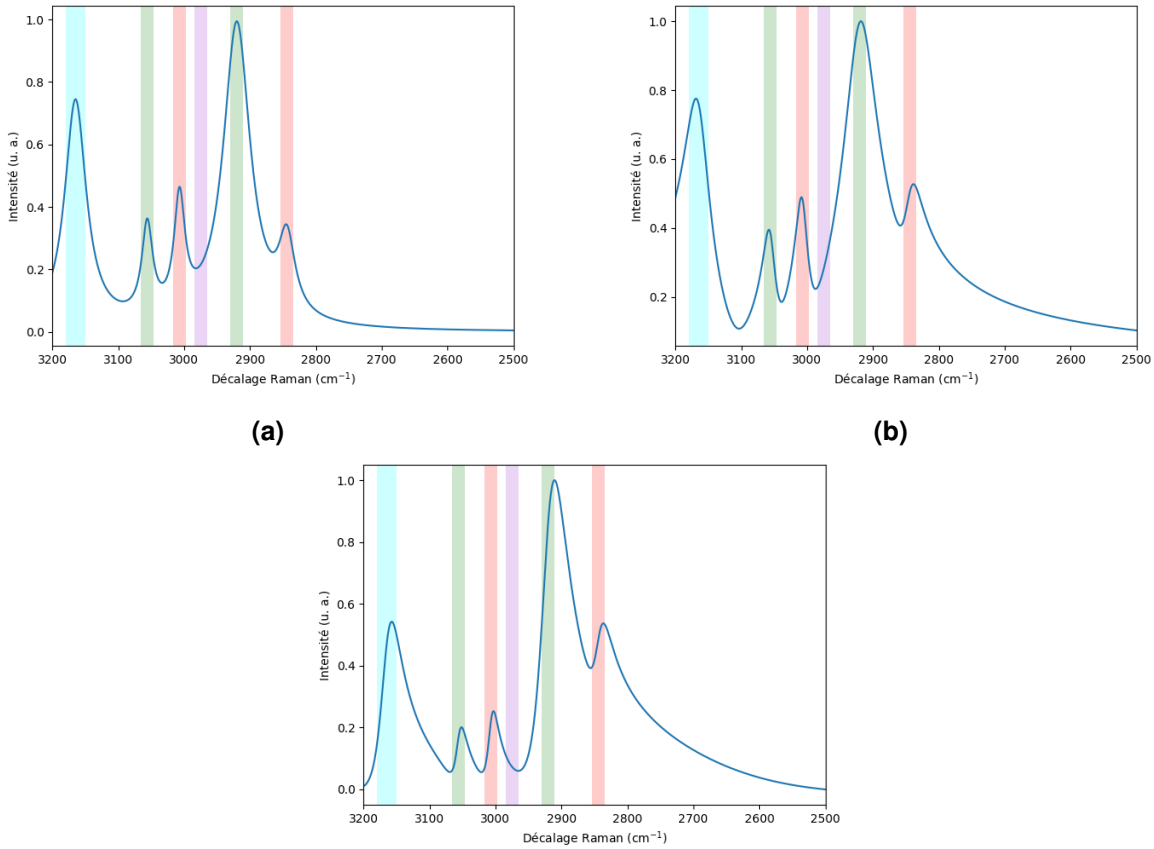


FIGURE 3.19 – Spectres du composant cytoplasme, (a) partie imaginaire de $\chi_R^{(3)}$, (b) module de $\chi_R^{(3)}$, (c) spectre CARS.

3.3.1.2 Concentrations des composants

Pour générer les concentrations, chaque pixel est associé à un composant à l'aide d'un vecteur où 1 correspond à son composant et 0 pour les autres. Ces concentrations sont ensuite mélangés dans une fenêtre de taille $\mathcal{V} \times \mathcal{V}$ en utilisant une pondération κ aléatoire :

$$c_{i,j,k} = \sum_{l=i-\mathcal{V}/2}^{i+\mathcal{V}/2} \sum_{m=j-\mathcal{V}/2}^{j+\mathcal{V}/2} c_{l,m,k} * \kappa_{l,m,k}. \tag{3.10}$$

Une fois le mélange effectué, les concentrations sont normalisées pour que la somme fasse 1.

Les concentrations obtenues sont présentées en figure 3.22. Le mélange lisse les changements de composant, les membranes étant peu épaisses, les différents

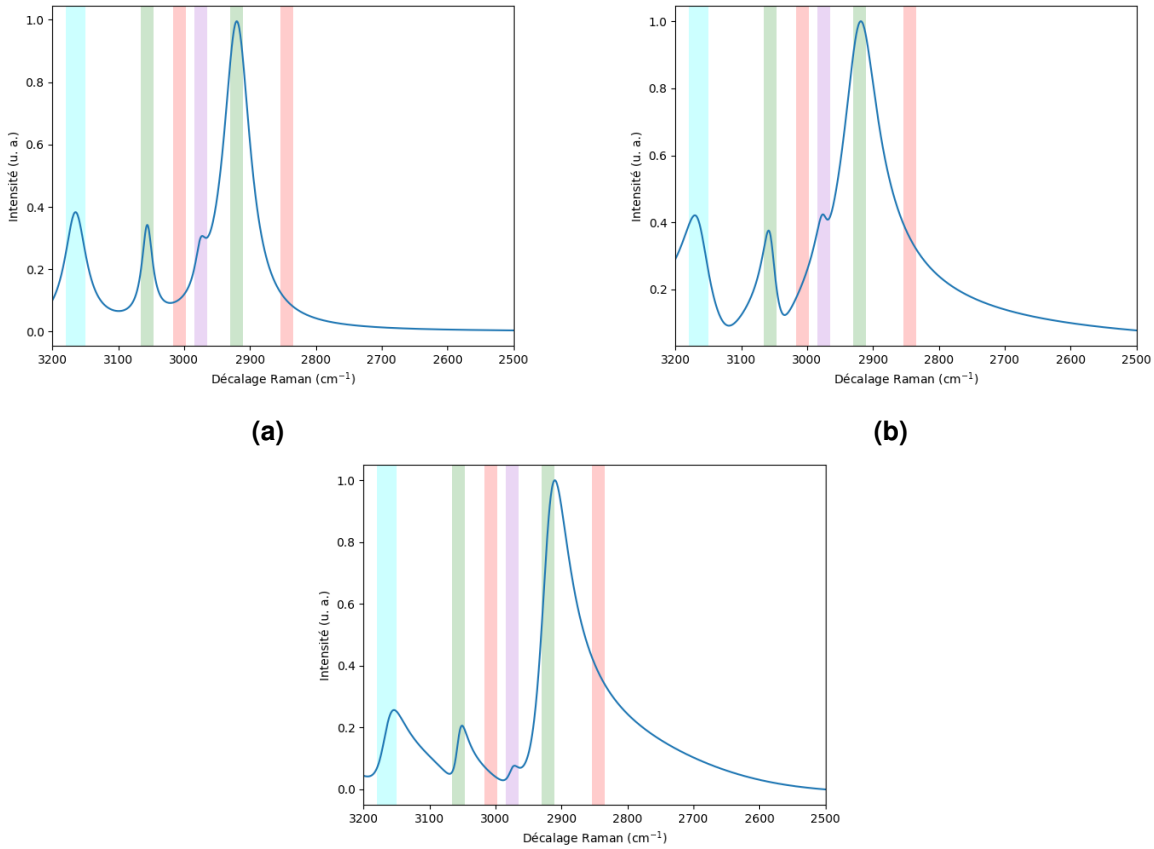


FIGURE 3.20 – Spectres du composant noyau, (a) partie imaginaire de $\chi_R^{(3)}$, (b) module de $\chi_R^{(3)}$, (c) spectre CARS.

composants qui les entourent se mélangent toujours à elles. Ainsi, il n'existe pas de pixels dans l'image seulement constitué de membranes contrairement aux autres composants qui ont plusieurs pixels non mélangés.

Une fois les spectres et les concentrations des composants générés, les spectres du jeu de données D sont générés en multipliant C et S. Afin d'émuler la variation d'intensité présente au sein d'une cartographie, chaque spectre de D est multiplié par un scalaire aléatoire suivant une loi normale de moyenne 1 et de variance paramétrable. Pour finir, un bruit gaussien est ajouté pour obtenir un SNR prédéterminé. Dans le jeu de données qui sera utilisé dans la suite du manuscrit $\mathcal{V} = 7$, la variance de la loi normale est 0.2 et le SNR est de 15 dB. La figure 3.23 présente un spectre situé au niveau de la membrane entre le noyau et le cytoplasme. Nous pouvons voir que les pics des lipides sont très peu visibles dans le spectre sans bruit et presque indiscernable avec le bruit ; il est probable que les méthodes de résolution de la MCR auront des difficultés à trouver ce composant.

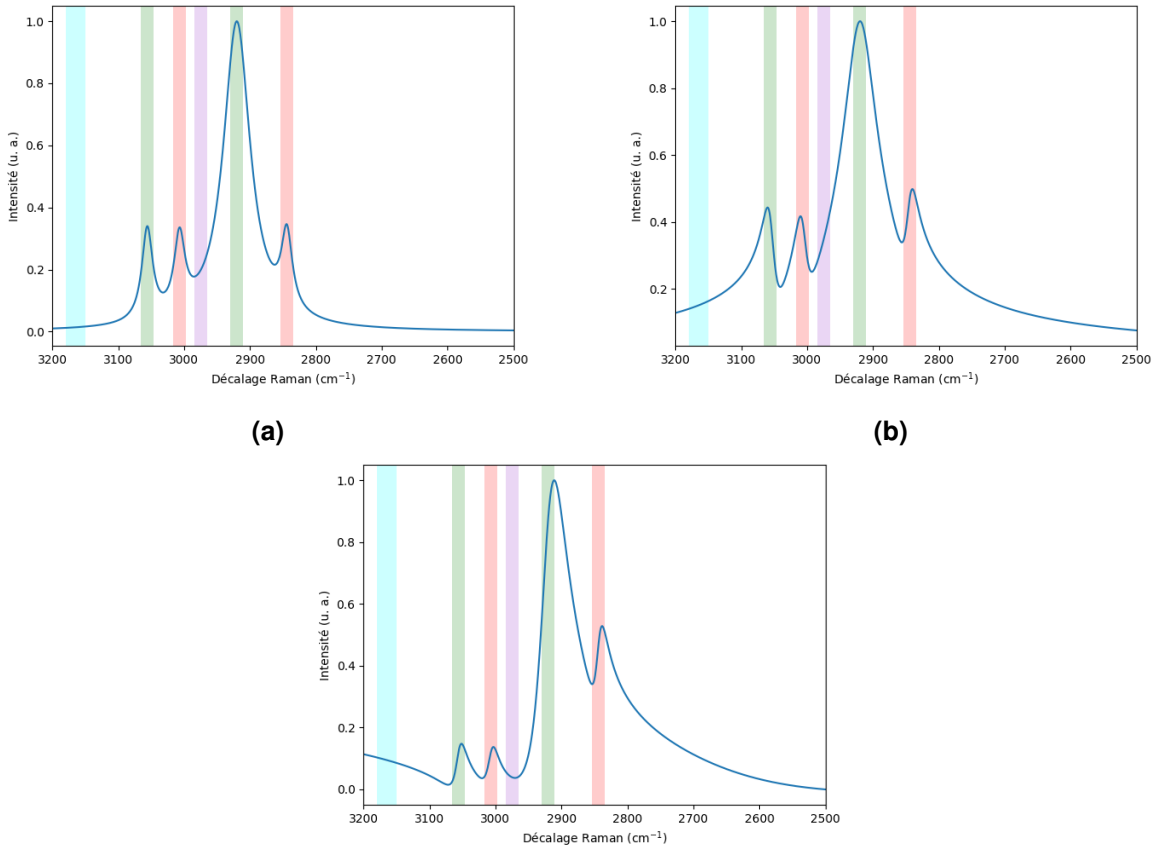


FIGURE 3.21 – Spectres du composant membrane, (a) partie imaginaire de $\chi_R^{(3)}$, (b) module de $\chi_R^{(3)}$, (c) spectre CARS.

3.3.1.3 Application de la MCR-ALS

Afin d'obtenir un résultat de référence, la MCR-ALS utilisant la régression par NNLS avec une contrainte de normalisation des concentrations et initialisé avec la VCA a été appliqué sur le jeu de données. Les spectres et concentrations calculés sont disponibles en figures 3.24 et 3.25. Nous pouvons constater que la membrane n'est pas retrouvée contrairement aux autres composants. Cet échec est imputable à l'absence de spectres ne contenant que de la membrane dans le jeu de données. Nous pouvons aussi constater comme limite des résultats une forte variation d'intensité des concentrations entre deux pixels bien que leurs spectres soient similaires à la variation d'intensité et au bruit près. Spectralement, nous pouvons constater que le spectre associé à la membrane contient du bruit au contraire des trois autres.

Afin d'obtenir une quantification de l'efficacité de la méthode, la SAD des spectres et l'EQM des concentrations calculées par la MCR-ALS par rapport à la réalité du

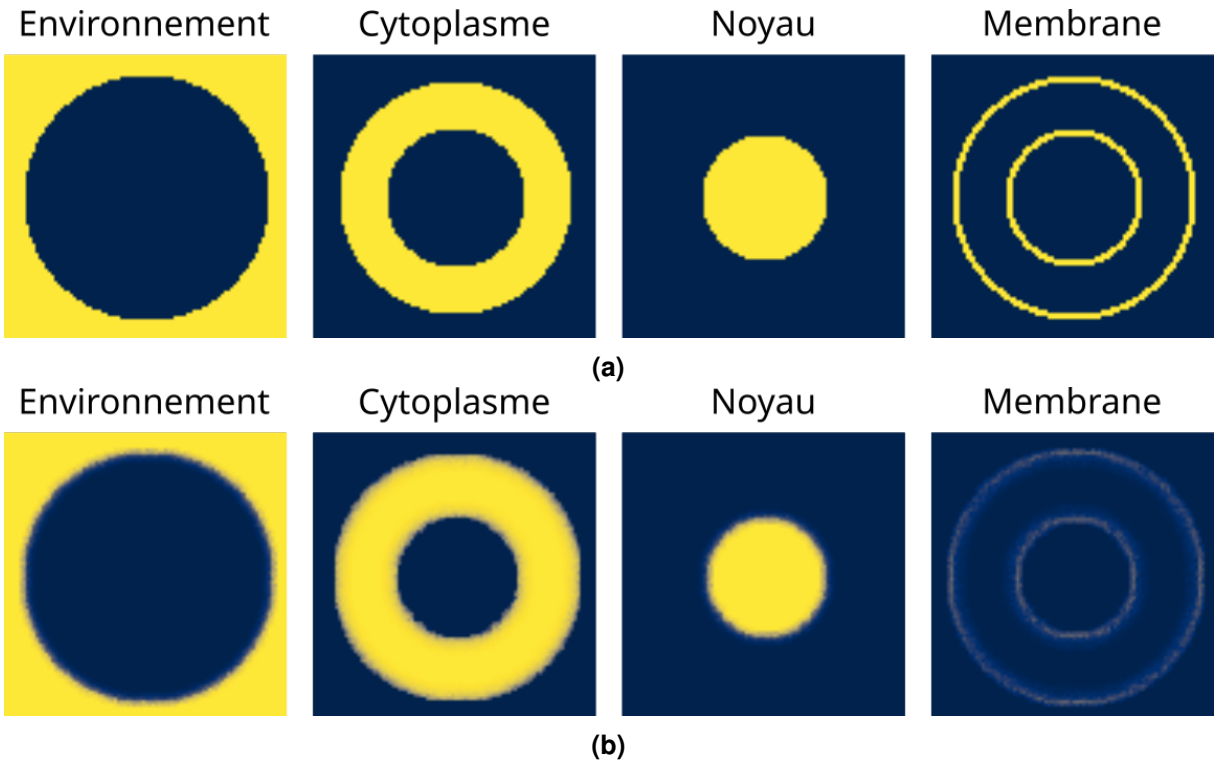


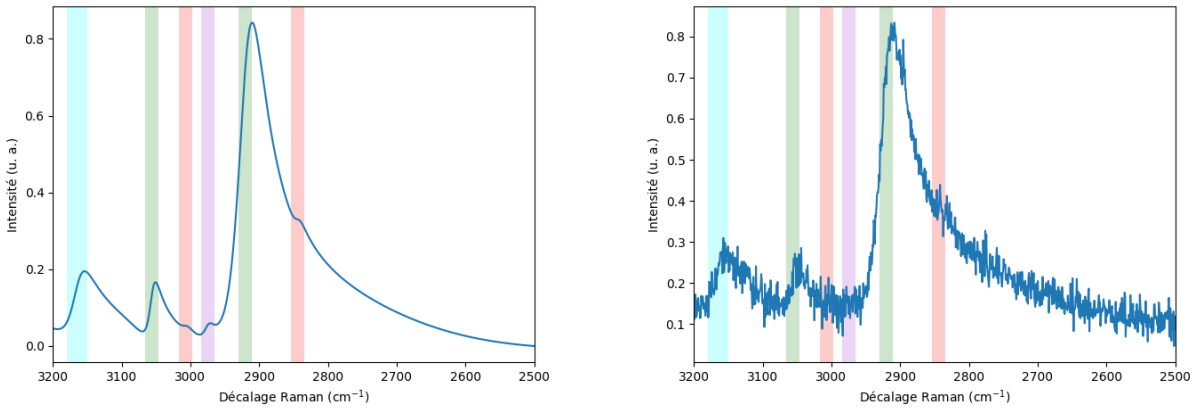
FIGURE 3.22 – Concentrations du jeu de données artificiel, (a) concentrations initiales, (b) concentrations après mélange.

jeu de données ont été calculées. Les résultats sont montrés dans le tableau 3.6. Le spectre le mieux retrouvé est celui du cytoplasme avec une SAD de 0.225, suivi par l'environnement et le noyau avec respectivement 0.252 et 0.2705. Sans surprise, la membrane a une SAD plus importante égale à 0.912. En ce qui concerne les concentrations, le noyau est le mieux reconstruit avec une erreur égale 0.0262. Ensuite viennent l'environnement et le cytoplasme avec respectivement 0.094 et 0.111. Pour finir, la membrane a une EQM de 0.197. La différence d'erreur moins importante sur les concentrations que les spectres peut s'expliquer par le fonctionnement de la MCR-ALS qui optimise pour diminuer l'erreur quadratique.

Métrique	Env.	Cyto.	Noyau	Membrane
SAD	0.252	0.225	0.2705	0.912
EQM	0.094	0.111	0.0262	0.197

TABLEAU 3.6 – SAD et EQM calculées avec les résultats de la MCR-ALS. Env. signifie environnement et cyto. signifie cytoplasme.

Maintenant que le jeu de données artificiel a été présenté et un résultat de référence obtenu avec la MCR-ALS, les différentes expérimentations pour développer



(a) **(b)**

FIGURE 3.23 – Exemple de spectre mélangeant le noyau, une membrane et le cytoplasme. (a) Spectre CARS sans bruit, (b) spectre CARS bruité.

un AE pour appliquer la MCR vont pouvoir être effectuées.

3.3.2 Application de prétraitements aux données

Pour évaluer quels prétraitements sont nécessaires, le modèle est entraîné avec trois jeux de données différents représentant les mêmes données plus ou moins traitées : le premier jeu de données correspond aux données brutes, le deuxième aux données débruitées. Le débruitage consiste en la projection sur les K premiers vecteurs propres de l'ACP des données non centrées réduites, puis à la reconstruction des spectres à partir des données projetées.

Le spectre du pixel utilisé pour la figure 3.23 débruité est présenté dans la figure 3.26. Ainsi la SAD entre le spectre débruité et le spectre sans bruit est de 0.292 contre 0.304 pour la SAD entre le spectre bruité et le spectre sans bruit.

