Université de Lille – Sciences et Technologies

École Doctorale des Sciences de la matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de l'Université de Lille

Discipline :

Optique, Lasers, Physico-Chimie et Atmosphère

par

Silvère ANDRÉ

Apports de la spectroscopie Raman et de la chimiométrie au suivi *in situ* de cultures cellulaires : nouvelles perspectives en biotechnologie

Soutenance le 6 décembre 2016

Rapporteurs	
Dr. Annie MARC	Directeur de Recherche CNRS, Université de Lorraine
Dr. Ganesh D. SOCKALINGUM	Professeur, Université de Reims Champagne-Ardenne
Examinateurs	
Dr. Philippe JACQUES	Professeur, Université de Liège
Dr. Yves LIJOUR	Maître de conférences, Université de Bretagne Occidentale
Dr. Cyril RUCKEBUSCH	Professeur, Université de Lille
Invités	
Dr. Éric CALVOSA	Société Sanofi Pasteur
Dr. Zahia HANNAS	Société Mérial
Directeur de thèse	
Dr. Ludovic DUPONCHEL	Professeur, Université de Lille
Co-encadrant de thèse	
Dr. Olivier DEVOS	Maître de conférences, Université de Lille

Université de Lille – Sciences et Technologies

École Doctorale des Sciences de la matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de l'Université de Lille

Discipline :

Optique, Lasers, Physico-Chimie et Atmosphère

par

Silvère ANDRÉ

Apports de la spectroscopie Raman et de la chimiométrie au suivi *in situ* de cultures cellulaires : nouvelles perspectives en biotechnologie

Soutenance le 6 décembre 2016

Rapporteurs	
Dr. Annie MARC	Directeur de Recherche CNRS, Université de Lorraine
Dr. Ganesh D. SOCKALINGUM	Professeur, Université de Reims Champagne-Ardenne
Examinateurs	
Dr. Philippe JACQUES	Professeur, Université de Liège
Dr. Yves LIJOUR	Maître de conférences, Université de Bretagne Occidentale
Dr. Cyril RUCKEBUSCH	Professeur, Université de Lille
Invités	
Dr. Éric CALVOSA	Société Sanofi Pasteur
Dr. Zahia HANNAS	Société Mérial
Directeur de thèse	
Dr. Ludovic DUPONCHEL	Professeur, Université de Lille
Co-encadrant de thèse	
Dr. Olivier DEVOS	Maître de conférences, Université de Lille

Remerciements

Les différents travaux de recherche présentés au cours de ce manuscrit ont été réalisés au sein du Laboratoire de Spectrochimie Infrarouge et Raman (LASIR, UMR CNRS 8516). Ainsi, je tiens à remercier Hervé VEZIN, Directeur de Recherche CNRS et actuel directeur du laboratoire LASIR.

Je tiens ensuite à remercier Annie MARC, Directeur de Recherche CNRS de l'Université de Lorraine, et Ganesh D. SOCKALINGUM, Professeur de l'Université de Reims Champagne-Ardenne, tous deux rapporteurs de cette thèse, pour leurs remarques et les discussions qui en découlèrent au sujet du manuscrit. Je tiens aussi à remercier Philippe JACQUES, Professeur de l'Université de Liège, pour avoir examiné ce travail de recherche et pour ses enseignements au sein de Polytech'Lille durant la thèse. De même, je remercie Yves LIJOUR, Maître de conférences de l'Université de Bretagne Occidentale, pour avoir accepté d'examiner cette thèse. Au-delà de son rôle de jury, il a su guider mes choix durant mes études de Master. Par ailleurs, j'en profite pour saluer tous les membres de la formation OPEx, notamment Philippe, Alicia, David, Karine. Enfin, je remercie Cyril RUCKEBUSH, Professeur de l'Université de Lille, pour avoir accepté d'être examinateur de cette thèse. En plus de son rôle de jury, il aura été très présent durant toutes mes années au laboratoire, participant beaucoup à la bonne ambiance générale de notre groupe.

Je tiens maintenant à exprimer toute ma gratitude au Professeur Ludovic DUPONCHEL, non seulement pour avoir dirigé cette thèse pendant trois ans, mais aussi pour tous ses enseignements portés depuis mon arrivée au sein du laboratoire, durant mon stage de Master. Il aura été pour moi un véritable guide tout au long de mon parcours depuis mon arrivée dans le Nord. Je remercie également Olivier DEVOS pour avoir encadré cette thèse, avec qui nos échanges ont souvent permis d'avancer.

Je remercie maintenant tous les membres du projet CellPAT, auquel nous avons participé durant l'accomplissement de cette thèse. Je pense particulièrement à Zahia HANNAS et Éric CALVOSA, tous deux présents le jour de ma soutenance, mais aussi à Lydia SAINT-CRISTAU, Sylvain LAGRESLE, Anne-Marie BEAUCHARD et Anthony DA SILVA qui m'ont accueillis dans leurs groupes lors de mes différentes missions. Je n'oublie pas les autres membres, notamment Fanny, Pierre-Alexis, Thierry, Marion, Raphaël, Sabine. Je pense également aux différents stagiaires avec qui j'ai travaillé : Lolita, David, Sophie et j'en oublie surement…

Remerciements

Je tiens maintenant à remercier les membres du petit groupe « Chimiométrie » du LASIR, principalement non-permanents (les permanents ayant déjà été remerciés : Ludovic, Olivier, Cyril). Ainsi, je remercie Maya ABOU FADEL, dit La Princesse, que j'ai côtoyé depuis mon arrivée au laboratoire jusque son départ une fois devenue docteur. Je remercie Romain BERNEX qui m'aura inculqué la culture *gaming*. Je remercie également Mattéo BRYCKAERT, *confrère* et ami du sous-sol souvent venu partager de bons moments autour d'un café, d'un thé ou d'une bière, au sein ou en dehors du laboratoire. Je tiens aussi à remercier Siewert HUGELIER, véritable ami rencontré au LASIR en début de deuxième année de thèse. Ce belgo-belgicain aura toujours été présent dans les meilleurs moments passés depuis maintenant trois ans, notamment lors de longues marches dans Lille, téléphone en main durant la période Pokémon GO, ou encore pendant notre séjour à Chamonix lorsqu'il sauva mes nuits. Toutefois, Siewert reste indissociable de mon ami de longue date, Thomas « Lebiket » SIGNOUR avec qui j'ai beaucoup partagé, entre autres un toit, un check, un serveur Minecraft, de longues soirées en vocal sur Skype, et j'en passe beaucoup car je pourrais facilement réécrire 150 pages pour conter tous nos bons moments.

Je ne manque pas de remercier tous les autres membres que nous avons pu rencontrer au sein de ce petit groupe, surtout Remo (*mio amico italiano*), Amélie (qui veillait sur Thomas, Maya et moi), Julia (petite Russe archi-dynamique qui nous gavait de biscuits), Ya-Juan (ma voisine avec qui j'ai beaucoup appris sur la culture chinoise), Patrizia (petite stagiaire italienne), Bruno (ce fan de catch en fin de doctorat pendant mon stage), et maintenant Rafa avec qui j'ai déjà passé de bons moments.

Je n'oublie pas de remercier tous les membres du laboratoire avec qui j'ai passé de bons moments durant et en dehors des heures de travail : Samantha, Vincent, Matthieu, Myriam, Michel, Émilie, Julien, Gwendoline, Cécilia, Ismail, Lucie, Diksha et j'en passe, la plupart faisant partie du « groupe de l'autre bâtiment d'en face ». J'en profite pour remercier tous les autres membres du laboratoire que je n'ai pas cité.

Je remercie maintenant mes amis, William GUÉRIN (ou *Doctor GUERIN*) ainsi que sa fiancée Élise VANBIERVLIET, ainsi qu'Anne-Charlotte BOZEC (conjointe de Thomas). Will et Thomas sont mes deux amis les plus proches avec qui nous formons une fine équipe depuis nos années brestoises et leurs soutiens à tous pendant les moments difficiles auront été d'une grande importance.