3.3.2.1 Application aux données brutes

La figure 3.27 contient les spectres calculés sur les données brutes. À l'issue des entraînements, 10 spectres ont été associés à l'environnement, 20 au cytoplasme, 5 au noyau et 5 aux membranes. Ce déséquilibre est marqueur d'une difficulté du modèle à trouver les spectres réels. De plus, nous pouvons voir sur la figure 3.27, qu'au sein d'un même composant, de nombreuses signatures spectrales différentes sont présentes. La variation la plus importante est sur le cytoplasme ou deux échelles de spectres sont présentes. En effet, des spectres plus bruités et avec une variation des intensités

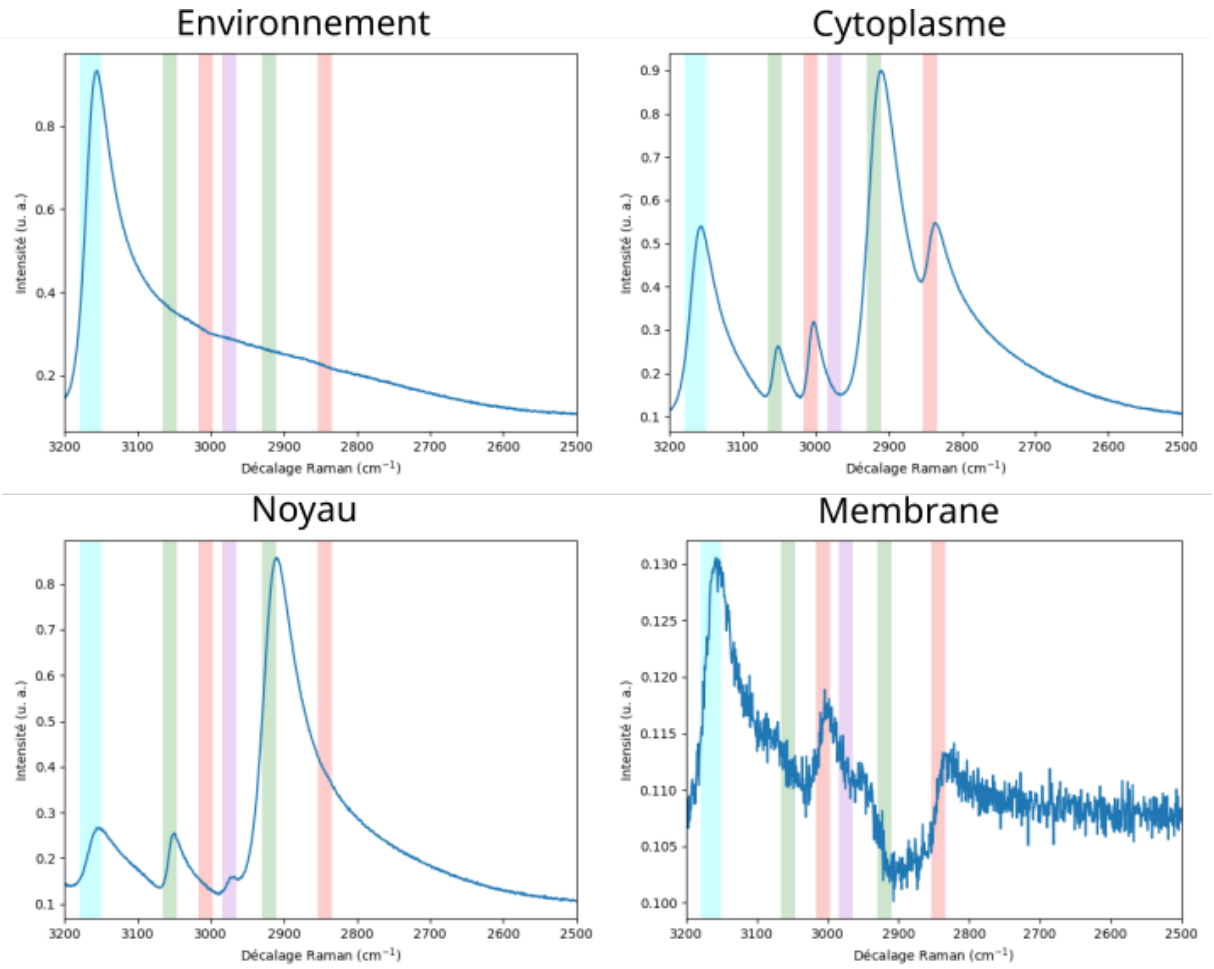


FIGURE 3.24 – Les spectres calculés par la MCR-ALS.

moins importantes sont présents. Les pics des différents composants n'ont pas les bonnes intensités les uns par rapport aux autres ou n'ont pas les bonnes largeurs de bande. De plus, la contrainte de non-négativité n'étant pas intégrée dans le modèle, les spectres possèdent des valeurs négatives.

Les concentrations moyennes présentées en figure 3.28 montrent une moins grande variation de concentrations pixels à pixels que dans le cas de la MCR-ALS. De plus, les concentrations des membranes ne forment pas un bruit sur toute l'image mais illumine légèrement le cytoplasme et l'environnement. Cependant, ces concentrations sont moins sélectives, elles présentent trois niveaux d'intensité selon si le pixel se situe dans l'environnement, le cytoplasme ou le noyau. Cet effet est imputable à la variabilité des spectres calculées dans les différents entraînements résultant en des composants aussi différents sur les concentrations.

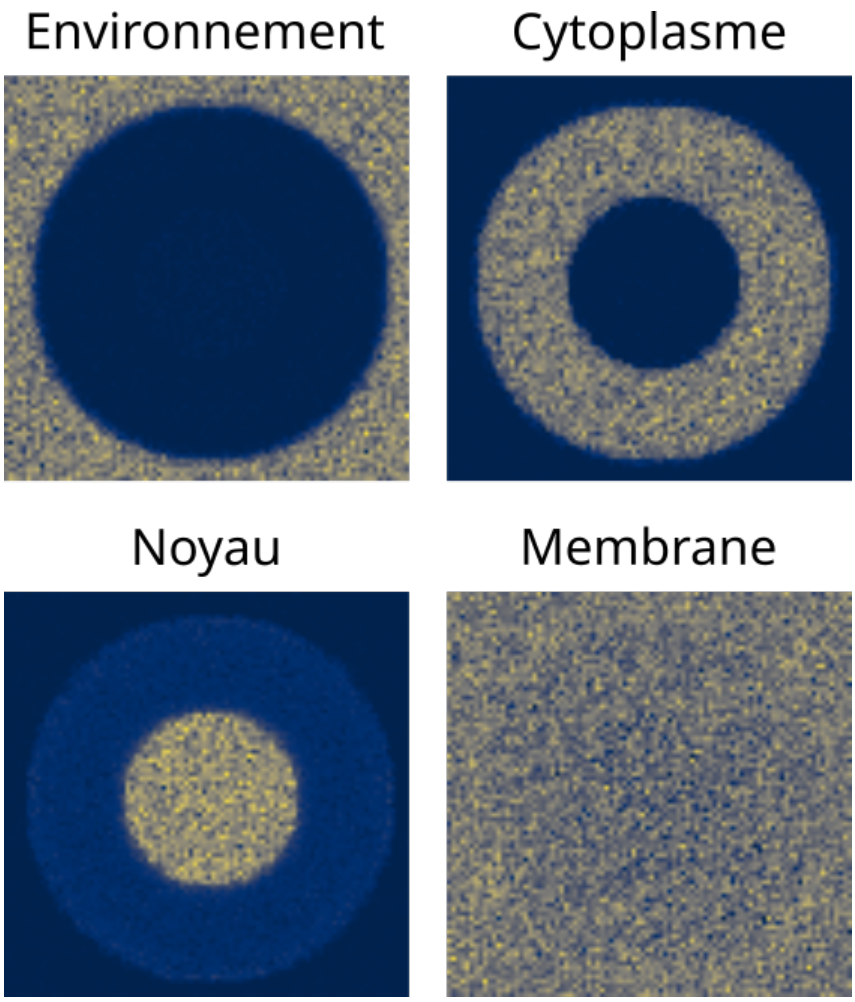


FIGURE 3.25 – Les concentrations calculées par la MCR-ALS.

3.3.2.2 Application aux données débruitées

La figure 3.29 montre les spectres calculés sur les données débruitées et la figure 3.30 montre les concentrations moyennes. Une analyse similaire à celle faite sur les données bruitées peut être faite. Une importante variabilité est présente au sein des résultats ce qui limite l'analyse qui peut être faite des composants et le débruitage ne semble pas aider à améliorer la qualité des spectres et concentrations trouvés.

Le tableau 3.7 présente les moyennes et écart-types des SAD et EQM obtenus entre les composants calculés et les composants réels du jeu de données. Les métriques confirment la forte variabilité pouvant aller jusqu'à presque la valeur de la moyenne. C'est le cas, par exemple, de l'environnement sur les données débruitées qui a un EQM moyen de 0.126 avec un écart-type 0.123. Nous pouvons cependant constater que l'approche par réseau de neurones obtient de meilleurs résultats que la

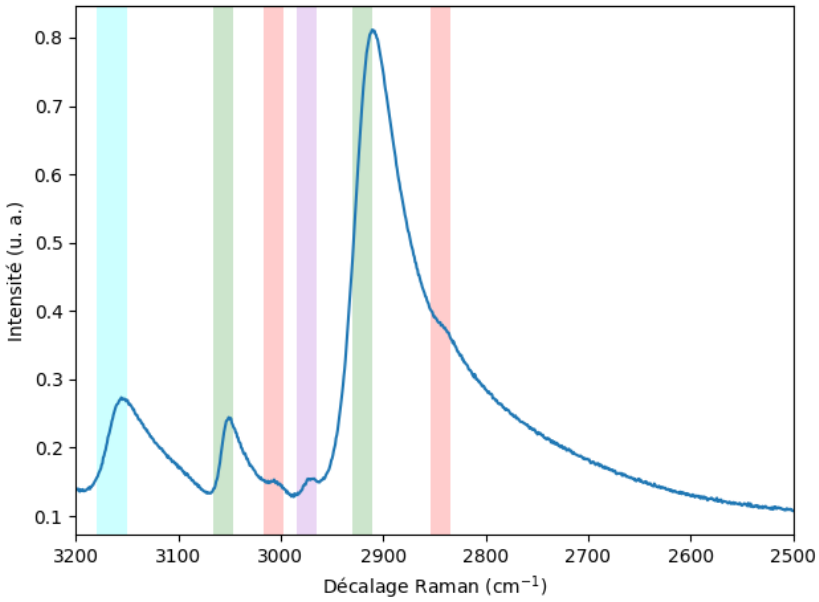


FIGURE 3.26 – spectre mélangeant le noyau, une membrane et le cytoplasme débruité.

MCR-ALS sur le composant membrane. Si l'on étudie l'évolution des métriques entre les deux jeux de données, celles-ci sont très similaires. Le prétraitement des données n'est pas nécessaire pour un niveau de SNR de 15 dB. Il est donc décidé de continuer les expérimentations sur les données bruitées.

Données	Métrique	Env.	Cyto.	Noyau	Membrane
Bruit	SAD	0.266 ± 0.100	0.582 ± 0.218	0.160 ± 0.073	0.510 ± 0.157
	EQM	0.105 ± 0.101	0.275 ± 0.077	0.083 ± 0.037	0.037 ± 0.030
Débruit.	SAD	0.260 ± 0.164	0.616 ± 0.278	0.200 ± 0.036	0.502 ± 0.125
	EQM	0.126 ± 0.123	0.303 ± 0.076	0.0759 ± 0.025	0.044 ± 0.037

TABLEAU 3.7 – Moyennes et écart-types des SAD et EQM calculées à partir des données brutes et débruitées sur 10 entraînements. Env. signifie environnement, cyto. signifie cytoplasme et débruit. signifie débruitées.

3.3.3 Choix de la fonction de coût

Pour étudier l'impact de la fonction de coût sur l'apprentissage, 10 entraînements utilisant l'EQM plutôt que la SAD sont effectués. Les spectres obtenus à l'issue des entraînements sont disponibles en figure 3.31 et les concentrations moyennes en figure 3.32. 20 spectres sont associés à l'environnement, 5 au cytoplasme, 6 au noyau et 9 aux membranes. La variabilité des spectres est encore plus importante que pour les résultats

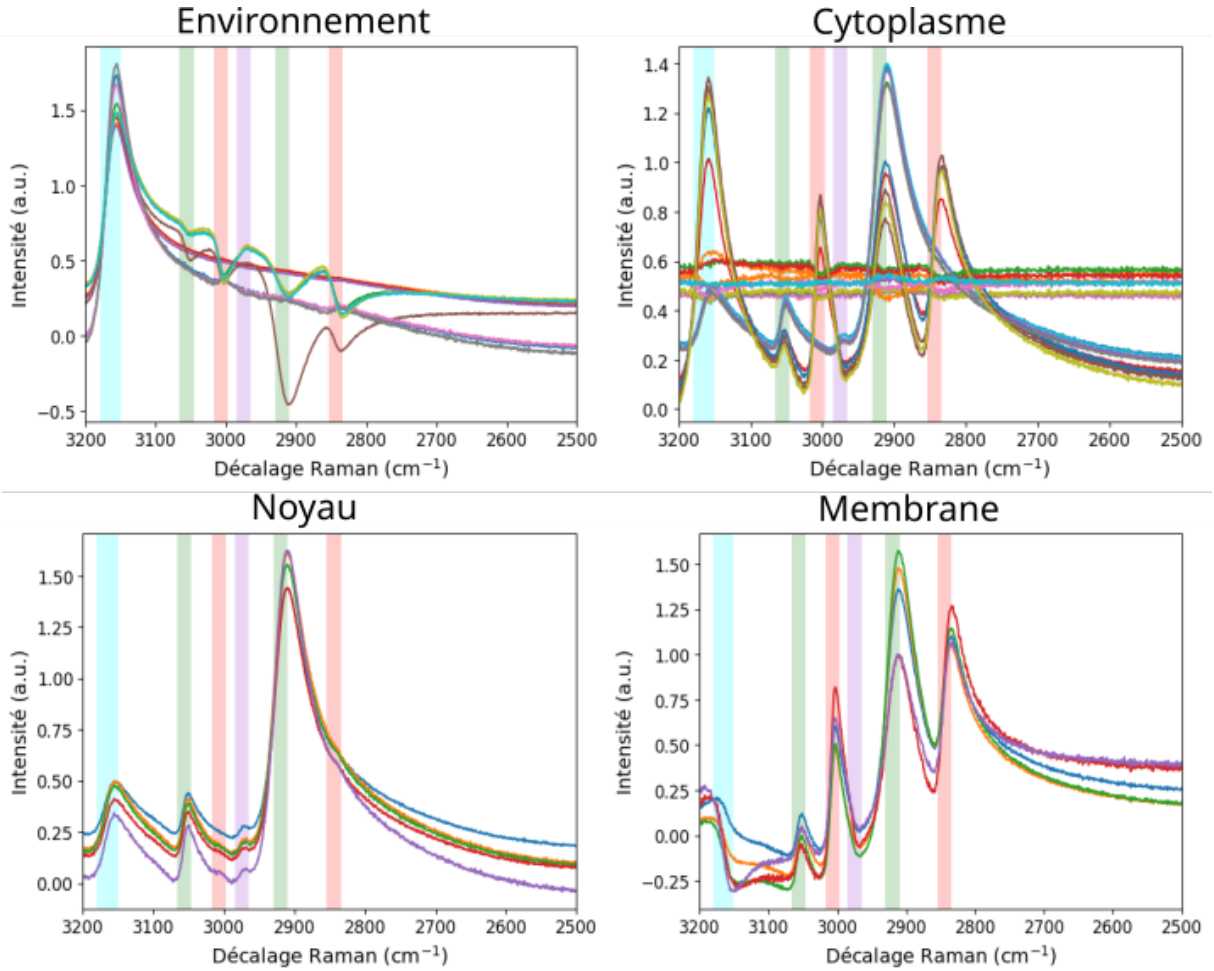


FIGURE 3.27 – Les spectres calculés par 10 entraînements sur les données bruitées.

obtenus avec la SAD et les concentrations discriminent peu des régions de l'image.

les moyennes et écart-types des SAD et EQM obtenus présentés en tableau 3.8 montrent qu'à la fois les spectres et le concentrations sont moins proches de la réalité que ceux calculés en utilisant la SAD en fonction de coût détaillés dans le tableau 3.7. Par exemple, la SAD moyenne en utilisant l'EQM de l'environnement est de 0.599 ± 0.250 contre 0.266 ± 0.100 en utilisant la SAD.

Métrique	Env.	Cyto.	Noyau	Membrane
SAD	0.599 ± 0.250	0.611 ± 0.441	0.414 ± 0.221	0.512 ± 0.354
EQM	0.229 ± 0.046	0.269 ± 0.025	0.083 ± 0.025	0.054 ± 0.037

TABLEAU 3.8 – Moyennes et écart-types des SAD et EQM calculées en utilisant l'EQM comme fonction de coût sur 10 entraînements. Env. signifie environnement, cyto. signifie cytoplasme.

Les résultats obtenus avec l'EQM indiquent que cette métrique n'est pas appropriée pour la MCR appliquée à des données CARS par AE. La SAD est donc utilisée

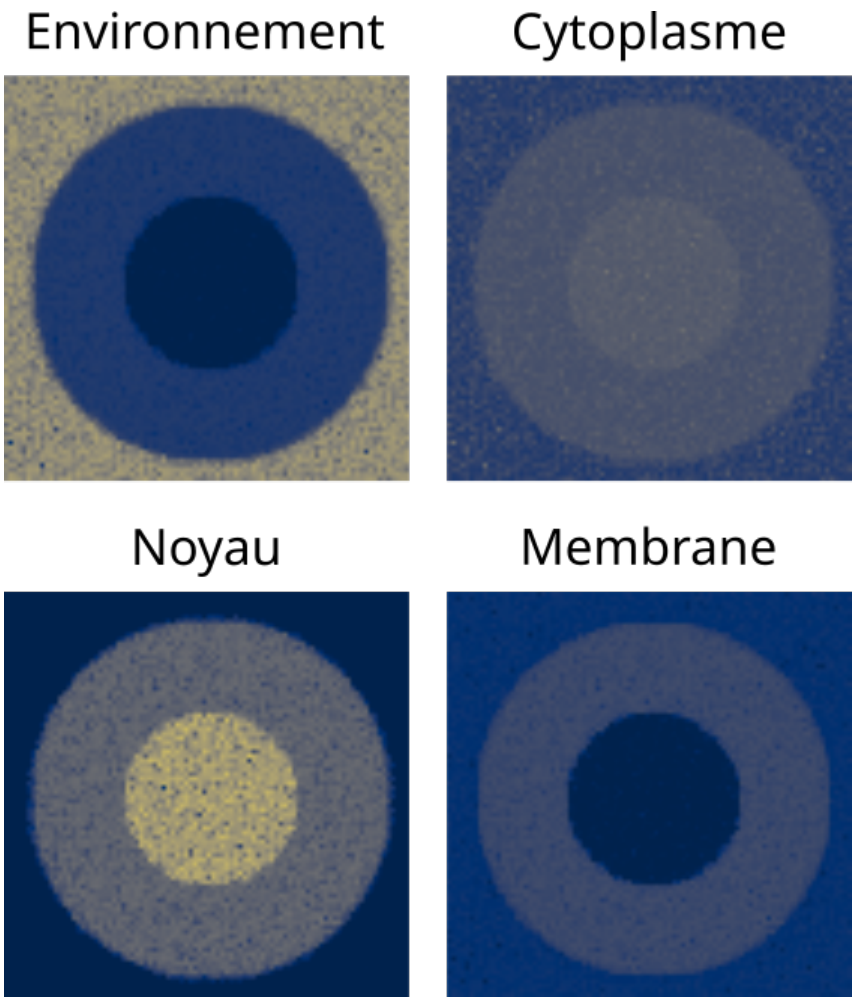


FIGURE 3.28 – Concentrations moyennes calculées à partir de 10 entraînements sur les données bruitées.

pour les expériences suivantes.

3.3.4 Initialisation du décodeur

Afin d'estimer l'impact de l'initialisation du décodeur sur les résultats, 10 entraînements en initialisant le décodeur avec les spectres calculés par la VCA sont effectués. Les spectres obtenus sont disponibles en figure 3.33. Contrairement aux entraînements précédents seulement 3 composants sont trouvés à l'issue des entraînements. Cela signifie que plusieurs dimensions des données encodées sont associés au même composant lors de la classification des spectres par l'algorithme du plus proche voisin. 11 spectres sont associés à l'environnement, 24 au cytoplasme et 5 aux membranes. C'est donc le cytoplasme qui est trouvé en double et aucun spectre n'est plus proche

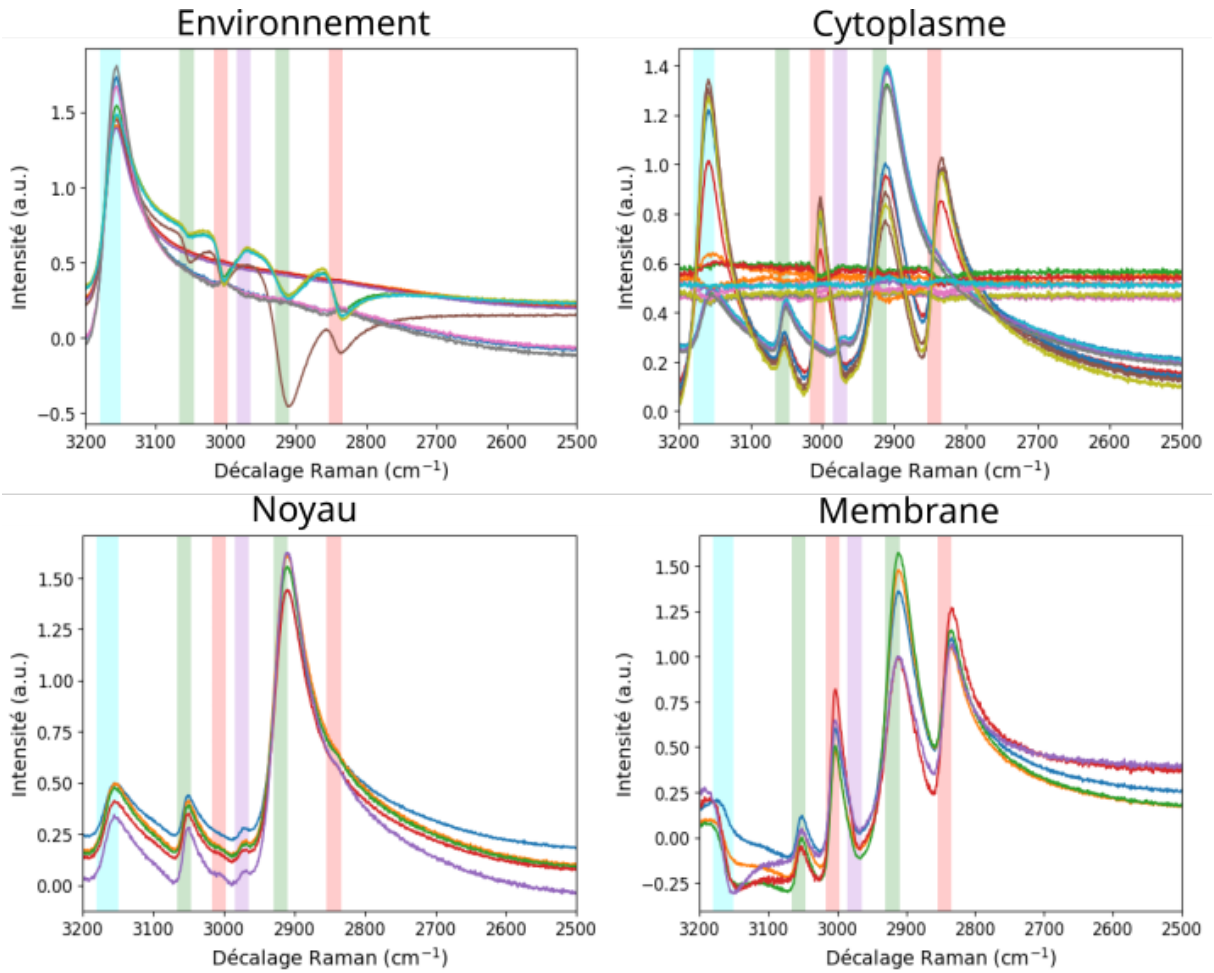


FIGURE 3.29 – Les spectres calculés par 10 entraînements sur les données débruitées.

du noyau que des autres. Cependant, la variabilité des spectres trouvés est inférieure à celle des paramétrages précédents. L'environnement contient cependant deux spectres qui n'ont rien à voir avec le spectre réel mettant en évidence l'échec de certains entraînements. Le cytoplasme possède la plus grande variabilité de résultats ce qui est cohérent avec le nombre important de spectres lui étant associé.

Les moyennes des concentrations sont présentées en figure 3.34. Les concentrations moyennes de l'environnement et des membranes sont similaires à celles obtenues sans initialisation. Pour le cytoplasme, celle-ci présente au final une plus importante concentration dans le noyau. Cela s'explique par le fait que le spectre du noyau présente des pics de protéines similaires au cytoplasme. La méthode tend alors à associer les deux composants ensembles. Aussi, le composant membrane est localisé dans tout le cytoplasme et non seulement à ses frontières.

Le tableau 3.9 présente les métriques obtenues à partir des différents entraî-

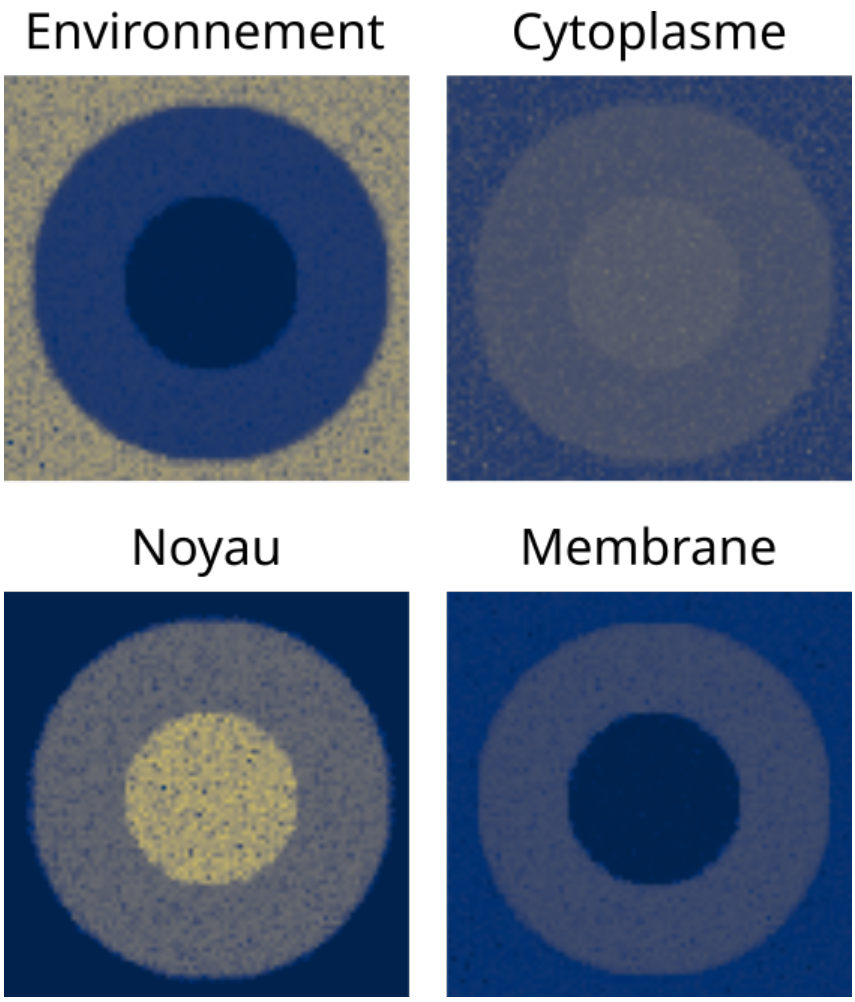


FIGURE 3.30 – Concentrations moyennes calculées à partir de 10 entraînements sur les données débruitées.

nements. Si nous comparons au tableau 3.7, nous pouvons constater que seule la métrique membrane obtient de meilleurs résultats. En raison de la perte d'un composant lorsque le décodeur est initialisé avec la VCA, les expériences suivantes seront menées avec le décodeur initialisé aléatoirement.

Métrique	Env.	Cyto.	Noyau	Membrane
SAD	0.375 ± 0.168	0.487 ± 0.236		0.344 ± 0.043
EQM	0.084 ± 0.067	0.272 ± 0.098		0.026 ± 0.003

TABLEAU 3.9 – Moyennes et écart-types des SAD et EQM calculées en initialisant le décodeur avec la VCA sur 10 entraînements. Env. signifie environnement, cyto. signifie cytoplasme.

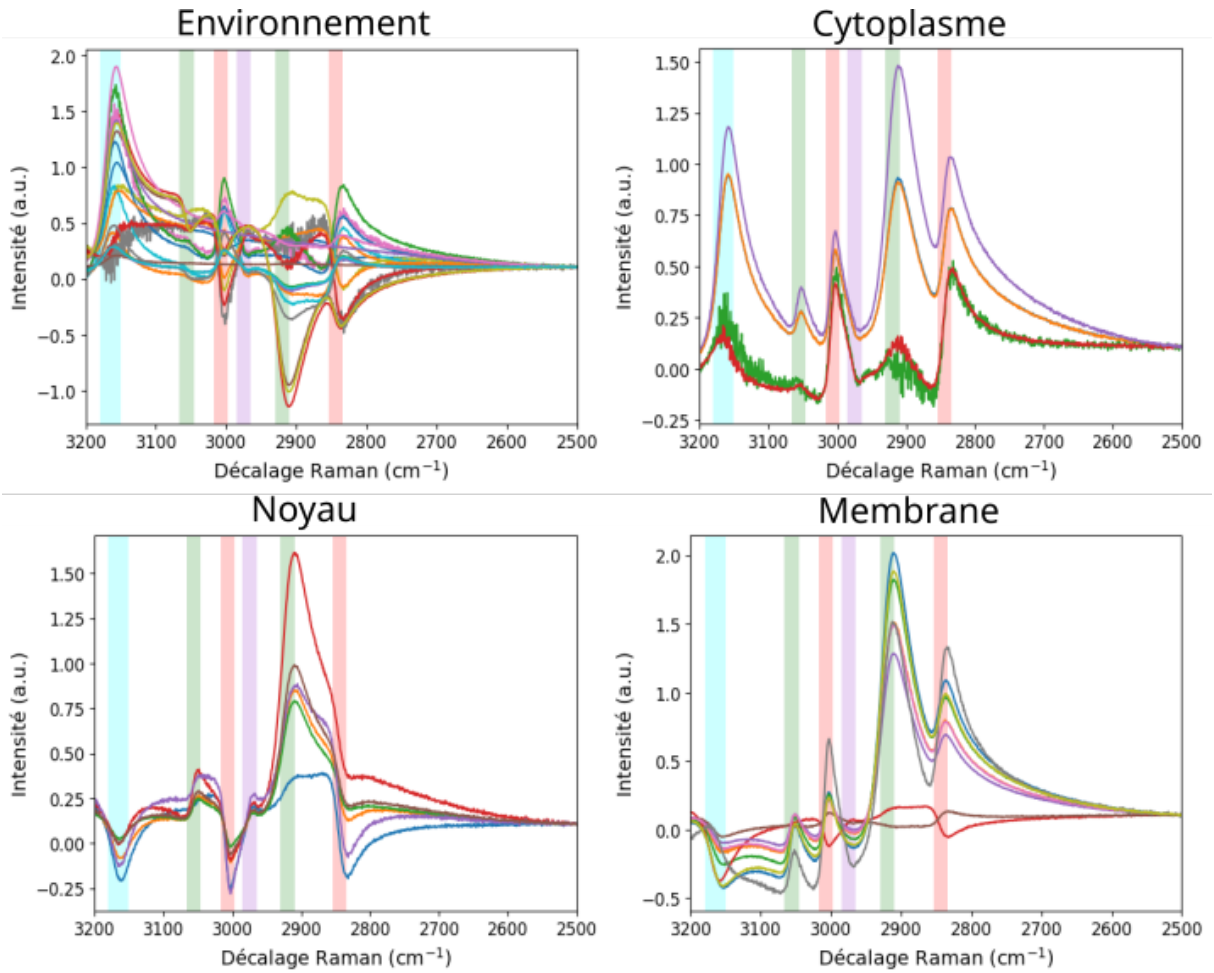


FIGURE 3.31 – Les spectres calculés par 10 entraînements utilisant l'EQM comme fonction de coût.

3.3.5 Intégration de la non-négativité des spectres

La non-négativité des spectres est abordée avec deux approches différentes : l'application d'une fonction d'activation non-négative aux paramètres de l'unique couche dense du décodeur ainsi que l'utilisation d'un décodeur non linéaire.

3.3.5.1 Fonction d'activation non-négative

Deux fonctions ont été appliquées aux paramètres du décodeur pour forcer la non-négativité des spectres : la fonction absolue et la fonction ReLU. Ces fonctions sont appliquées directement aux paramètres du décodeur avant la multiplication avec l'espace latent. Bien que non conventionnelle, cette approche permet d'assurer la non-négativité dans la structure du modèle par l'application de fonctions différentiables

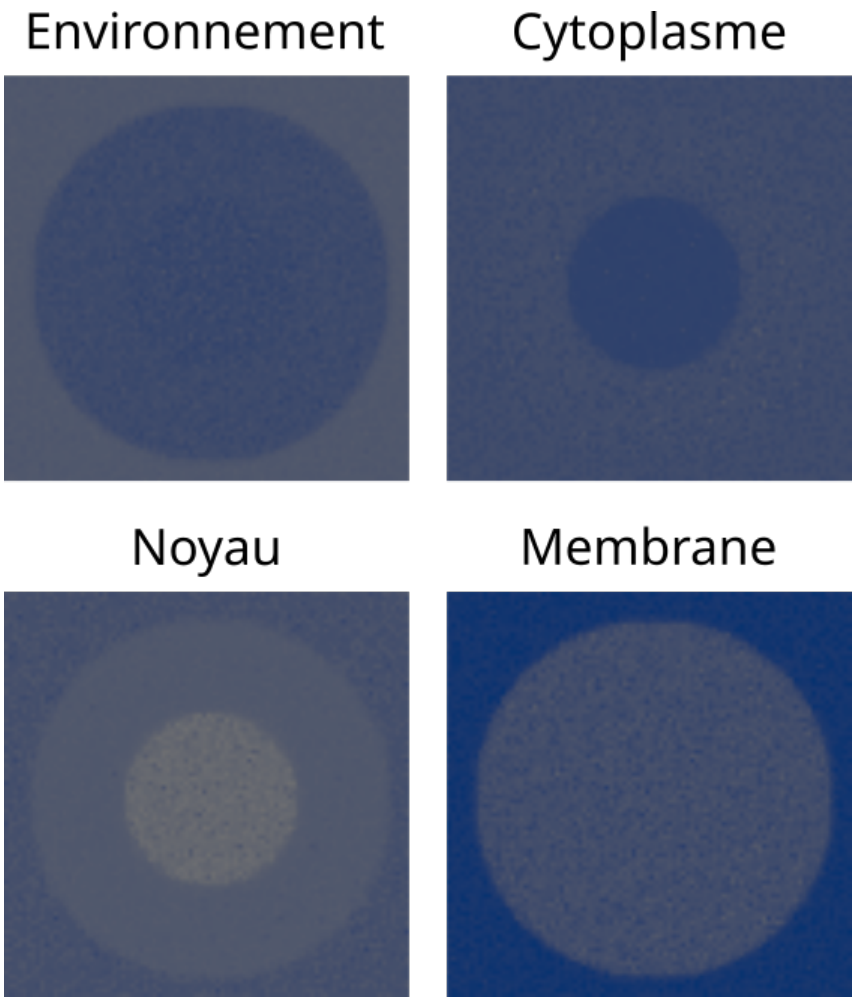


FIGURE 3.32 – Concentrations moyennes calculées à partir de 10 entraînements utilisant l'EQM comme fonction de coût.

n'empêchant pas l'apprentissage des paramètres par descente de gradient. La reconstruction des spectres d'entrées devient alors :

$$\hat{X} = C\vartheta(W)^T, \tag{3.11}$$

avec ϑ, la fonction d'activation et W, les paramètres du décodeur. Ainsi, $S = \vartheta(W)$.

La fonction absolue utilisée comme contrainte de non-négativité permet d'obtenir les résultats de la figure 3.35. 11 spectres sont associés à l'environnement, 22 au cytoplasme, 4 au noyau et 3 aux membranes. Le déséquilibre entre les différents composants est donc toujours fortement présent. Des variations importantes sont présentes au sein de l'environnement et le cytoplasme. Un faible pic d'ADN est bien observable dans les spectres du noyau. Nous pouvons aussi constater que la contrainte

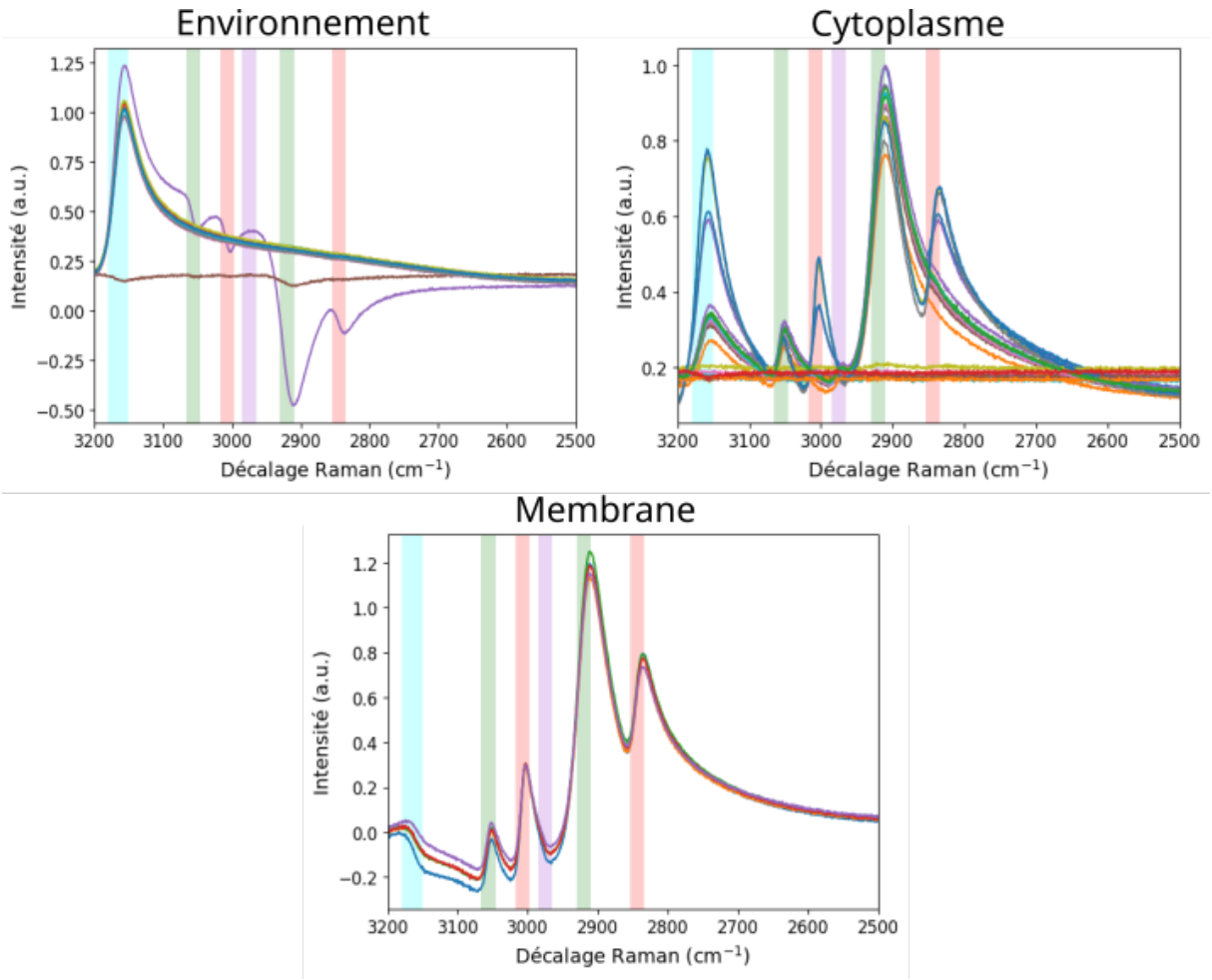


FIGURE 3.33 – Les spectres calculés par 10 entraînements initialisés par la VCA.

de non-négativité est bien respectée.

Les moyennes des concentrations obtenues sont présentées en figure 3.36. L'environnement est bien mis en avant dans ces concentrations. Le cytoplasme, par contre, illumine tout l'intérieur de la cellule. Ce phénomène peut être imputé aux différents types de spectres différents catégorisés comme cytoplasme. Les concentrations du noyau sont principalement situées dans le noyau mais aussi en plus faible intensité dans le cytoplasme. Les concentrations des membranes sont réparties de manière homogène dans le cytoplasme.

Les métriques de SAD et d'EQM par rapport à la vérité du jeu de données sont présentées en 3.10. Celles-ci sont très proches de celles obtenues sans la contrainte de non-négativité. Ainsi, l'application de la fonction absolue permet de respecter la contrainte de non-négativité sans impacter la qualité des résultats que ça soit vers une amélioration ou une dégradation de ceux-ci.