Je remercie aussi ma famille, venue des contrés bretonnes pour assister à ma soutenance, notamment mon frère Cibel, ma sœur Enya, ma mère Myriam et mon père Joseph (et Lulu) accompagnés de mon filleul Liam. Leur présence ce 6 décembre était un réel réconfort, mais c'est pendant tout mon parcours qu'ils auront été les plus importants,

soutenant perpétuellement mes choix et m'encourageant toujours dans les voies que je choisissais.

Enfin, j'exprime mes remerciements les plus chaleureux à ma compagne Carole MORVAN (et la DTDD), pour tout le soutien, les encouragements, le réconfort et la vie que j'ai pu avoir durant toute cette thèse. Ces années auraient été infiniment plus dures sans elle alors merci à ma papuche.

Résumé

Apports de la spectroscopie Raman et de la chimiométrie au suivi *in situ* de cultures cellulaires : nouvelles perspectives en biotechnologie

Dans les années 2000, les grandes instances de sécurité sanitaire ont proposé l'initiative PAT (Process Analytical Technology) pour inciter les acteurs des milieux pharmaceutiques et agroalimentaires à améliorer leurs méthodes de production et de contrôle des produits manufacturés en utilisant de nouvelles techniques innovantes. Cette thèse s'inscrit dans cette démarche et propose d'exploiter la spectroscopie Raman, couplée aux outils chimiométriques pour le suivi en temps réel de cultures cellulaires à visée pharmaceutique. Acquis à l'aide d'une sonde optique à immersion, les spectres Raman in situ permettent de bénéficier d'une vue d'ensemble de l'état biochimique du bioprocédé au cours du temps. Ainsi, en appliquant des outils chimiométriques adéquats, il est possible de tirer profit des informations spectrales générées, notamment pour prédire les concentrations métaboliques au cours du temps.

Les travaux de recherche présentés dans cette thèse proposent tout d'abord de démontrer la nécessité d'optimiser l'acquisition spectrale et le traitement statistique des données pour différentes cultures cellulaires. Des modèles de régression robustes intégrant plusieurs sources de variabilité sont alors développés. Ils prennent en compte les variations inter-cultures, les changements de paramètres « procédés », le repositionnement des sondes optiques, voire le changement de lignée cellulaire. Enfin, ces mêmes spectres Raman sont utilisés pour développer des outils de contrôle statistique des procédés afin de détecter en temps réel des états anormaux comme par exemple la contamination ou encore pour prédire la concentration d'un produit d'intérêt comme les anticorps.

Mots clés : spectroscopie Raman, chimiométrie, cellules – cultures et milieux de cultures, microbiologie pharmaceutique, processus biotechnologiques – surveillance, sondes optiques.

Abstract

Contributions of Raman spectroscopy and chemometrics to *in situ* cell culture monitoring: new perspectives in biotechnology

In the 2000s, the major sanitary safety authorities proposed the Process Analytical Technology initiative (PAT) encouraging pharmaceutical and agri-food industries to enhance their own processes and the control of manufacturing products by using new and innovative techniques. This thesis is directly related to this framework and proposes to use Raman spectroscopy, coupled to chemometric tools, to monitor cell cultures for pharmaceutical purposes. The *in situ* Raman spectra, acquired using an optical immersion probe, allow getting an overview of the biochemical state of the process in time. Thus, by applying the appropriate chemometric methods, it is possible to obtain biological information from the spectra, including the prediction of metabolite concentrations throughout the cultures.

The research work presented here proposes to highlight the necessity to optimize spectral acquisition and statistical preprocessing of the data provided by different cell cultures. Then, robust regression models are developed, taking into account different sources of variability, such as inter-cultures variations, process parameters changes, optical probe repositioning and cell line variations. Finally, these spectra are used to determine the antibody concentration during the culture and to develop new tools for statistical process control of the batches. In this way, any abnormal behavior such as contamination can be detected.

Keywords: Raman spectroscopy, chemometrics, cells and culture media, pharmaceutical microbiology, bioprocesses monitoring, optical probes.

Table des matières

Liste des abréviations

2-OG	2-oxoglutarate
ACP / PCA	Analyse en Composantes Principales / *Principal Component Analysis*
ACPS	*Advisory Committee for Pharmaceutcial Science*
ADN	Acide DésoxyriboNucléique
AKG	α-cétoglutarate
ALA	Alanine
ANTI	*Antibody concentration*
ARN	Acide RiboNucléique
ASN	Aspargine
ATP	Adénosine Triphosphate
BIOM	Biomasse
CCD	*Charge-Coupled Device*
CDER	*Center for Drug Evaluation and Research*
CHO	*Chinese Hamster Ovary*
CMF	Cytométrie en flux
COW	*Correlation Optimized Warping*
CYS	Cystéine
DAS	*Double Antibody Sandwich*
DHAP	Dihydroxyacétone phosphate
DO	*Dissolved Oxygen*
DQM	*Derivative Quotient Math*
EIT	Échelle Internationale de Température
ELISA	*Enzyme-Linked ImmunoSorbent Assay*
F1, 6P	Fructose-bis-phosphate
$FADH_2$	Flavine adénine dinucléotide
FDA	*Food and Drug Administration*
FSC	*Forward Scatter*, Diffusion axiale
G6P	Glucose-6-phosphate
GA	*Genetic Algorithms*
GAP	Glycéraldéhyde phosphate
GLC	Glucose
GLN	Glutamine
GLU	Glutamate
GLY	Glycine
HEK	*Human Embryonic Kidney*
HPLC	*High Pressure Liquid Chromatography*
ICH	*Internation Council for Harmonisation*
IR	Infrarouge
MAL	Malate
MIR	Moyen Infrarouge
MPCA	*Multiway Principal Component Analysis*
MSC	*Multiplicative Scatter Correction*
MSPC	*Multivariate Statistical Process Control*
NADH	Nicotinamide adénine dinucléotide

Nd-YAG	*Neodymium-doped Yttrium Aluminium Garnet*
NH_3	Ammoniac
NH_4^+	Ammonium
NIPALS	*Non-linear Iterative Partial Least Squares*
OCDE	Organisation de Coopération et de Développement Économiques
OMS	Organisation Mondiale de la Santé
OXA	Oxaloacétate
PAT	*Process Analytical Technology*
PC	*Principal Component*
PIR/NIR	Proche Infrarouge / *Near Infrared*
PLS	*Partial Least Squares*
PYR	Pyruvate
QbD	*Quality by Design*
QR	Quotient Respiratoire
RIST	*RadioImmunoSorbent Test*
RMN	Résonnance Magnétique Nucléaire
RMSE	*Root Mean Square Error*
RMSEC	*Root Mean Square Error of Calibration*
RMSECV	*Root Mean Square Error of Cross-Validation*
RMSEP	*Root Mean Square Error of Prediction*
RPE	Résonnance Paramagnétique Électronique
SER	Sérine
SF	*Spodoptera Frugiperda*
SIMPLS	*Straightforward Implementation of the PLS method / Statistically Inspired Modification of PLS method*
SNR	*Signal to Noise Ratio*
SNV	*Standard Normal Variate*
SPC	*Set Point Control*
SPC	*Statistical Process Control*
SSC	*Side Scatter*, Diffusion latérale
SVF	Sérum de Veau Fœtal
SVM	*Support Vector Machine*
TCA	*Tricarboxylic Acid*
TCD	*Total Cell Density*
TD	Turbine à Disques
UPLC	*Ultra Performance Liquid Chromatography*
UV	Ultraviolet
UVE	*Uninformative Variable Elimination*
VCD	*Viable Cell Density*
VIR	*Virus concentration*
WLS	*Weighted Least Squares*