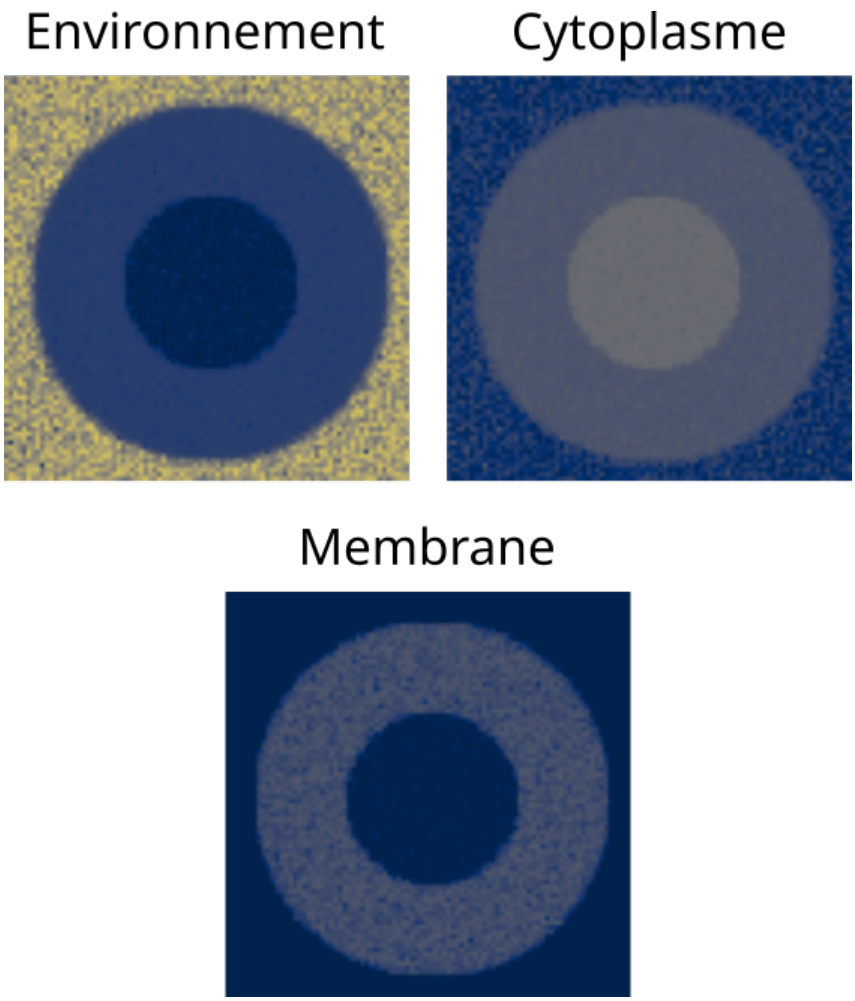


FIGURE 3.34 – Moyennes des concentrations calculées par 10 entraînements initialisés par la VCA.

Métrique	Env.	Cyto.	Noyau	Membrane
SAD	0.276 ± 0.150	0.557 ± 0.220	0.174 ± 0.067	0.458 ± 0.131
EQM	0.085 ± 0.058	0.256 ± 0.080	0.079 ± 0.021	0.033 ± 0.011

TABLEAU 3.10 – Moyennes et écart-types des SAD et EQM calculées en appliquant la fonction absolue aux paramètres du décodeur sur 10 entraînements. Env. signifie environnement, cyto. signifie cytoplasme.

L'application de la fonction ReLU aux paramètres du décodeur produit les spectres de la figure 3.37. Contrairement aux résultats obtenus avec la fonction absolue, la fonction ReLU diminue fortement la qualité des spectres. Le bruit et la variation des spectres très importants excluent le choix de la fonction ReLU pour appliquer la contrainte de non-négativité.

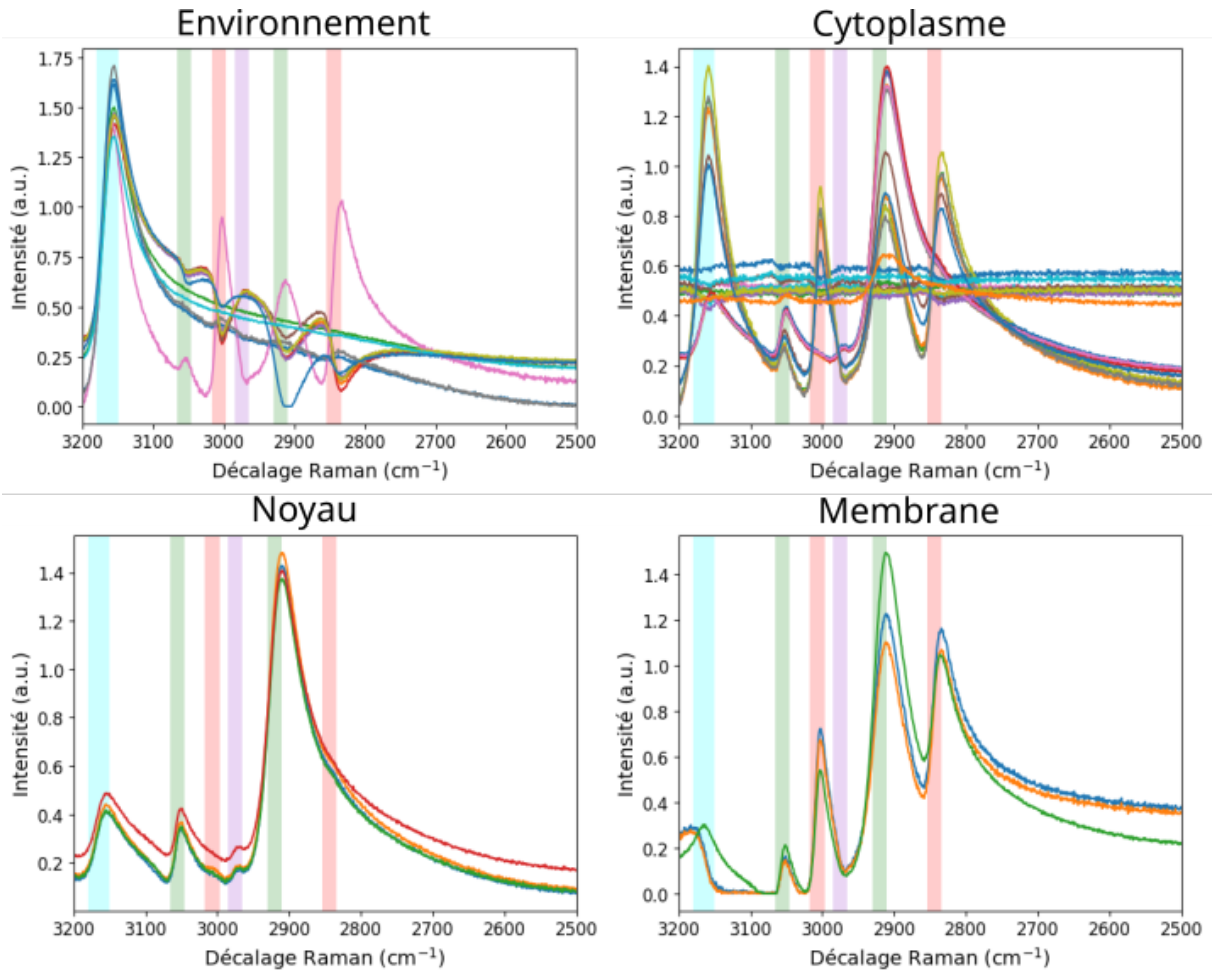


FIGURE 3.35 – Les spectres calculés par 10 entraînements avec la fonction absolue.

3.3.5.2 Utilisation d'un décodeur non linéaire

La seconde approche utilisée pour appliquer la contrainte de non-négativité aux spectres est l'utilisation d'un décodeur non linéaire. Pour ce faire, il nous faut repartir de la formulation générale de la MCR :

$$D = \Phi(C, S), \tag{3.12}$$

Φ étant une fonction combinant C et S pour reconstruire les données d'entrée D. En utilisant un AE, C peut être calculé par l'encodeur $\mathcal{E}$. Le décodeur $\mathcal{D}$ permet alors de reconstruire les données d'entrée une fois appliquée à C. La MCR devient alors :

$$D = \mathcal{D}(C). \tag{3.13}$$

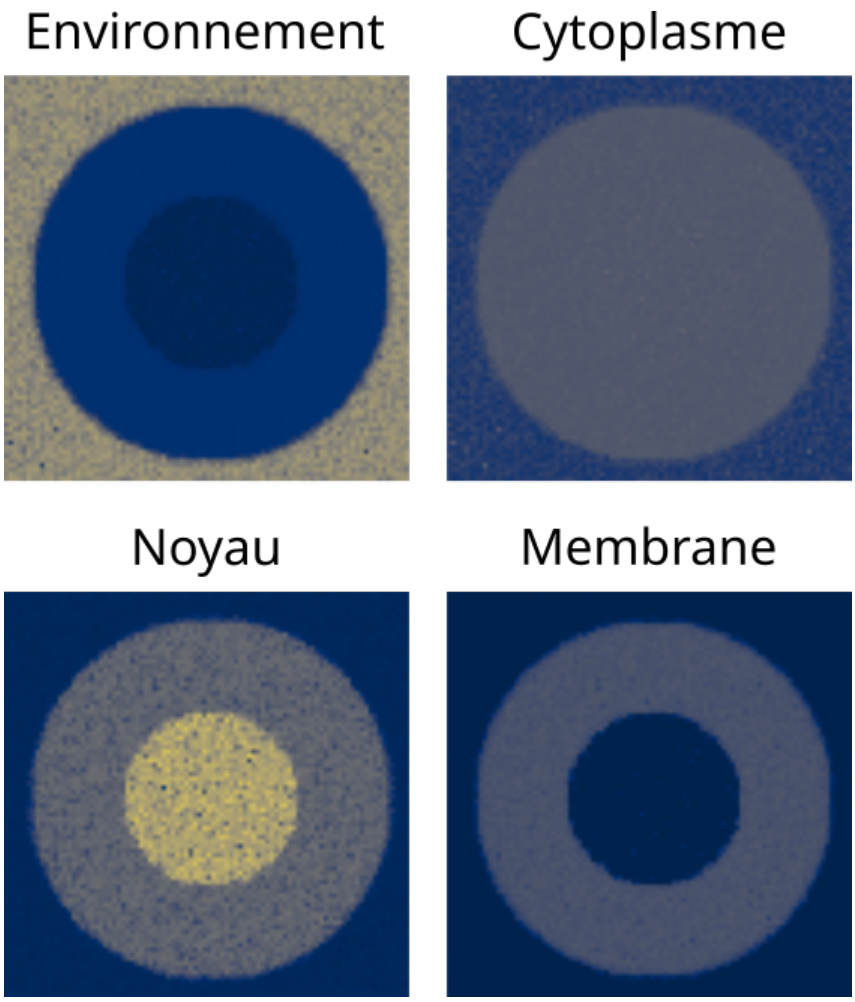


FIGURE 3.36 – Moyennes des concentrations calculées par 10 entraînements avec la fonction absolue.

Si $\mathcal{D}$ est un décodeur purement spectral, il est possible d'extraire les spectres appris par C en faisant décoder la matrice identité I_K. En effet, chaque ligne de la matrice correspond à un pixel « pur », un pixel contenant un seul composant. En décodant le pixel, le spectre du composant est alors obtenu.

Bloc	Couche	Descr.	Norm. lots	Abandon	Activation.
Encodeur	Dense	16	✓	0.2	ReLU
	Dense	K			*Softmax*
Décodeur	Dense	16	✓	0.2	ReLU
	Dense	N			Sigmoïde

TABLEAU 3.11 – Architecture du modèle avec décodeur non linéaire.

Pour implémenter le décodeur non linéaire, le modèle présenté en section 3.3 est modifié pour devenir le modèle présenté en table 3.11. Le modèle est symétrique, le

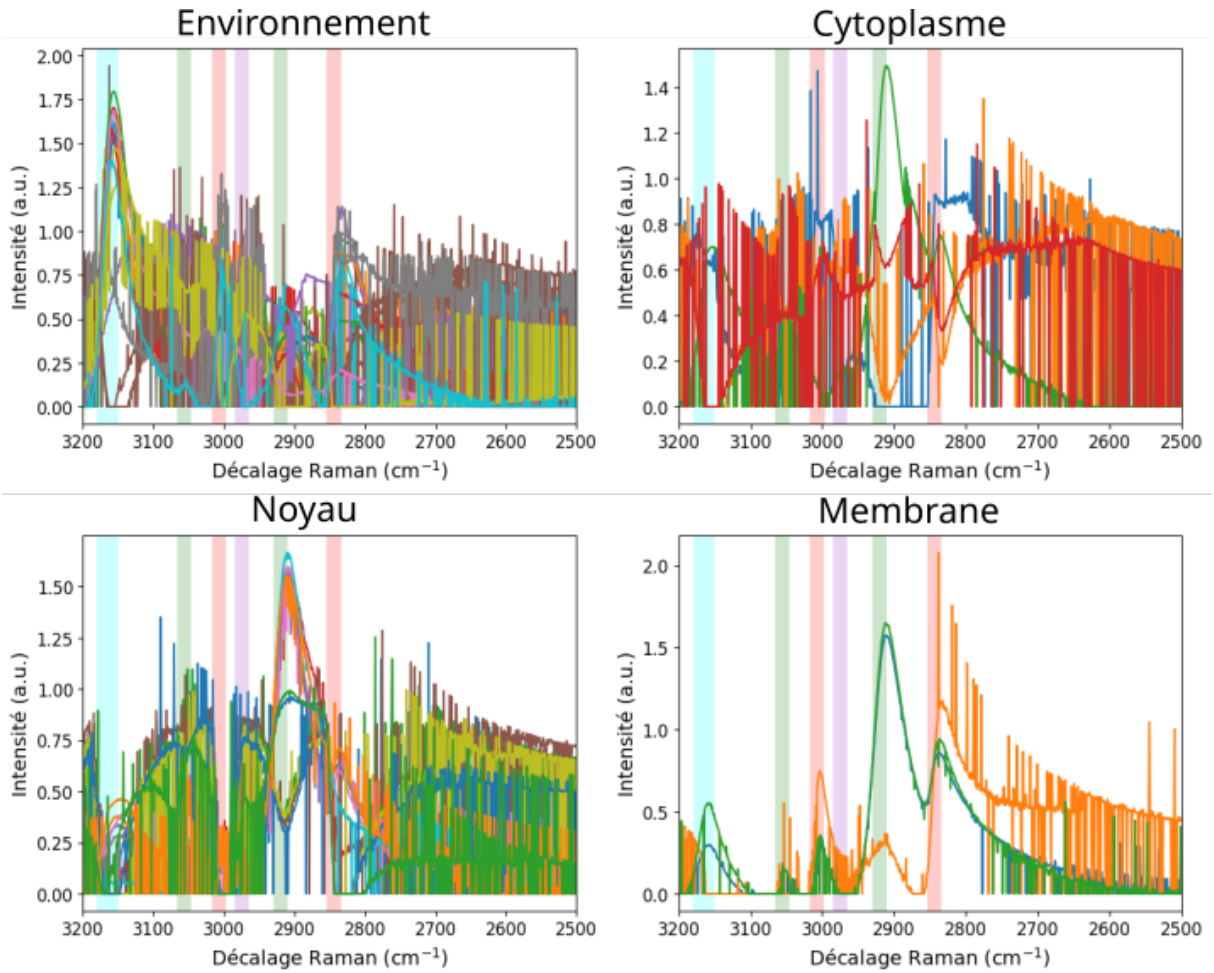


FIGURE 3.37 – Les spectres calculés par 10 entraînements avec la fonction ReLU.

décodeur contient autant de couches que l'encodeur et le nombre de descripteurs est inversé. Ce choix permet de limiter la complexité du modèle. De plus, il n'y a, *a priori*, pas d'arguments en faveur d'un décodeur plus complexe que l'encodeur. La première couche du décodeur calcule 16 descripteurs et la deuxième les N canaux spectraux des données d'entrée. Tout comme dans l'encodeur, la couche intermédiaire utilise une normalisation du lot, un abandon avec une probabilité de 0.2 et la fonction d'activation ReLU. La couche finale utilise la fonction d'activation non linéaire sigmoïde afin d'implémenter la contrainte de non-négativité. La fonction sigmoïde est choisie comme fonction d'activation finale car elle permet d'appliquer la contrainte de non-négativité des spectres tout en ne faisant pas disparaître les résultats négatifs contrairement à la fonction ReLU.

Les spectres obtenus par le modèle sont disponibles en figure 3.38. Seulement deux composants sont retrouvés dans les spectres calculés : l'environnement avec 12 spectres et le cytoplasme avec 28 spectres. Les spectres de l'environnement

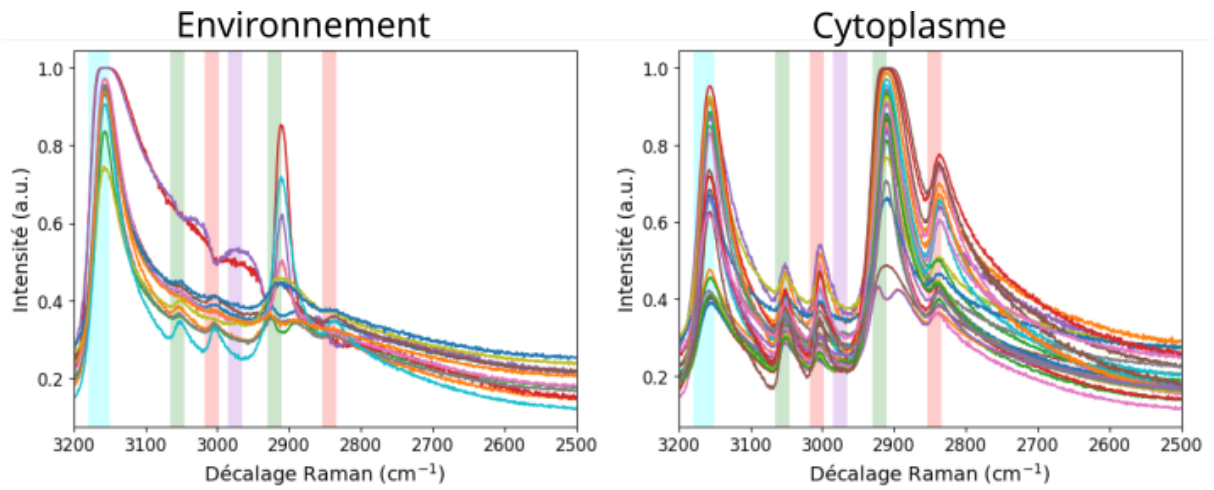


FIGURE 3.38 – Les spectres calculés par 10 entraînements avec un décodeur non linéaire.

contiennent un pic de lipides à 2920 cm^{-1} qui ne devraient pas être présents. Le cytoplasme contient tous les pics possibles à l'exception de celui de l'ADN.

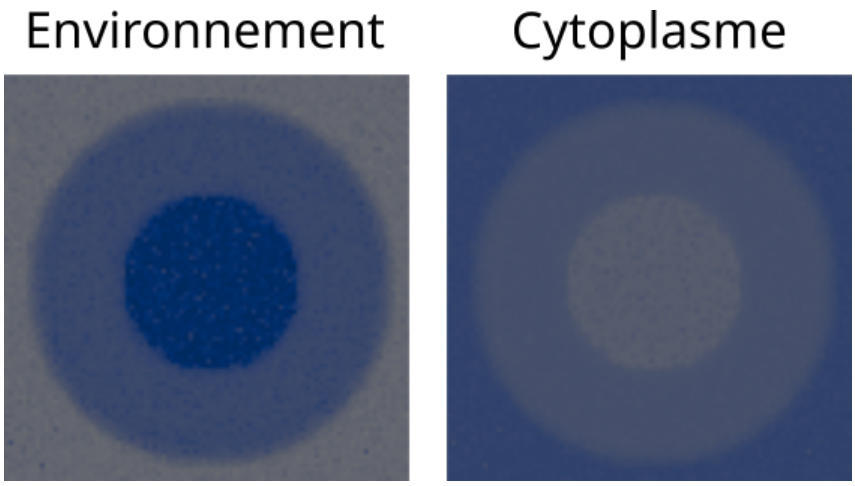


FIGURE 3.39 – Moyennes des concentrations calculées par 10 entraînements avec un décodeur non linéaire.

Les moyennes des concentrations calculées sont montrées en figure 3.39. En raison de leur variation importante, aucun des composants n'atteint la valeur maximale de 1 dans des régions de l'image. De plus, le cytoplasme est légèrement plus intense dans le noyau que dans le cytoplasme.

Si l'on étudie les métriques d'évaluation en tableau 3.12, les valeurs sont similaires à celles obtenues avec le décodeur linéaire avec toutefois une augmentation de l'écart-type sur les SAD. La diminution du nombre de composants trouvés ne rend pas la qualité de ceux extraits meilleure. Une hypothèse sur la difficulté à faire ressortir les composants

biologiques est le manque de contrainte de régularisation sur l'apprentissage du modèle. La dernière expérimentation de cette étude consiste en l'ajout de la contrainte spatiale pour améliorer le calcul des concentrations.

Métrique	Env.	Cyto.	Noyau	Membrane
SAD	0.276 ± 0.150	0.557 ± 0.220		
EQM	0.085 ± 0.058	0.256 ± 0.080		

TABLEAU 3.12 – Moyennes et écart-types des SAD et EQM calculées avec un décodeur non linéaire sur 10 entraînements. Env. signifie environnement, cyto. signifie cytoplasme.

3.3.6 Utilisation d'un encodeur convolutif

Pour intégrer la contrainte spatiale aux concentrations, un encodeur convolutif est utilisé. Le modèle devient alors celui décrit dans le tableau 3.13. Deux couches convolutives sont utilisées. La première utilise un filtre de taille 3×3 avec 16 descripteurs, une normalisation des lots, une probabilité d'abandon de 0,2 et la fonction d'activation ReLU. La deuxième couche utilise un filtre de taille 1×1, ce qui correspond au calcul d'une couche dense, avec K descripteurs et applique la contrainte de non-négativité et normalisation en appliquant la fonction *softmax*. Pour effectuer l'entraînement, des patchs de taille 30×30 espacés de 10 pixels sont extraits de l'image complète, créant un jeu de données de 49 images. La taille des lots est de 8 patchs. Deux décodeurs sont évalués, le décodeur non linéaire de la section 3.3.5.2 et le décodeur aux paramètres contraints par la fonction absolue de la section 3.3.5.1.

Couche	Descr.	Norm. lots	Abandon	Activation.
Conv(3×3)	16	✓	0.2	ReLU
Conv(1×1)	K			*Softmax*

TABLEAU 3.13 – Architecture de l'encodeur convolutif.

3.3.6.1 Avec un décodeur utilisant la fonction absolue

Le décodeur dont les paramètres sont rendus positifs par la fonction absolue ayant obtenu de meilleurs résultats avec l'encodeur spectral, il est évalué en premier lieu pour être combiné à un encodeur spatial. Les spectres obtenus à l'issue de 10 entraînements sont présentés en figure 3.40. Nous pouvons constater que les performances du modèle sont fortement détériorées avec des spectres qui ne sont pas analysables car trop

bruités. Des pics sont tout de même visibles comme celui de l'eau pour l'environnement et le cytoplasme ou encore le pic de protéines à 2920 cm^{-1} pour le cytoplasme, le noyau et la membrane. La complexité d'entraînement ajouté par l'utilisation de couches convolutives ne permet pas d'utiliser une telle contrainte directement aux paramètres. Nous décidons donc d'évaluer le décodeur non linéaire.

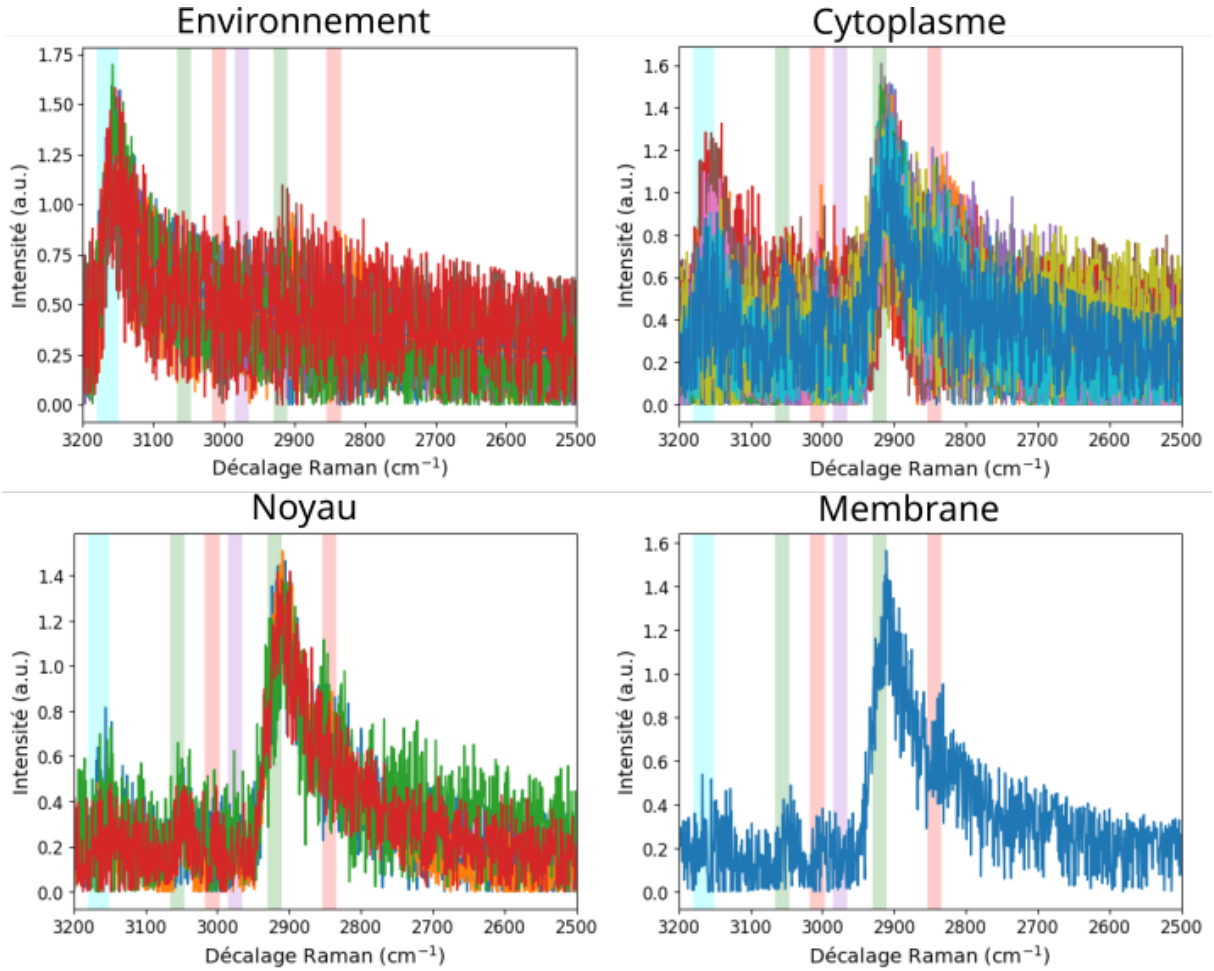


FIGURE 3.40 – Les spectres calculés par 10 entraînements utilisant un encodeur convolutif et un décodeur dont les paramètres sont contraints par la fonction absolue.

3.3.6.2 Avec un décodeur non linéaire

Les spectres calculés par le modèle avec un décodeur non linéaire sont présentés en figure 3.41. À l'instar du modèle avec un encodeur sans convolution, seulement deux composants sont retrouvés. Cependant, ceux-ci varient moins qu'avec l'encodeur purement spectral. 9 spectres sont associés à l'environnement et 31 au cytoplasme. L'environnement possède un pic de protéines protéines à 2920 cm^{-1} qui ne

devrait pas être présent.

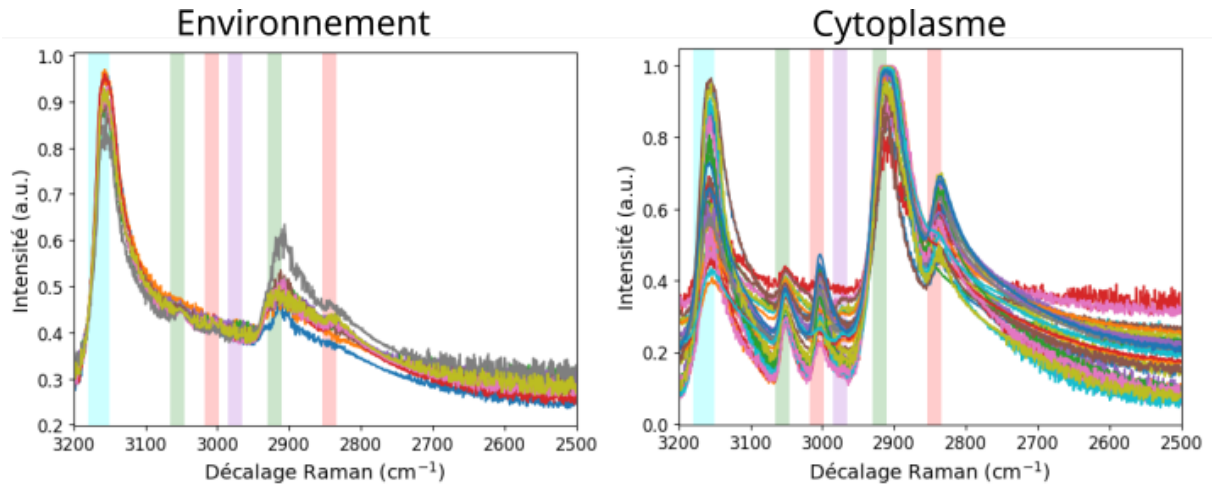


FIGURE 3.41 – Les spectres calculés par 10 entraînements utilisant un encodeur convolutif avec un décodeur non linéaire.

Les moyennes des concentrations sont disponibles en figure 3.42. Les composants représentent principalement l'environnement de la cellule et la cellule dans son entièreté. La présence du pic à 2920 cm^{-1} dans les spectres de l'environnement est due aux pixels à l'intérieur de la cellule qui ont une forte concentration associée à l'environnement. Cette décomposition binaire peut être interprétée comme une amélioration par rapport au modèle spectral puisque ce dernier ne trouvait que deux types de composants sans dissocier complètement la cellule de son environnement. La décomposition binaire effectuée par le modèle peut être imputée à des descripteurs ayant enlevé de l'information vibrationnelle à l'issue de la première couche convolutive. Les métriques présentées dans le tableau 3.14 confirment la forte baisse de variabilité des résultats. En raison du pic à 2920 cm^{-1}, le spectre de l'environnement est moins correct que sans l'encodeur convolutif. La SAD du cytoplasme est améliorée de même que les EQM des deux composants.

Métrique	Env.	Cyto.	Noyau	Membrane
SAD	0.539 ± 0.041	0.330 ± 0.116		
EQM	0.056 ± 0.024	0.246 ± 0.043		

TABLEAU 3.14 – Moyennes et écart-types des SAD et EQM calculées avec un encodeur convolutif et un décodeur non linéaire sur 10 entraînements. Env. signifie environnement, cyto. signifie cytoplasme.

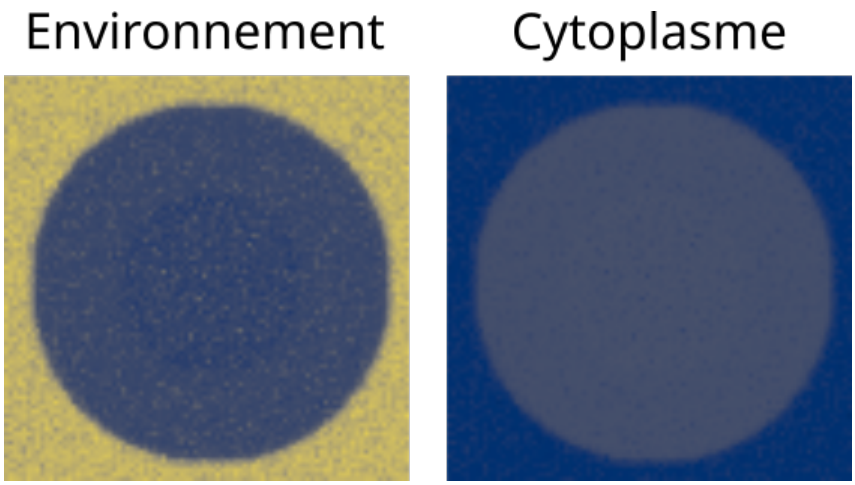


FIGURE 3.42 – Moyennes des concentrations calculés par 10 entraînements utilisant un encodeur convolutif.

3.3.6.3 Application à des donnnées cellulaires

Le modèle avec décodeur non linéaire ayant obtenu les meilleures résultats, il est appliqué à la cellule de référence pour étudier son comportement sur des données réelles. En l'absence de vérité terrain, l'algorithme K-moyennes est utilisé pour classer les spectres en 5 groupes. Les spectres trouvés à l'issue de 10 entraînements sont

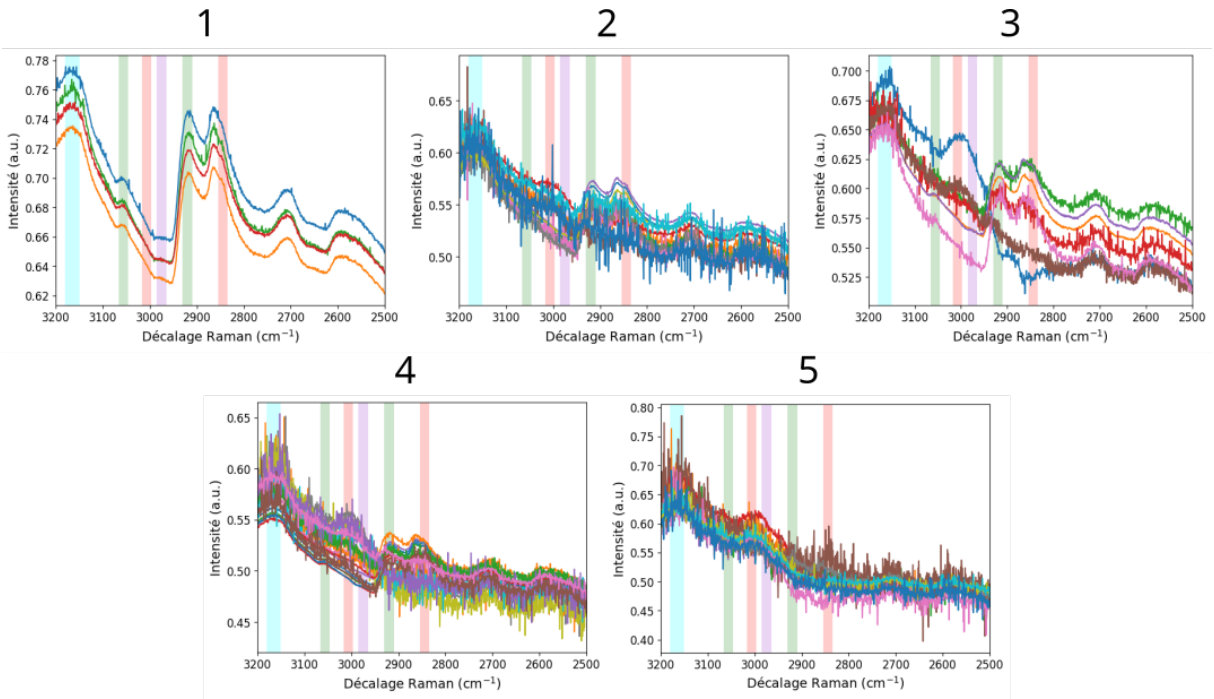


FIGURE 3.43 – Les spectres calculés par 10 entraînements sur la cellule de référence.

présentés en figure 3.43. 4 spectres sont associés au premier composant, 11 au

second, 7 au troisième, 17 au quatrième et 11 au cinquième. Malheureusement, le modèle semble échouer sur ces données. Seul le premier composant se démarque réellement des autres. Bien que les autres spectres soient décalés en intensité, ils partagent les mêmes pics vibrationnels. Le composant signe dans l'eau ainsi que dans les protéines à 3056 et 2920 cm^{-1}. Un pic de lipides est aussi présent à 2844 cm^{-1}. Les autres composants sont très bruités, pour les composants 2, 4 et 5, ou ont une trop importante variabilité pour le composant 3.

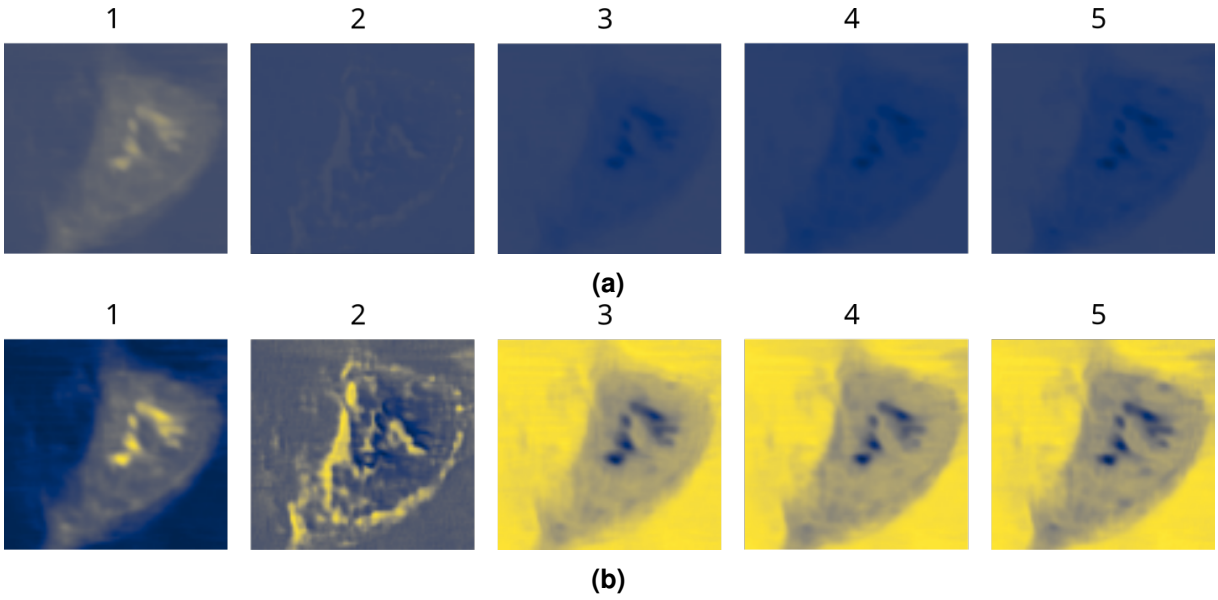


FIGURE 3.44 – Moyennes des concentrations calculés par 10 entraînements sur la cellule de référence. (a) Concentrations avec une échelle de couleur normalisée. (b) Concentrations avec une échelle de couleur non normalisée.

La figure 3.44 montre les concentrations obtenues. L'échelle de couleur normalisée rendant difficile la lecture des concentrations, nous décidons d'utiliser une échelle de couleur non normalisée et de se cantonner à une analyse qualitative des résultats. Le composant 1 est intense au sein de la cellule et atteint son maximum dans les nucléoles de la cellule, ce qui est cohérent avec les spectres du composant. Le composant 2 n'est pas très clair, il semble mettre en avant la membrane plasmique mais contient aussi du signal à l'intérieur du noyau, aux limites des nucléoles. Les composants 3, 4 et 5 sont tous localisés dans les mêmes régions : l'environnement et l'intérieur de la cellule à l'exception des nucléoles.

Bien que le modèle n'atteint pas les performances de la MCR-ALS, il est à noter qu'il réussit à extraire un spectre caractérisant, en partie, les nucléoles sans utiliser de méthodes d'initialisation.

3.3.7 Bilan

Suite aux différentes expérimentations effectuées, il est maintenant possible de définir comment construire un modèle d'AE pour effectuer la MCR sur des données CARS. Tout d'abord, le modèle évalué n'est pas sensible aux débruitages des spectres, il n'est donc pas nécessaire de débruiter les jeux de données avant l'entraînement. Le bruit étant similaire sur tous les spectres, ce dernier n'est pas retenu lors de la réduction de dimension. Une fonction de coût comme la SAD est à privilégier par rapport à une fonction de coût basée sur les distances euclidiennes comme la EQM. Un modèle n'utilisant pas d'initialisation est à favoriser, en effet, l'initialisation permet de diminuer la variation des résultats entre différents entraînements mais elle rend le modèle très dépendant de la méthode d'initialisation utilisée. Pour intégrer la non-négativité des spectres, deux approches ont été évaluées :

— l'application de fonctions d'activation aux paramètres du décodeur,

— l'utilisation d'un décodeur non linéaire finissant par une fonction non-négative.

L'application de la fonction absolue aux paramètres du décodeur permet d'intégrer la contrainte de non négativité sans améliorer ou détériorer les résultats. L'encodeur non linéaire est aussi une approche envisageable mais a plus de difficulté à trouver les différents composants. Lorsque la contrainte spatiale est ajoutée par l'utilisation d'un encodeur spatial convolutif, l'approche par fonction appliquée aux paramètres du décodeur échoue. Le décodeur non linéaire est donc à privilégier. L'apport de la convolution améliore légèrement les résultats par rapport à un encodeur dense dans le cas du décodeur non linéaire.

Une piste pour améliorer les résultats serait de définir des ensembles de pixels de nature similaire pouvant être traités ensemble. Ce regroupement de pixel est couramment appelé super-pixel. Un algorithme pour trouver les super-pixels d'une image est l'algorithme *simple linear iterative clustering* (SLIC) [100]. Ce dernier regroupe des pixels au sein d'un même ensemble selon une distance prenant en compte à la fois les valeurs portées par les pixels et leur proximité spatiale dans l'image. Utiliser les super-pixels permettrait alors de définir des graphes dont les nœuds seraient les pixels formant le super-pixel, et les arêtes relierait les pixels voisins dans l'image. Ces graphes pourraient ensuite servir à entraîner un réseau de neurones sur graphe [103]. Cette approche présente deux avantages. Le premier est l'utilisation d'une convolution la plus juste possible car non limité par la forme des filtres utilisés dans un réseau de neurones convolutif, la seconde est de proposer un modèle qui ne dépend pas de la dimensionnalité spatiale des données. En effet, l'algorithme SLIC s'étend de manière assez directe

à des calculs de super-voxels sur des volumes. Ainsi, la même architecture pourrait traiter de la même manière des cartographies 2D et 3D.