Introduction

En mai 2013 a été lancé le 16ème appel à projet de Recherche et Développement (R&D) du Fonds Unique Interministériel (FUI), ayant pour but de financer des projets de recherche labélisés par les pôles de compétitivité. Ces pôles, issus de la nouvelle politique industrielle française initiée en 2004, s'inscrivent directement dans une économie mondiale plus concurrentielle et ont pour objectif de regrouper et mobiliser différents acteurs autour de thématiques ciblées favorables à la croissance et la création d'emplois sur des marchés porteurs. Les projets proposés lors des appels FUI doivent nécessairement être collaboratifs et associer au moins deux entreprises à un organisme de recherche ou de formation. Au total, 68 des 123 projets présentés lors du 16ème appel du FUI décrochèrent un financement de l'État, pour un total de 51 M€. Parmi eux, le projet CellPAT, labellisé par Lyonbiopôle, décrocha une aide à hauteur de 1,2 M€ pour un budget global de plus de 3,3 M€.

Ce projet s'inscrit directement dans la volonté d'amélioration des techniques de contrôle qualité, initiée par les grandes instances sanitaires au début des années 2000. En effet, l'agence américaine de contrôle des denrées alimentaires et médicamenteuses (ou *US Food and Drug Administration*, FDA) proposa une nouvelle façon de travailler et de contrôler les produits développés par les entreprises en privilégiant la mise en place de procédures de contrôle suivant une démarche scientifique et ordonnée, initiative qui porte le nom de *Process Analytical Technology* (PAT). L'approche classique pour garantir la qualité d'un produit est alors basée sur la maîtrise de la qualité des matières premières et la robustesse des techniques de transformation. Toutefois, au sein des entreprises pharmaceutiques productrices de vaccins, la matière première est variable par nature puisqu'il s'agit de cellules vivantes. Il est donc nécessaire de développer des outils adaptés en fonction des connaissances acquises par l'expérience sur les systèmes mis en jeu, ce qui porte aussi le nom de *Quality by Design* (QbD), pour le contrôle en temps réel de la qualité des produits transformés.

Le projet CellPAT est centré sur cette demande émise par les grandes organisations sanitaires dans le domaine des biotechnologies impliquées dans la fabrication de vaccins. Afin de pouvoir développer des techniques de contrôle qualité performantes et aptes à réaliser le suivi en temps réel des produits manufacturés, il est nécessaire de faire intervenir des technologies innovantes. Pour le suivi et le contrôle des cultures de cellules, il existe déjà certains outils capables de travailler en ligne et en temps réel, notamment en ce qui concerne les conditions de culture (pH, température, etc.). Toutefois, les paramètres

biochimiques importants, tels que les concentrations métaboliques ou les densités cellulaires ne sont pas suivis en temps réel car les méthodes de mesure ne procurent pas de réponse immédiate. Ceci empêche donc toute rétroaction sur les systèmes de culture, notamment à cause du délai qui existe entre l'avancement des cultures et la réception des informations métaboliques. Le projet CellPAT tend donc à développer des méthodes de mesure des différents paramètres métaboliques en direct et sans prélèvement afin d'être pleinement capable de piloter les procédés de cultures cellulaires pour anticiper sur les teneurs du produit attendu.