3.4 Conclusion

Dans ce chapitre, les AE ont été introduits ainsi que leur adaptation pour accomplir la MCR. L'état de l'art de leur utilisation dans le contexte de l'HSI a été effectué avec une revue des différentes approches d'intégration des contraintes, d'initialisation, du choix de la fonction de coût ou encore de la non linéarité.

Deux modèles de référence, EndNet et CNNAEU, ont été étudiés en détail pour voir leurs limites. EndNet est très dépendant de l'initialisation pour obtenir des résultats stables. CNNAEU, bien qu'efficace sur les données HSI, échoue sur les données CARS.

Pour mieux comprendre la source des difficultés rencontrées par les AE pour effectuer la MCR sur les données CARS, différentes expériences ont été effectuées. Celles-ci ont permis d'établir comment un auto-encodeur pour la MCR. L'étude de l'ajout de débruitage a montré que le modèle était peu sensible au débruitage. Le bruit étant similaire sur tous les spectres, il est supprimé lors de la réduction de dimension. Le choix de la fonction de coût est crucial, la fonction SAD est bien plus efficace que la fonction EQM et devrait être privilégiée. L'initialisation du décodeur permet d'obtenir des résultats très stables mais rend fortement dépendant le modèle d'une méthode d'initialisation extérieure. Bien qu'étant moins efficace avec un encodeur purement spectral, un décodeur non linéaire donne de meilleurs résultats pour intégrer la non-négativité des spectres. En effet, ce dernier est moins perturbé lors de l'ajout d'un encodeur appliquant une contrainte spatiale que l'approche d'appliquer une fonction non-négative aux paramètres du décodeur. L'ajout d'un encodeur convolutif améliore légèrement les résultats lorsqu'il est combiné avec un décodeur non linéaire. Cependant, des problèmes persistent, empêchant l'AE d'obtenir des résultats robustes et égalant la MCR-ALS. Appliqué à des données réelles, le modèle échoue à faire une MCR efficace mais réussit tout de même à extraire le signal des nucléoles dans la majorité des entraînements. Ces résultats montrent que les AE présentent un fort potentiel pour appliquer la MCR grâce à la possibilité d'obtenir des spectres qui peuvent être analysés sans utiliser de méthodes d'initialisation, principale limite de la MCR-ALS. Des travaux restent à mener pour obtenir des résultats variant moins entre les entraînements et décorrélant mieux les différents composants.

Une autre manière d'intégrer la contrainte spatiale aux données est l'utilisation de super-pixels [100]. Un super-pixel est un regroupement de formes quelconques de

pixels voisins partageant une proximité dans leurs valeurs. Ils présentent l'avantage de regrouper des pixels proches et permettent d'éviter la combinaison entre deux pixels très différents. Ils peuvent être utilisés pour transformer l'image à analyser en un ensemble de graphes pouvant servir à entraîner un auto-encodeur sur graphe. Des travaux dans ce sens sont en cours mais n'ont pas pu être intégrés dans ce manuscrit.

Conclusion générale

Conclusion

La possibilité de visualiser leurs objets d'étude est essentiel pour les biologistes. Les méthodes d'imagerie par fluorescence couramment utilisées, comme IFD et IFI, altèrent les échantillons à analyser et la préparation nécessaire est longue et parfois fastidieuse.

Des alternatives ont alors été développées pour contourner ces limites. Parmi elles, la microspectroscopie CARS permet d'acquérir de l'information sur la composition chimique à partir d'un phénomène vibrationnel non linéaire. Les acquisitions de ce type de microspectroscopie contiennent généralement une riche information spectrale qui n'est pas toujours pleinement exploitée.

Les méthodes de MCR visent à exploiter la richesse spectrale des jeux de données pour caractériser ses principaux composants par leur signature spectrale ainsi que par la quantification de leur concentration en chaque point de mesure. Habituellement résolu par la méthode MCR-ALS, l'aspect spatial des données n'est généralement pas pris en compte dans les méthodes de MCR alors qu'utiliser cette spécificité des données peut apporter une information supplémentaire améliorant les résultats.

Les travaux de cette thèse ont apportés les contributions suivantes :

— l'application de techniques de MCR à des données CARS de cellules et tissus proposant une forme d'imagerie pour la biologie permettant d'associer aux images des spectres renseignant sur la composition chimique des organismes au sein de l'échantillon observé,

— l'intégration d'une contrainte de segmentation par contour actif au sein de la MCR-ALS permettant d'extraire la cellule de son environnement tout en appliquant la méthode de décomposition,

— l'introduction de l'utilisation d'AE pour appliquer la MCR à des données CARS

avec une étude mettant en avant les limites des modèles actuelles pour proposer des résultats plus robustes.

Les travaux du chapitre 2 amènent une discussion sur la méthodologie pour sélectionner le nombre de composants recherchés. L'approche habituelle repose sur les valeurs propres de la SVD des données. Nous proposons de répéter la méthode de MCR avec plusieurs nombres de composants et de choisir comme nombre final le point d'inflexion de la courbe dessinée par les valeurs de métrique de reconstruction. L'application de la MCR-ALS à des cellules et l'analyse des concentrations et spectres trouvés ont permis d'identifier, l'environnement, le noyau ou encore les nucléoles. L'application à un tissu a permis de retrouver le noyau et le cytoplasme des cellules le composant ainsi que la MEC structurant le tissu. En appliquant la MCR-ALS à une cellule surexprimant TrkB « saine », puis en se servant des spectres trouvés comme dictionnaire de projection pour une cellule surexprimant TrkB et exposé au BDNF, il est possible de visualiser les modifications induites par l'ajout de BDNF. Les travaux sur l'ajout de la contrainte de segmentation ont mis en évidence l'impossibilité de construire un réseau de neurones efficace en l'absence de base de données d'entraînement et qu'une approche par contour actif est à privilégier dans ce contexte. L'implémentation de la méthode CSV a permis la segmentation avec succès de cellules lors de l'application de la MCR-ALS.

Dans le chapitre 3, nous avons d'abord cherché à transposer l'état de l'art utilisé en HSI pour effectuer l'opération de démélange, problème équivalent à la MCR, par l'utilisation d'auto-encodeurs. La difficulté à obtenir des résultats aussi convaincants que dans le cas d'images satellitaires hyperspectrales nous a amené à mener une étude sur la construction d'un AE pour la MCR et la robustesse des résultats à l'aide d'un jeu de données artificiel. Les différentes expérimentations ont montrées qu'un bruit au sein des données ne perturbe pas l'entraînement, que la fonction de coût SAD est plus adaptée que l'EQM et qu'initialiser le décodeur du modèle n'améliore pas significativement la qualité du modèle. Pour implémenter la non-négativité des spectres, utiliser une fonction non-négative sur les paramètres du décodeur fonction avec un encodeur dense mais échoue quand l'encodeur est composée de couches convolutives. Les différents résultats obtenus montrent un manque de robustesse dans la méthode dont les résultats varient trop entre différents entraînements. Cependant, elle permet de s'affranchir d'une méthode d'initialisation externe et a réussie à identifier certains éléments comme les nucléoles. Ces résultats de notre approche a mis en évidence un vrai potentiel dans dans l'analyse biologiques, cependant des travaux restent nécessaires pour obtenir une solution robuste.

Perspectives

Comme tout projet de recherche, le sujet n'est pas clos et plusieurs perspectives ont été identifiées. Tout d'abord, l'utilisation de super-pixels [100] pour définir des graphes peut servir à entraîner un réseau de neurones sur graphe. Les super-pixels regroupent ensemble des pixels qui sont à la fois connexes spatialement et similaires spectralement sans contrainte sur la géométrie des groupes formés. Cette approche permettrait d'obtenir une régularisation spatiale la plus juste possible car non limitée par la forme des filtres comme c'est le cas avec les réseaux de neurones convolutifs. De plus, l'utilisation de graphes permettrait de s'abstraire du nombre de dimension spatiale et d'utiliser le même modèle pour une cartographie composée d'une seule tranche que pour une cartographie de tout un volume. Les cartographies de tissus étant couramment effectuées sur son épaisseur en plus de la largeur et longueur, un modèle capable d'appliquer la contrainte spatiale aussi bien sur une tranche que sur un volume serait un apport non-négligeable.

La seconde perspective consiste en un travail sur le lien entre la fonction définie par l'encodeur et celle définie par le décodeur. En effet, le nombre de paramètres important formant un réseau de neurones permet d'obtenir de nombreuses solutions numériquement équivalentes. Or l'encodeur et le décodeur représentent, à l'issue d'un entraînement idéal, une fonction et son inverse. Expliciter la relation entre ces deux blocs pour un nombre de couches quelconque tout en respectant les contraintes imposées par le modèle MCR permettrait de réduire fortement le nombre de paramètres à apprendre et mettrait des contraintes réduisant le nombre de solutions possibles.

Une autre perspective de ces travaux est de reformuler la MCR dans le contexte précis de la microspectroscopie CARS en intégrant le fonctionnement du phénomène pour construire la matrice des spectres. Nous savons qu'un spectre CARS est composé d'un bruit de fond non-résonnant et d'une partie résonnante complexe. Le calcul des spectres pourrait être simplifié en déterminant en amont le bruit de fond non-résonnant et en faisant chercher au modèle seulement la partie résonnante des spectres des composants. Cette approche aurait l'avantage de permettre de supprimer le bruit de fond non-résonnant des spectres calculés et de faciliter leur analyse.

Enfin, afin de valider le plus justement possible les méthodes et permettre des études plus poussées, il semble important de travailler à la création de jeux de données CARS d'échantillons biologiques dont les spectres des composants et les concentrations connus. Pour créer ces jeux de données, des organismes biologiques simples pourraient servir aux mesures et les mesures CARS pourraient être accompagnées de

mesures reposant sur d'autres phénomènes physiques pour aider à la caractérisation des différents pixels acquis.

A

Annexes

Sommaire

A.1 Relation de Kramers-Kronig

Si un système est linéaire, il peut être caractérisé dans le domaine temporel par sa réponse impulsionnelle $h(t)$ dans le domaine temporel :

$$o(t) = g(t) * h(t) = \int_{-\infty}^{\infty} g(\tau)h(t-\tau)d\tau, \tag{A.1}$$

avec g et o respectivement l'entrée et la sortie du système. En utilisant la transformée de Fourier, il est possible de caractériser ce système dans le domaine fréquentielle par sa fonction de transfert $H(\omega)$:

$$O(\omega) = H(\omega)G(\omega). \tag{A.2}$$

Si h est un nombre réel pur, les parties imaginaires et réelles de H sont symétriques. De plus, si ce système est causal, c'est-à-dire un système dont la réponse ne dépend que des entrées passées, la réponse impulsionnelle h et la sortie et o sont nulles pour $t \leq 0$. Il est alors possible de redéfinir H de la manière suivante :

$$H(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t)u(t)e^{-i\omega t}dt = \mathcal{F}[h(t)u(t)], \tag{A.3}$$

avec u la fonction de Heaviside prenant pour valeur 0 si $t < 0$, 1 sinon et $\mathcal{F}$ la transformée de Fourier. En appliquant la relation entre transformée de Fourier et convolution :

$$H(\omega) = \frac{H(\omega)}{\sqrt{2\pi}} * \mathcal{F}[u(t)], \tag{A.4}$$

En développant la convolution, en éliminant les fréquences négatives et en considérant la réponse impulsionnelle comme réelle pure, une relation entre la partie réelle et imaginaire finit par apparaître :

$$\begin{aligned} \mathrm{Re}(H(\omega)) &= \frac{PV}{\pi} \int_{-\infty}^{\infty} \frac{\mathrm{Im}(H(\omega'))}{\omega' - \omega} d\omega', \\ \mathrm{Im}(H(\omega)) &= -\frac{PV}{\pi} \int_{-\infty}^{\infty} \frac{\mathrm{Re}(H(\omega'))}{\omega' - \omega} d\omega'. \end{aligned} \tag{A.5}$$

Dans ces équations, PV est la valeur principale de Cauchy.

A.2 Développement de $\psi(f(\omega))$

$$\psi(f(\omega)) = \frac{1}{\sqrt{2\pi}}(F[u(t)] * f(\omega)). \tag{A.6}$$

La transformée de Fourier de la fonction de Heaviside s'écrit :

$$F[u(t)] = \frac{1}{i\sqrt{2\pi}\omega} + \sqrt{\frac{\pi}{2}}\delta(\omega), \tag{A.7}$$

avec δ la distribution de Dirac. En remplaçant $F[u(t)]$ par son expression et explicitant la convolution, $\psi(f(\omega))$ devient :

$$\begin{aligned}
\psi(f(\omega)) &= \frac{1}{\sqrt{2\pi}}\left[\left(\frac{1}{i\sqrt{2\pi}\omega} + \sqrt{\frac{\pi}{2}}\delta(\omega)\right) * f(\omega)\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[PV\int_{-\infty}^{+\infty}\left(\frac{1}{i\sqrt{2\pi}x} + \sqrt{\frac{\pi}{2}}\delta(x)\right) f(\omega - x)dx\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[PV\int_{-\infty}^{+\infty}\frac{f(\omega - x)}{i\sqrt{2\pi}x}dx + \int_{-\infty}^{+\infty}\sqrt{\frac{\pi}{2}}\delta(x)f(\omega - x)dx\right] \qquad \text{(A.8)} \\
&= \frac{1}{\sqrt{2\pi}}\left[PV\int_{-\infty}^{+\infty}\frac{f(\omega - x)}{i\sqrt{2\pi}x}dx + \sqrt{\frac{\pi}{2}}f(\omega)\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{1}{i\sqrt{2\pi}}PV\int_{-\infty}^{+\infty}\frac{f(\omega - x)}{x}dx + \sqrt{\frac{\pi}{2}}f(\omega)\right] \\
\psi(f(\omega)) &= \frac{1}{\sqrt{2\pi}}\left[\frac{1}{i\sqrt{2\pi}}PV\int_{-\infty}^{+\infty}\frac{f(\omega - x)}{x}dx + \sqrt{\frac{\pi}{2}}f(\omega)\right]. \qquad \text{(A.9)}
\end{aligned}$$

En utilisant la méthode d'Intégration par changement de variable, il est possible de définir :

$$\omega' = \omega - x \tag{A.10}$$

$$x = \omega - \omega' \tag{A.11}$$

$$\begin{aligned}
\frac{d\omega'}{dx} &= \frac{d(\omega - x)}{dx} = -1 \\
\Rightarrow d\omega' &= -dx.
\end{aligned} \tag{A.12}$$

Ainsi, l'équation A.9 devient finalement :

$$\begin{aligned}
\psi(f(\omega)) &= \frac{1}{\sqrt{2\pi}}\left[\frac{1}{i\sqrt{2\pi}}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega-\omega'}-d\omega'+\sqrt{\frac{\pi}{2}}f(\omega)\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{1}{i\sqrt{2\pi}}PV\int_{-\infty}^{+\infty}-\frac{f(\omega')}{\omega-\omega'}d\omega'+\sqrt{\frac{\pi}{2}}f(\omega)\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{1}{i\sqrt{2\pi}}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega'+\sqrt{\frac{\pi}{2}}f(\omega)\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{-i}{\sqrt{2\pi}}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega'+\sqrt{\frac{\pi}{2}}f(\omega)\right] \qquad \text{(A.13)} \\
&= -\frac{i}{2\pi}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega'+\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}}f(\omega) \\
&= -\frac{i}{2\pi}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega'+\frac{1}{\sqrt{2\pi}}\frac{\sqrt{2\pi}}{2}f(\omega) \\
&= -\frac{i}{2\pi}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega'+\frac{1}{2}f(\omega) \\
\psi(f(\omega)) &= \frac{1}{2}\left(\frac{-i}{\pi}PV\int_{-\infty}^{+\infty}\frac{f(\omega')}{\omega'-\omega}d\omega'+f(\omega)\right). \qquad \text{(A.14)}
\end{aligned}$$

A.3 Application de la MCR-ALS à une cellule HEK-293 vivante en interphase

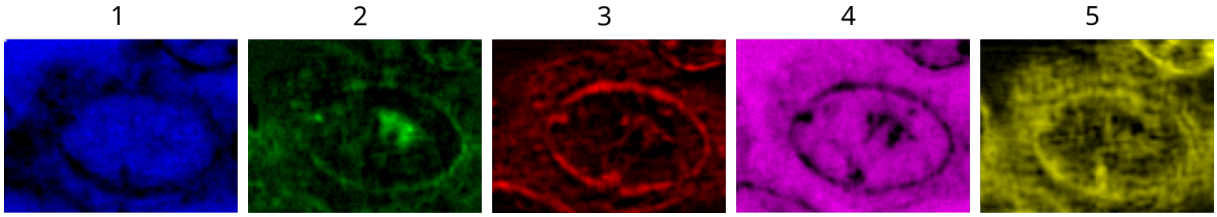


FIGURE A.1 – Les 5 concentrations calculées par la MCR-ALS sur une cellule HEK-293 vivante en interphase.

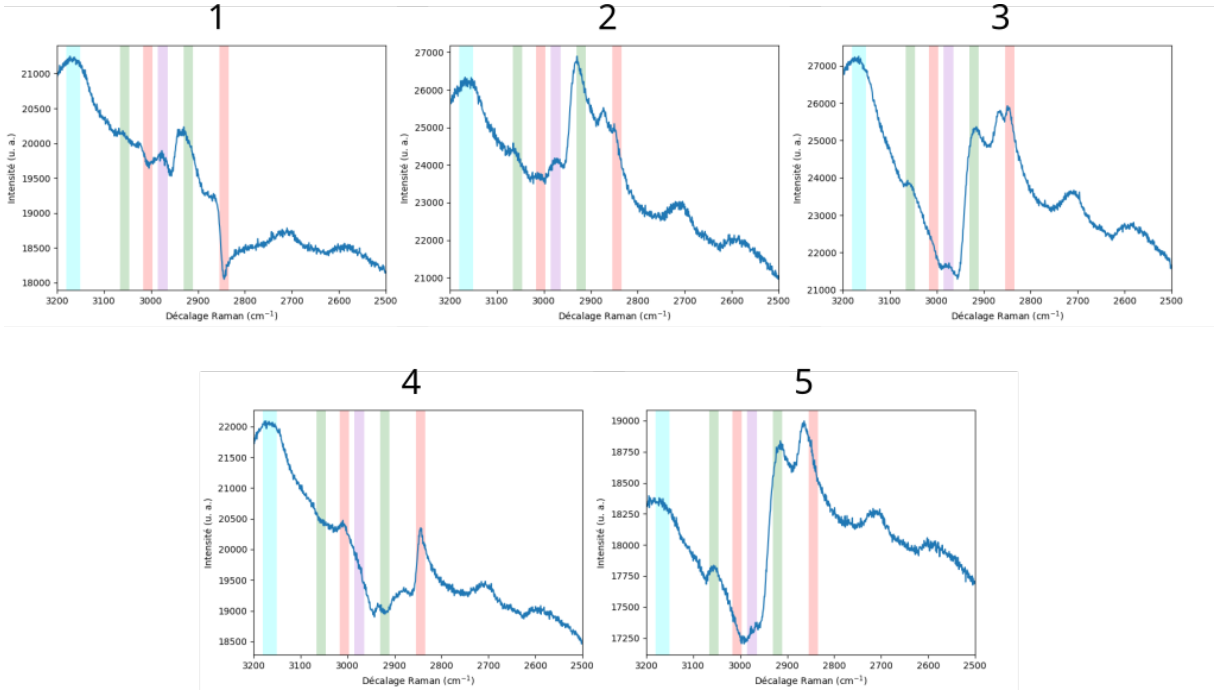


FIGURE A.2 – Les 5 spectres calculées par la MCR-ALS sur une cellule HEK-293 vivante en interphase.

A.4 Application de la MCR-ALS à un tissu adipeux blanc de souris

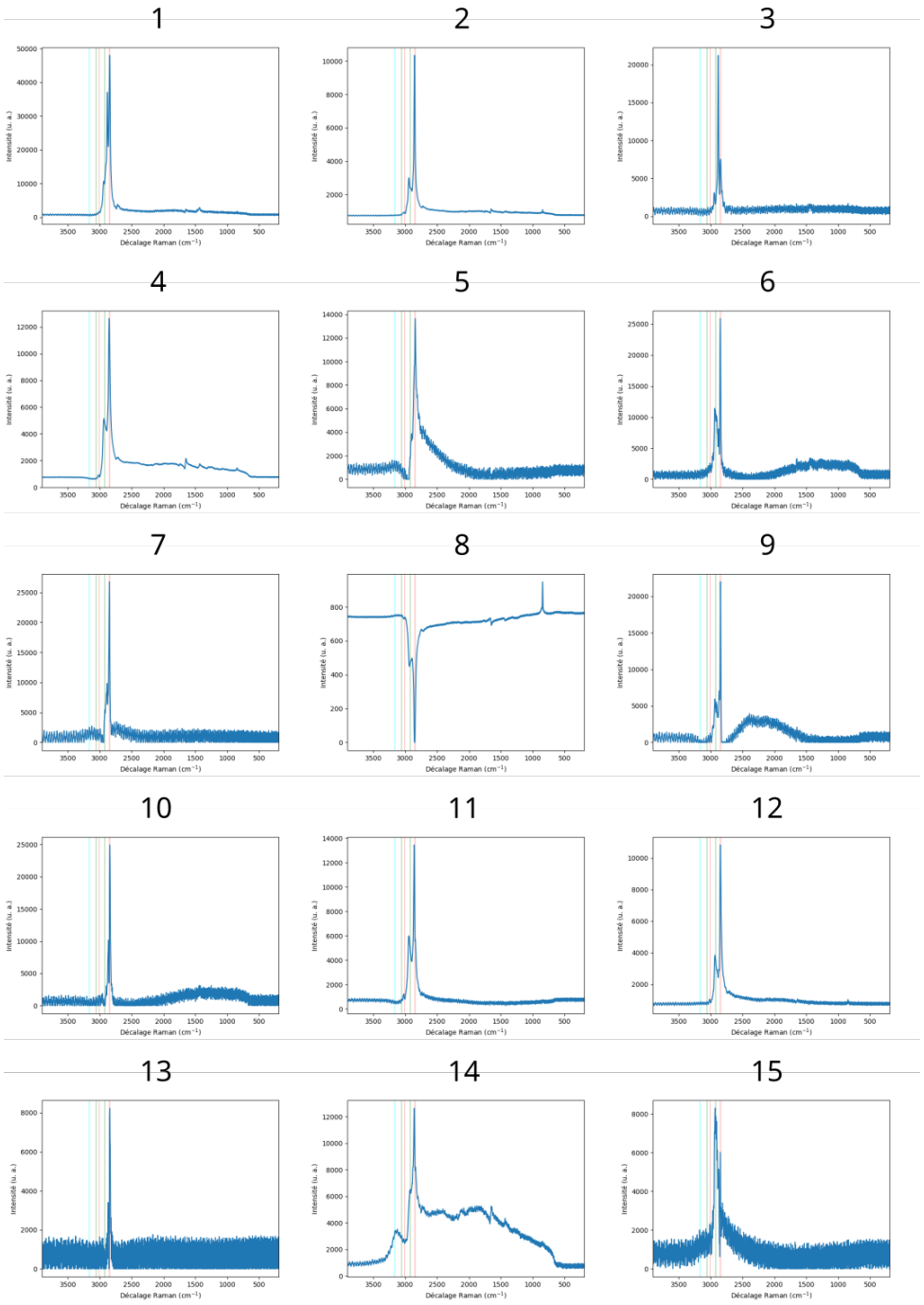


FIGURE A.3 – Les spectres calculés par la MCR-ALS à partir d'un tissu adipeux blanc de souris.

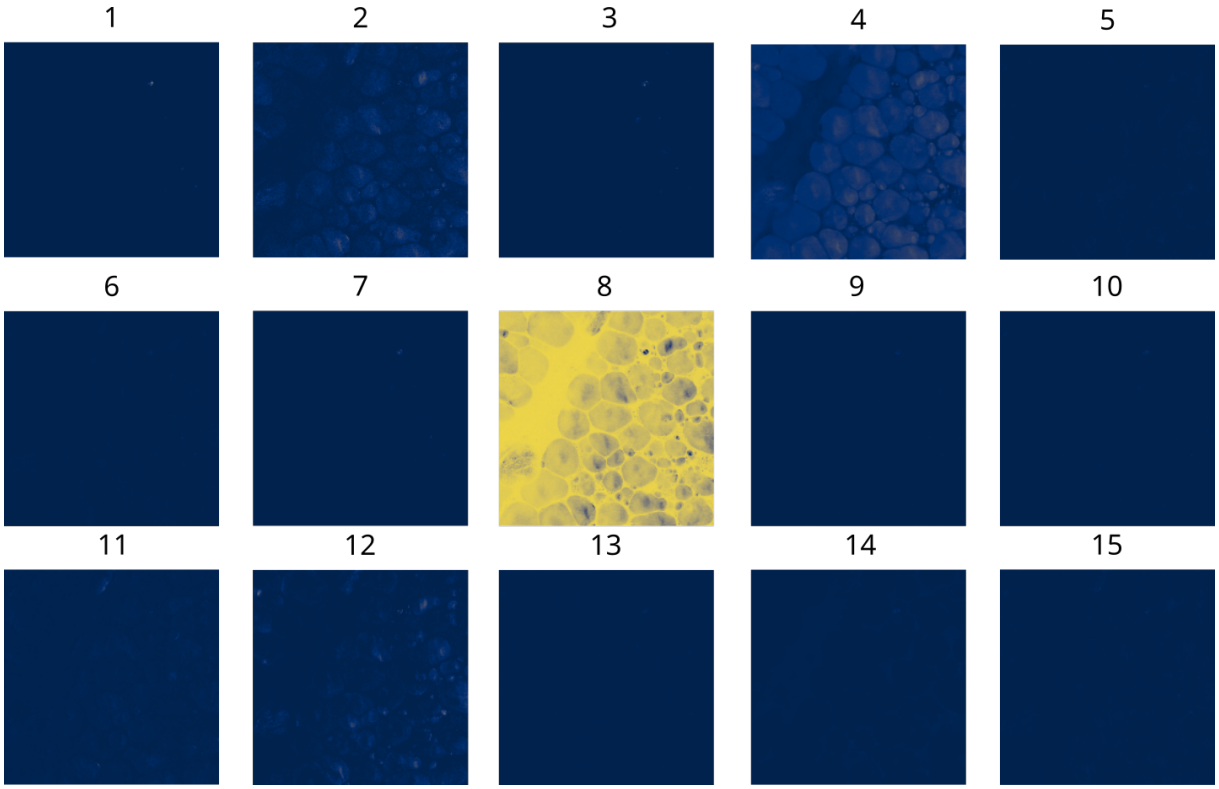


FIGURE A.4 – Les concentrations calculés par la MCR-ALS à partir d'un tissu adipeux blanc de souris.

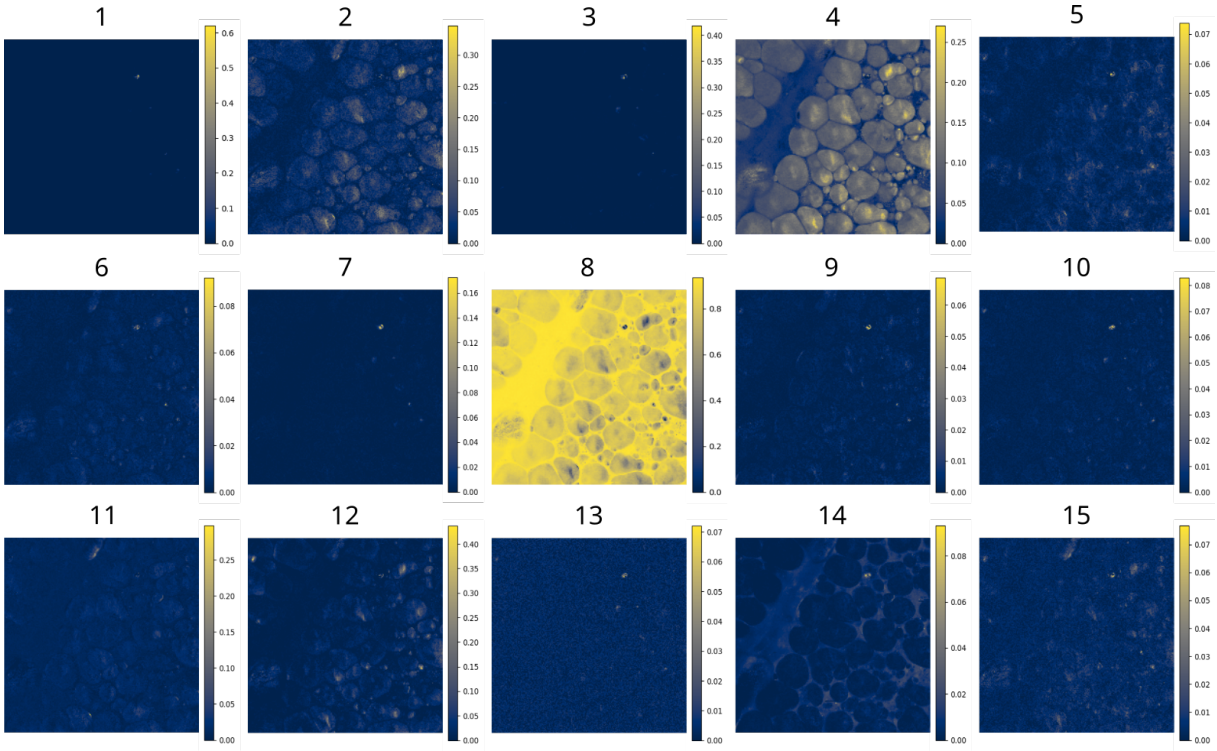


FIGURE A.5 – Les concentrations calculés par la MCR-ALS à partir d'un tissu adipeux blanc de souris avec une échelle de couleur non normalisée.

A.5 Paramétrisation de la contrainte de Chan-Sandberg-Vese

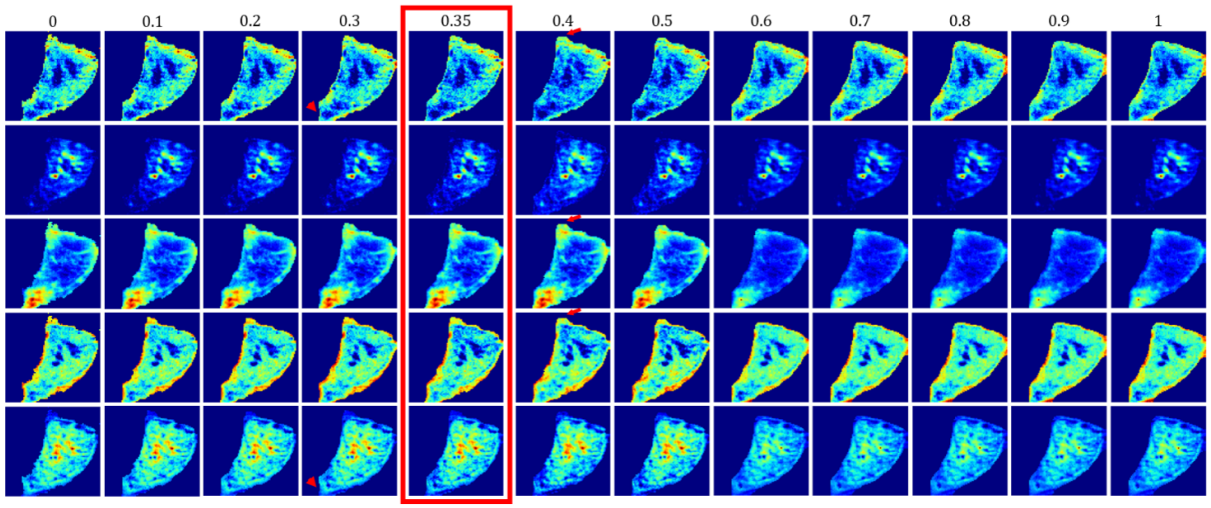


FIGURE A.6 – Résultats de la MCR-ALS avec contrainte CSV pour différentes valeurs v.

B

Bibliographie

Sommaire

Références

[1] Y. Liu, Y. J. Lee et M. T. Cicerone, « Broadband CARS spectral phase retrieval using a time-domain Kramers–Kronig transform, » *Opt. Lett., OL*, t. 34, nº 9, p. 1363-1365, 1er mai 2009, Publisher : Optica Publishing Group.

[2] E. Capitaine, « Nouveaux procédés de microspectroscopie Raman cohérente à bande ultralarge, » thèse de doct., Université de Limoges, 2017.

[3] T. Guerenne-Del Ben, « La microspectroscopie CARS appliquée au suivi de l'activation d'un récepteur associé à la cancérogenèse, » thèse de doct., Université de Limoges, 2019.

[4] T. Guerenne-Del Ben, V. Couderc, L. Duponchel, V. Sol, P. Leproux, J.-M. Petit *et al.*, « Multiplex coherent anti-Stokes Raman scattering microspectroscopy detection of lipid droplets in cancer cells expressing TrkB, » *Scientific reports*, t. 10, nº 1, p. 1-12, 2020.

[5] A. Tixier-Vidal, « Les compartiments membranaires de la cellule eucaryote, » *médecine/sciences*, t. 18, nº 10, p. 1004-1011, 2002.

[6] J. G. Carlton, H. Jones et U. S. Eggert, « Membrane and organelle dynamics during cell division, » *Nature Reviews Molecular Cell Biology*, t. 21, nº 3, p. 151-166, 2020.

[7] T. Guerenne-Del Ben, Z. Rajaofara, V. Couderc *et al.*, « Multiplex coherent anti-Stokes Raman scattering highlights state of chromatin condensation in CH region, » *Scientific reports*, t. 9, nº 1, p. 1-10, 2019.

[8] J. Kapuscinski, « DAPI : a DNA-specific fluorescent probe, » *Biotechnic & histochemistry*, t. 70, nº 5, p. 220-233, 1995.

[9] C. V. Raman et K. S. Krishnan, « A new type of secondary radiation, » *Nature*, t. 121, nº 3048, p. 501-502, 1928.

[10] I. R. M. Ramos, A. Malkin et F. M. Lyng, « Current advances in the application of Raman spectroscopy for molecular diagnosis of cervical cancer, » *BioMed research international*, t. 2015, 2015.

[11] J. SMULKO et M. WRÓBEL, « Noise sources in Raman spectroscopy of biological objects, » in *Dynamics and Fluctuations in Biomedical Photonics XIV*, International Society for Optics et Photonics, t. 10063, 2017, 100630Q.

[12] P. MAKER et R. TERHUNE, « Study of optical effects due to an induced polarization third order in the electric field strength, » *Physical Review*, t. 137, nº 3A, A801, 1965.

[13] R. BEGLEY, A. HARVEY et R. L. BYER, « Coherent anti-Stokes Raman spectroscopy, » *Applied Physics Letters*, t. 25, nº 7, p. 387-390, 1974.

[14] M. MÜLLER et J. M. SCHINS, « Imaging the thermodynamic state of lipid membranes with multiplex CARS microscopy, » *The Journal of Physical Chemistry B*, t. 106, nº 14, p. 3715-3723, 2002.

[15] R. R. ALFANO et S. SHAPIRO, « Observation of self-phase modulation and small-scale filaments in crystals and glasses, » *Physical Review Letters*, t. 24, nº 11, p. 592, 1970.

[16] E. M. VARTIAINEN, « Phase retrieval approach for coherent anti-stokes raman scattering spectrum analysis, » *J. Opt. Soc. Am. B*, t. 9, nº 8, p. 1209, 1er août 1992.

[17] S. HAYKIN et S. KESLER, « Prediction-error filtering and maximum-entropy spectral estimation, » in *Nonlinear Methods of Spectral Analysis*, S. HAYKIN, éd., t. 34, Series Title : Topics in Applied Physics, Berlin, Heidelberg : Springer Berlin Heidelberg, 1979, p. 9-72.

[18] H. A. RINIA, M. BONN, M. MÜLLER et E. M. VARTIAINEN, « Quantitative CARS spectroscopy using the maximum entropy method : the main lipid phase transition, » *ChemPhysChem*, t. 8, nº 2, p. 279-287, 2 fév. 2007.

[19] R. ARORA, G. I. PETROV et V. V. YAKOVLEV, « Analytical capabilities of coherent anti-Stokes Raman scattering microspectroscopy, » *Journal of Modern Optics*, t. 55, nº 19, p. 3237-3254, 10 nov. 2008, Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/09500340802168639.

[20] C. H. CAMP, Y. J. LEE et M. T. CICERONE, « Quantitative, comparable coherent anti-stokes raman scattering (CARS) spectroscopy : correcting errors in phase retrieval : quantitative, comparable CARS spectroscopy, » *J. Raman Spectrosc.*, t. 47, nº 4, p. 408-415, avr. 2016.

[21] T. HÄRKÖNEN, L. ROININEN, M. T. MOORES et E. M. VARTIAINEN, « Bayesian quantification for coherent anti-stokes raman scattering spectroscopy, » *J. Phys. Chem. B*, t. 124, nº 32, p. 7005-7012, 13 août 2020.

[22] R. Houhou, P. Barman, M. Schmitt, T. Meyer, J. Popp et T. Bocklitz, « Deep learning as phase retrieval tool for CARS spectra, » *Opt. Express*, t. 28, nº 14, p. 21 002, 6 juill. 2020.

[23] Z. Wang, K. O' Dwyer, R. Muddiman, T. Ward, C. H. Camp Jr. et B. M. Hennelly, « VECTOR : very deep convolutional autoencoders for non-resonant background removal in broadband coherent anti-stokes raman scattering, » *Journal of Raman Spectroscopy*, t. n/a, n/a.

[24] C. M. Valensise, A. Giuseppi, F. Vernuccio, A. De la Cadena, G. Cerullo et D. Polli, « Removing non-resonant background from CARS spectra via deep learning, » *APL Photonics*, t. 5, nº 6, p. 061 305, 2020.

[25] R. Junjuri, A. Saghi, L. Lensu et E. M. Vartiainen, « Convolutional neural network-based retrieval of Raman signals from CARS spectra, » *Optics Continuum*, t. 1, nº 6, p. 1324-1339, 2022.

[26] C. Krafft, T. Knetschke, R. H. Funk et R. Salzer, « Studies on stress-induced changes at the subcellular level by Raman microspectroscopic mapping, » *Analytical chemistry*, t. 78, nº 13, p. 4424-4429, 2006.

[27] T. Ichimura, L.-d. Chiu, K. Fujita *et al.*, « Visualizing cell state transition using Raman spectroscopy, » *PLoS One*, t. 9, nº 1, e84478, 2014.

[28] H. Kano, T. Maruyama, J. Kano *et al.*, « Ultra-multiplex CARS spectroscopic imaging with 1-millisecond pixel dwell time, » *OSA Continuum*, t. 2, nº 5, p. 1693, 15 mai 2019.

[29] E. Capitaine, N. O. Moussa, C. Louot *et al.*, « Fast epi-detected broadband multiplex CARS and SHG imaging of mouse skull cells, » *Biomed. Opt. Express*, t. 9, nº 1, p. 245, 1er jan. 2018.

[30] D. Boildieu, T. Guerenne-Del Ben, L. Duponchel *et al.*, « Coherent anti-Stokes Raman scattering cell imaging and segmentation with unsupervised data analysis, » *Front. Cell Dev. Biol.*, t. 10, août 2022.

[31] X. Nan, J.-X. Cheng et X. S. Xie, « Vibrational imaging of lipid droplets in live fibroblast cells with coherent anti-Stokes Raman scattering microscopy, » *Journal of lipid research*, t. 44, nº 11, p. 2202-2208, 2003.

[32] C. Matthäus, T. Chernenko, J. A. Newmark, C. M. Warner et M. Diem, « Label-free detection of mitochondrial distribution in cells by nonresonant Raman microspectroscopy, » *Biophysical journal*, t. 93, nº 2, p. 668-673, 2007.

[33] P. Matousek, E. R. Draper, A. E. Goodship, I. P. Clark, K. L. Ronayne et A. W. Parker, « Noninvasive Raman spectroscopy of human tissue in vivo, » *Applied spectroscopy*, t. 60, nº 7, p. 758-763, 2006.

[34] Y. TAKEI, R. HIRAI, A. FUKUDA *et al.*, « Visualization of intracellular lipid metabolism in brown adipocytes by time-lapse ultra-multiplex CARS microspectroscopy with an onstage incubator, » *The Journal of Chemical Physics*, t. 155, nº 12, p. 125 102, 2021.

[35] Y. J. LEE, S. L. VEGA, P. J. PATEL, K. A. AAMER, P. V. MOGHE et M. T. CICERONE, « Quantitative, label-free characterization of stem cell differentiation at the single-cell level by broadband coherent anti-Stokes Raman scattering microscopy, » *Tissue Engineering Part C : Methods*, t. 20, nº 7, p. 562-569, 2014.

[36] T. ICHIMURA, L.-d. CHIU, K. FUJITA *et al.*, « Visualizing the appearance and disappearance of the attractor of differentiation using Raman spectral imaging, » *Scientific reports*, t. 5, nº 1, p. 1-10, 2015.