Trois principaux axes sont au cœur du développement des procédures de pilotage des cultures de cellules, dont tout d'abord la mise en place de capteurs *in situ* dans les bioprocédés de culture, permettant ainsi de réaliser le suivi en temps réel de différents paramètres métaboliques. Ensuite, l'identification de marqueurs spécifiques par étude métabolomique est nécessaire pour caractériser les procédés de culture de cellules et apporter plus d'informations pour la validation des procédés de mesure en temps réel. Enfin, le développement de modèles informatiques des processus de bioproduction prenant en compte les caractéristiques physiques et biologiques des cultures permet d'accéder au pilotage des cultures de cellules en temps réel. Cinq partenaires sont engagés au sein du projet CellPAT, dont trois entreprises biopharmaceutiques : Sanofi Pasteur (Marcy-l'Étoile, Rhône 69, France), Mérial (Lyon, Rhône 69, France) et Transgène (Illkirch-Graffenstaden, Bas-Rhin 67, France). Celles-ci participent notamment aux cultures des cellules dont plusieurs types sont engagés : cellules mammifères *Chinese Hamster Ovary* (CHO) et cellules embryonnaires *Human Embryonic Kidney* (HEK) pour la société Sanofi Pasteur, cellules cancéreuses HeLa pour la société Transgène et enfin, cellules d'insectes *Spodoptera Frugiperda* (Sf9) pour la société Mérial. Les deux autres partenaires du projet sont The CoSMo Company (Lyon, Rhône 69, France), société de modélisation de systèmes complexes, et le Laboratoire de Spectrochimie Infrarouge et Raman (LASIR), unité mixte de recherche située à l'Université de Lille (Villeneuve d'Ascq, Nord 59, France).

La thèse présentée ici s'inscrit directement dans ce projet collaboratif pour la mise en place des capteurs de mesure *in situ* des concentrations métaboliques. Les récentes avancées en matière de développement de systèmes d'acquisitions spectrales permettent de réaliser des mesures au sein même des bioréacteurs de culture. Couplées aux outils chimiométriques (traitement statistique des données), ces méthodes présentent un puissant potentiel pour le suivi en ligne de paramètres biochimiques. En effet, la spectroscopie présente un avantage non négligeable du fait de sa non-destructivité par rapport aux échantillons analysés, ce qui permet donc de développer des capteurs directement immergés dans les bioréacteurs de cultures. De plus, malgré un essor commercial important

des sondes électrochimiques, fluorimétriques et infrarouge (moyen), ce sont les spectroscopies Proche Infrarouge (PIR ou NIR pour *Near Infrared*) et Raman qui sont les nouveaux outils employés par les producteurs biopharmaceutiques pour le dosage d'analytes et de paramètres classiques de contrôle de la croissance cellulaire.

Les sondes optiques associées à ces spectroscopies présentent de nombreux avantages, outre la non-destructivité des échantillons analysés, à travers la rapidité de mesure, ainsi que les caractères écologiques (pas d'utilisation de solvant ou de réactif, donc pas de production de déchets) et économiques (faible coût d'analyse). Toutefois, bien que son application soit encore très limitée en phase industrielle, la spectrométrie Raman présente un avantage non-négligeable sur la spectrométrie NIR en sa capacité à être peu perturbée par les contributions spectrales de l'eau. En effet, les cultures de cellules sont réalisées en milieu aqueux qui présente de faibles signaux Raman pouvant être facilement extraits des analyses. Néanmoins, il faut aussi noter que la spectrométrie Raman est sujette à des phénomènes de fluorescence contrairement à la spectrométrie NIR, mais ces effets peuvent être corrigés par des traitements numériques post-acquisition. Ainsi, ces dernières années, les outils chimiométriques ont de plus en plus systématiquement été appliqués aux données spectroscopiques. En effet, ces puissants outils permettent non seulement de corriger les spectres, mais aussi de traiter les importants volumes de données générés par les différentes méthodes spectroscopiques.

Les travaux de recherche présentés au cours de ce manuscrit présentent donc les différentes études spectroscopiques et chimiométriques mises en œuvre pour réaliser le suivi des paramètres importants des différents bioprocédés de culture cellulaire à l'aide de la spectrométrie Raman et de la chimiométrie. Le premier chapitre est ainsi consacré à la présentation et à l'explication des mécanismes biologiques impliqués au sein de ces travaux de recherche. Nous détaillons également les moyens déployés pour le pilotage des conditions de culture ainsi que les méthodes de référence pour la mesure des paramètres métaboliques et des densités cellulaires. Le deuxième chapitre est dévoué à la spectrométrie et la chimiométrie. Nous reprenons les principes de la spectrométrie Raman, détaillant la théorie et présentant les avancées technologiques en matière de développement de sondes à immersion installées pour l'acquisition de spectres *in situ* sur des bioréacteurs de culture. De plus, nous présentons les outils chimiométriques mis en œuvre pour les traitements spectraux (corrections des spectres) et les modélisations des paramètres métaboliques à travers différentes techniques multivariées. Le troisième chapitre, première partie présentant nos résultats, est dédié à la mise en place de la stratégie d'acquisition des spectres Raman. Nous présentons alors le dispositif de mesure et optimisons plusieurs paramètres critiques tels que le temps d'exposition et le nombre d'accumulations lors des acquisitions spectrales.