[37] R. FURUTA, N. KURAKE, K. TAKEDA *et al.*, « Lipid droplets exhaustion with caspases activation in HeLa cells cultured in plasma-activated medium observed by multiplex coherent anti-Stokes Raman scattering microscopy, » *Biointerphases*, t. 12, nº 3, p. 031 006, 2017.

[38] C. V. DESSAI, A. PLISS, A. N. KUZMIN, E. P. FURLANI et P. N. PRASAD, « Coherent Raman spectroscopic imaging to characterize microglia activation pathway, » *Journal of biophotonics*, t. 12, nº 5, e201800133, 2019.

[39] C. SCALFI-HAPP, M. UDART, C. HAUSER et A. RÜCK, « Investigation of lipid bodies in a colon carcinoma cell line by confocal Raman microscopy, » *Medical Laser Application*, t. 26, nº 4, p. 152-157, 2011.

[40] H. ABRAMCZYK, J. SURMACKI, M. KOPEĆ, A. K. OLEJNIK, K. LUBECKA-PIETRUSZEWSKA et K. FABIANOWSKA-MAJEWSKA, « The role of lipid droplets and adipocytes in cancer. Raman imaging of cell cultures : MCF10A, MCF7, and MDA-MB-231 compared to adipocytes in cancerous human breast tissue, » *Analyst*, t. 140, nº 7, p. 2224-2235, 2015.

[41] D. CHATURVEDI, S. A. BALAJI, V. K. BN, F. ARIESE, S. UMAPATHY et A. RANGARAJAN, « Different phases of breast cancer cells : Raman study of immortalized, transformed, and invasive cells, » *Biosensors*, t. 6, nº 4, p. 57, 2016.

[42] F. DRAUX, C. GOBINET, J. SULÉ-SUSO, M. MANFAIT, P. JEANNESSON et G. D. SOCKALINGUM, « Raman imaging of single living cells : probing effects of non-cytotoxic doses of an anti-cancer drug, » *Analyst*, t. 136, nº 13, p. 2718-2725, 2011.

[43] S. F. EL-MASHTOLY, D. PETERSEN, H. K. YOSEF *et al.*, « Label-free imaging of drug distribution and metabolism in colon cancer cells by Raman microscopy, » *Analyst*, t. 139, nº 5, p. 1155-1161, 2014.

[44] H. K. YOSEF, L. MAVARANI, A. MAGHNOUJ, S. HAHN, S. F. EL-MASHTOLY et K. GERWERT, « In vitro prediction of the efficacy of molecularly targeted cancer therapy by Raman spectral imaging, » *Analytical and bioanalytical chemistry*, t. 407, n° 27, p. 8321-8331, 2015.

[45] A. MIGNOLET, B. WOOD et E. GOORMAGHTIGH, « Intracellular investigation on the differential effects of 4 polyphenols on MCF-7 breast cancer cells by Raman imaging, » *Analyst*, t. 143, n° 1, p. 258-269, 2018.

[46] H. AKIL, A. PERRAUD, M.-O. JAUBERTEAU et M. MATHONNET, « Tropomyosin-related kinase B/brain derived-neurotrophic factor signaling pathway as a potential therapeutic target for colorectal cancer, » *World journal of gastroenterology*, t. 22, n° 2, p. 490, 2016.

[47] K. PEARSON, « On lines of closes fit to system of points in space, London, E dinb, » *Dublin Philos. Mag. J. Sci*, t. 2, p. 559-572, 1901.

[48] H. HOTELLING, « Analysis of a complex of statistical variables into principal components., » *Journal of educational psychology*, t. 24, n° 6, p. 417, 1933.

[49] R. L. THORNDIKE, « Who belongs in the family, » in *Psychometrika*, Citeseer, 1953.

[50] J. B. TENENBAUM, V. d. SILVA et J. C. LANGFORD, « A global geometric framework for nonlinear dimensionality reduction, » *science*, t. 290, n° 5500, p. 2319-2323, 2000.

[51] N. S. ALTMAN, « An introduction to kernel and nearest-neighbor nonparametric regression, » *The American Statistician*, t. 46, n° 3, p. 175-185, 1992.

[52] E. W. DIJKSTRA *et al.*, « A note on two problems in connexion with graphs, » *Numerische mathematik*, t. 1, n° 1, p. 269-271, 1959.

[53] F. A. KRUSE, A. LEFKOFF, J. BOARDMAN *et al.*, « The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data, » *Remote sensing of environment*, t. 44, n° 2-3, p. 145-163, 1993.

[54] R. TAULER, B. KOWALSKI et S. FLEMING, « Multivariate curve resolution applied to spectral data from multiple runs of an industrial process, » *Analytical chemistry*, t. 65, n° 15, p. 2040-2047, 1993.

[55] J. C. HAMILTON et P. J. GEMPERLINE, « Mixture analysis using factor analysis. II : self-modeling curve resolution, » *Journal of chemometrics*, t. 4, n° 1, p. 1-13, 1990.

[56] N. KESHAVA et J. F. MUSTARD, « Spectral unmixing, » *IEEE signal processing magazine*, t. 19, n° 1, p. 44-57, 2002.

[57] C. L. LAWSON et R. J. HANSON, *Solving least squares problems*. Society for Industrial et Applied Mathematics, 1995.

[58] W. WINDIG et J. GUILMENT, « Interactive self-modeling mixture analysis, » *Anal. Chem.*, t. 63, nº 14, p. 1425-1432, 15 juill. 1991.

[59] J. NASCIMENTO et J. DIAS, « Vertex component analysis : a fast algorithm to unmix hyperspectral data, » *IEEE Trans. Geosci. Remote Sensing*, t. 43, nº 4, p. 898-910, avr. 2005.

[60] T. PEDERSON, « The nucleolus, » *Cold Spring Harbor perspectives in biology*, t. 3, nº 3, a000638, 2011.

[61] P. LEPROUX, D. BOILDIEU, Z. RAJAOFARA *et al.*, « Recent advances in cell and tissue imaging by multiplex CARS microspectroscopy, » in *International conference on Laser Applications in Life Science*, 2022.

[62] E. A. HOFFMAN, B. L. FREY, L. M. SMITH et D. T. AUBLE, « Formaldehyde crosslinking : a tool for the study of chromatin complexes, » *Journal of Biological Chemistry*, t. 290, nº 44, p. 26 404-26 411, 2015.

[63] S.-O. KIM, J. KIM, T. OKAJIMA et N.-J. CHO, « Mechanical properties of paraformaldehyde-treated individual cells investigated by atomic force microscopy and scanning ion conductance microscopy, » *Nano convergence*, t. 4, nº 1, p. 1-8, 2017.

[64] N. OTSU, « A threshold selection method from gray-level histograms, » *IEEE transactions on systems, man, and cybernetics*, t. 9, nº 1, p. 62-66, 1979.

[65] J. MACQUEEN, « Classification and analysis of multivariate observations, » in *5th Berkeley Symp. Math. Statist. Probability*, 1967, p. 281-297.

[66] T. CHAN et L. VESE, « An active contour model without edges, » in *Scale-Space Theories in Computer Vision*, M. NIELSEN, P. JOHANSEN, O. F. OLSEN et J. WEICKERT, éd., Berlin, Heidelberg : Springer, 1999, p. 141-151.

[67] D. P. KINGMA et J. BA, « Adam : A method for stochastic optimization, » *arXiv preprint arXiv :1412.6980*, 2014.

[68] S. SANTURKAR, D. TSIPRAS, A. ILYAS et A. MADRY, « How does batch normalization help optimization ? » *Advances in neural information processing systems*, t. 31, 2018.

[69] L. R. DICE, « Measures of the amount of ecologic association between species, » *Ecology*, t. 26, nº 3, p. 297-302, 1945.

[70] T. A. SORENSEN, « A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, » *Biol. Skar.*, t. 5, p. 1-34, 1948.

[71] A. DUFOUR, V. SHININ, S. TAJBAKHSH, N. GUILLEN-AGHION, J.-C. OLIVO-MARIN et C. ZIMMER, « Segmenting and tracking fluorescent cells in dynamic 3-d microscopy with coupled active surfaces, » *IEEE Trans. on Image Process.*, t. 14, nº 9, p. 1396-1410, sept. 2005.

[72] M. MASKA, O. DANEK, S. GARASA, A. ROUZAUT, A. MUNOZ-BARRUTIA et C. ORTIZ-DE-SOLORZANO, « Segmentation and shape tracking of whole fluorescent cells based on the chan–vese model, » *IEEE Trans. Med. Imaging*, t. 32, nº 6, p. 995-1006, juin 2013.

[73] G. LU, L. HALIG, D. WANG, X. QIN, Z. G. CHEN et B. FEI, « Spectral-spatial classification for noninvasive cancer detection using hyperspectral imaging, » *J. Biomed. Opt*, t. 19, nº 10, p. 106 004, 2 oct. 2014.

[74] P. GETREUER, « Chan-vese segmentation, » *Image Processing On Line*, t. 2, p. 214-224, 8 août 2012.

[75] S. OSHER et J. A. SETHIAN, « Fronts propagating with curvature-dependent speed : Algorithms based on Hamilton-Jacobi formulations, » *Journal of computational physics*, t. 79, nº 1, p. 12-49, 1988.

[76] T. F. CHAN, B. Y. SANDBERG et L. A. VESE, « Active contours without edges for vector-valued images, » *Journal of Visual Communication and Image Representation*, t. 11, nº 2, p. 130-141, 1er juin 2000.

[77] R. VITALE, S. HUGELIER, D. CEVOLI et C. RUCKEBUSCH, « A spatial constraint to model and extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral images, » *Analytica Chimica Acta*, t. 1095, p. 30-37, 2020.

[78] S. HUGELIER, O. DEVOS et C. RUCKEBUSCH, « On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis, » *Journal of Chemometrics*, t. 29, nº 10, p. 557-561, 2015.

[79] P. BALDI et K. HORNIK, « Neural networks and principal component analysis : Learning from examples without local minima, » *Neural networks*, t. 2, nº 1, p. 53-58, 1989.

[80] B. A. OLSHAUSEN et D. J. FIELD, « Emergence of simple-cell receptive field properties by learning a sparse code for natural images, » *Nature*, t. 381, nº 6583, p. 607-609, 1996.

[81] P. VINCENT, H. LAROCHELLE, I. LAJOIE, Y. BENGIO, P.-A. MANZAGOL et L. BOTTOU, « Stacked denoising autoencoders : Learning useful representations in

a deep network with a local denoising criterion., » *Journal of machine learning research*, t. 11, nº 12, 2010.

[82] R. SALAH, P. VINCENT, X. MULLER *et al.*, « Contractive auto-encoders : Explicit invariance during feature extraction, » in *Proc. of the 28th International Conference on Machine Learning*, 2011, p. 833-840.

[83] D. P. KINGMA et M. WELLING, « Auto-encoding variational bayes, » *arXiv preprint arXiv :1312.6114*, 2013.

[84] R. GUO, W. WANG et H. QI, « Hyperspectral image unmixing using autoencoder cascade, » in *2015 7th Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Tokyo, Japan : IEEE, juin 2015, p. 1-4.

[85] B. PALSSON, M. O. ULFARSSON et J. R. SVEINSSON, « Convolutional autoencoder for spectral–spatial hyperspectral unmixing, » *IEEE Trans. Geosci. Remote Sensing*, t. 59, nº 1, p. 535-549, jan. 2021.

[86] M. WANG, M. ZHAO, J. CHEN et S. RAHARDJA, « Nonlinear unmixing of hyperspectral data via deep autoencoder networks, » *IEEE Geosci. Remote Sensing Lett.*, t. 16, nº 9, p. 1467-1471, sept. 2019.

[87] S. SHI, M. ZHAO, L. ZHANG et J. CHEN, « Variational autoencoders for hyperspectral unmixing with endmember variability, » in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada : IEEE, 6 juin 2021, p. 1875-1879.

[88] R. A. BORSOI, T. IMBIRIBA et J. C. M. BERMUDEZ, « Deep generative endmember modeling : an application to unsupervised spectral unmixing, » *IEEE Trans. Comput. Imaging*, t. 6, p. 374-384, 2020.

[89] Y. SU, A. MARINONI, J. LI, J. PLAZA et P. GAMBA, « Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing, » *IEEE Geosci. Remote Sensing Lett.*, t. 15, nº 9, p. 1427-1431, sept. 2018.

[90] Y. QU et H. QI, « uDAS : an untied denoising autoencoder with sparsity for spectral unmixing, » *IEEE Trans. Geosci. Remote Sensing*, t. 57, nº 3, p. 1698-1712, mars 2019.

[91] Y. SU, J. LI, A. PLAZA, A. MARINONI, P. GAMBA et S. CHAKRAVORTTY, « DAEN : deep autoencoder networks for hyperspectral unmixing, » *IEEE Trans. Geosci. Remote Sensing*, t. 57, nº 7, p. 4309-4321, juill. 2019.

[92] S. OZKAN, B. KAYA et G. B. AKAR, « EndNet : sparse AutoEncoder network for endmember extraction and hyperspectral unmixing, » *IEEE Trans. Geosci. Remote Sensing*, t. 57, nº 1, p. 482-496, jan. 2019.

[93] Z. HAN, D. HONG, L. GAO, B. ZHANG et J. CHANUSSOT, « Deep half-siamese networks for hyperspectral unmixing, » *IEEE Geosci. Remote Sensing Lett.*, t. 18, nº 11, p. 1996-2000, nov. 2021.

[94] B. PALSSON, J. SIGURDSSON, J. R. SVEINSSON et M. O. ULFARSSON, « Hyperspectral unmixing using a neural network autoencoder, » *IEEE Access*, t. 6, p. 25 646-25 656, 2018.

[95] Z. HAN, D. HONG, L. GAO, J. YAO, B. ZHANG et J. CHANUSSOT, « Multimodal hyperspectral unmixing : insights from attention networks, » *IEEE Trans. Geosci. Remote Sensing*, p. 1-1, 2022.

[96] L. QI, F. GAO, J. DONG, X. GAO et Q. DU, « SSCU-net : spatial–spectral collaborative unmixing network for hyperspectral images, » *IEEE Trans. Geosci. Remote Sensing*, t. 60, p. 1-15, 2022.

[97] F. KHAJEHRAYENI et H. GHASSEMIAN, « Hyperspectral unmixing using deep convolutional autoencoders in a supervised scenario, » *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, t. 13, p. 567-576, 2020.

[98] D. HEINZ, C.-I. CHANG et M. L. ALTHOUSE, « Fully constrained least-squares based linear unmixing [hyperspectral image classification], » in *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No. 99CH36293)*, IEEE, t. 2, 1999, p. 1401-1403.

[99] D. C. HEINZ *et al.*, « Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery, » *IEEE transactions on geoscience and remote sensing*, t. 39, nº 3, p. 529-545, 2001.

[100] R. ACHANTA, A. SHAJI, K. SMITH, A. LUCCHI, P. FUA et S. SÜSSTRUNK, « SLIC superpixels compared to state-of-the-art superpixel methods, » *IEEE transactions on pattern analysis and machine intelligence*, t. 34, nº 11, p. 2274-2282, 2012.

[101] B. PALSSON, J. R. SVEINSSON et M. O. ULFARSSON, « Blind hyperspectral unmixing using autoencoders : a critical comparison, » *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, t. 15, p. 1340-1372, 2022.

[102] G. HINTON, N. SRIVASTAVA et K. SWERSKY, « Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, » *Cited on*, t. 14, nº 8, p. 2, 2012.

[103] Z. WU, S. PAN, F. CHEN, G. LONG, C. ZHANG et S. Y. PHILIP, « A comprehensive survey on graph neural networks, » *IEEE transactions on neural networks and learning systems*, t. 32, nº 1, p. 4-24, 2020.

Liste des Travaux

Conférences internationales à comité de lecture

D. BOILDIEU, D. HELBERT, E. CHAMPION, A. MAGNAUDEIX, P. LEPROUX et P. CARRÉ, « Segmentation integration in multivariate curve resolution applied to coherent anti-Stokes Raman scattering, » in *2021 Conference on Lasers and Electro-Optics Europe European Quantum Electronics Conference (CLEO/Europe-EQEC)*, juin 2021, p. 1-1.

Y. MURAKAMI, S. MIYAZAKI, D. BOILDIEU *et al.*, « Toward whole brain label-free molecular imaging with single-cell resolution sing ultra-broadband multiplex CARS microspectroscopy, » in *Advanced Chemical Microscopy for Life Science and Translational Medicine 2022*, SPIE, 2022, PC119731B.

D. BOILDIEU, D. HELBERT, A. MAGNAUDEIX, P. LEPROUX et P. CARRÉ, « Multivariate curve resolution with autoencoders for CARS microspectroscopy, » in *Computational Imaging Conference, IS&T Electronic Imaging*, 2023.

Journaux internationaux à comité de lecture

D. BOILDIEU, T. GUERENNE-DEL BEN, L. DUPONCHEL *et al.*, « Coherent anti-Stokes Raman scattering cell imaging and segmentation with unsupervised data analysis, » *Front. Cell Dev. Biol.*, t. 10, août 2022.

Méthodes de résolution de courbes multivariées pour la microspectroscopie CARS

Résumé : La visualisation d'échantillons biologiques comme les cellules ou les tissus est une pratique habituelle pour les biologistes. Cette opération nécessite le plus souvent l'ajout de marqueurs pour mettre en évidence les constituants ou molécules d'intérêts. Cependant, l'ajout de ces marqueurs nécessite plusieurs étapes de traitement et altère de manière irrémédiable l'échantillon observé. Une alternative permettant de s'abstraire des marqueurs est la microspectroscopie vibrationnelle. Cette méthode permet d'utiliser le phénomène de vibration des liaisons chimiques pour acquérir un spectre caractérisant la composition chimique de l'échantillon. L'utilisation de cette méthode en plusieurs points permet d'acquérir une cartographie avec une forte richesse spectrale. Afin d'exploiter cette richesse et caractériser au mieux la composition du spécimen étudié, la résolution de courbes multivariées (ou MCR de l'anglais *multivariate curve resolution*) a pour objectif de déterminer les composants présents en caractérisant leur signature spectrale et leur concentration en chaque point de mesure de la cartographie.

De nos jours, la MCR est essentiellement résolue par des régressions linéaires et ne tient pas compte de l'aspect spatial des données. Dans cette thèse, l'application de la résolution de courbes multivariées à des acquisitions de cellules et tissus utilisant la méthode de diffusion Raman anti-Stokes cohérente est introduite. Dans un second temps, une contrainte de segmentation est intégrée au sein de la MCR par l'implémentation d'une segmentation par contour actif. Pour finir, l'utilisation d'auto-encodeurs pour accomplir la MCR et intégré l'information spatiale est étudiée.

Les résultats obtenus ont permis de mettre en évidence la visualisation de différents organites présents au sein de cellules en accord avec l'état de l'art ainsi qu'une caractérisation de leur signature spectrale. L'ajout de la contrainte permet une segmentation efficace de cellules et, combiné avec les résultats sans segmentation, apporte une information supplémentaire pour l'analyse des résultats. L'étude des auto-encodeurs met en évidence leur potentiel pour appliquer la MCR tout en abordant les limites liées à l'initialisation aléatoire des poids du réseau.

Mots clés : Résolution de courbes multivariées, auto-encodeurs, contrainte spatiale, diffusion Raman anti-Stokes cohérente, imagerie cellulaire.

Multivariate curve resolution methods for CARS microspectroscopy

Abstract : Visualization of biological samples such as cells or tissues is a common practice for biologists. This operation usually requires the addition of markers to highlight the constituents or molecules of interest. However, the addition of these markers requires several processing steps and alters the observed sample. To avoid these steps, an alternative allowing is vibrational microspectroscopy. This method allows to use the vibration of chemical bonds to acquire a spectrum characterizing the chemical composition of the sample. The acquisition of several points allows to acquire a cartography with a strong spectral richness. In order to exploit this richness and characterize the composition of the specimen studied, the *multivariate curve resolution* (MCR) aims to determine the components present by characterizing their spectral signature and their concentration at each measurement point.

Nowadays, the MCR is essentially solved by linear regressions and does not take into account the spatial aspect of the data. In this thesis, the application of multivariate curve resolution to cell and tissue acquisitions with coherent anti-Stokes Raman scattering is introduced. In a second step, a segmentation constraint is integrated into the MCR by implementing an active contour segmentation. Finally, the use of autoencoders to accomplish the MCR and integrate spatial information is studied.

The results allowed to highlight the visualization of different organelles present within cells in agreement with the state of the art as well as a characterization of their spectral signature. The addition of the constraint allows an efficient segmentation of cells and, combined with the results without segmentation, brings additional information for the analysis of the results. The study of autoencoders highlights their potential to apply the MCR while addressing the limitations related to the random initialization of the network weights.

Keywords : Multivariate curve resolution, autoencoders, spatial constraint, coherent anti-Stokes Raman scattering, cell imaging.

Univ. de Limoges, CNRS, XLIM, UMR 7252, F-87000 Limoges, France