Au cours de ce chapitre, nous présentons également certaines contraintes liées à l'utilisation des sondes à immersion, notamment lors de l'installation de celles-ci sur les biogénérateurs. Le quatrième chapitre de ce manuscrit présente les avancées les plus importantes en matière de modélisation des biomarqueurs d'intérêt. D'une part, nous présentons les modèles de régression développés pour chaque type de cellule, considérant différentes sources de variabilité, ensuite nous évaluons l'influence des conditions physiques de culture (température et pH du contenu du bioréacteur, vitesse de rotation de l'hélice pour l'agitation du milieu de culture, niveau d'oxygénation). Il est indéniable que ces-dernières influent sur les cultures notamment au niveau des concentrations des métabolites et donc sur les densités cellulaires. En effet, si les paramètres de culture ne sont pas optimaux, d'une part la production cellulaire n'atteindra pas son niveau maximal, donc la densité cellulaire sera plus faible, et d'autre part les profils métaboliques seront différents. De plus, nous évaluons si ces variations des conditions de culture entrainent des perturbations au niveau du système d'acquisition des spectres. Enfin, toujours au sein du quatrième chapitre, nous évaluons la faisabilité de générer des modèles de régression pour les prédictions des paramètres métaboliques de différents types de cellule. Autrement dit, nous évaluons si la mutualisation de l'information spectrale détenue par plusieurs cultures de différentes cellules permet de développer des modèles de régression robustes aux changements de lignée cellulaire. Enfin, le cinquième et dernier chapitre présente les études complémentaires réalisées à partir des spectres Raman et des données biologiques acquis pour les travaux précédents. Ainsi, nous présentons nos recherches sur la création de modèles de régression permettant de prédire directement la concentration en anticorps ou le titre en virus infectieux présents en solution pendant les cultures cellulaires. Enfin, nous présentons nos travaux sur la mise en place d'un système de détection des déviations des cultures cellulaires par rapport à un état de référence à l'aide des techniques de contrôle statistique des procédés sur la base de données multivariées (ou MSPC pour *Multivariate Statistical Process Control*).

Chapitre 1
Culture de cellules pour la production de particules d'intérêt à visée pharmaceutique

1 Vaccins : les biotechnologies pour la santé

1.1 Enjeux des biotechnologies

Que sont les biotechnologies ? Il est difficile de ressortir une définition unique pour caractériser ce domaine. D'une manière générale, nous pouvons dire qu'elles composent « l'ensemble des méthodes et des techniques utilisant des composés vivants (molécules, organites, cellules, organismes) pour rechercher, modifier ou produire des substances chimiques ou des éléments d'origine végétale, animale ou microbienne ». Cette définition, bien que très vaste, résume l'idée générale des biotechnologies. Par ailleurs, l'administration américaine des produits alimentaires et médicamenteuses (ou *U.S. Food and Drug Administration*, FDA), entre autres responsable de la protection de la santé publique vis-à-vis des produits issus des biotechnologies, va dans le même sens en définissant ces-dernières comme « l'application des systèmes et organismes biologiques dans les techniques et procédés industriels » [1]. D'un point de vue plus économique, l'Organisation de Coopération et de Développement Économiques (OCDE) propose une unique définition accompagnée de sous-définitions afin de faciliter les pratiques statistiques (sondages) sur les biotechnologies. Ainsi, la définition générale indique que les biotechnologies sont « l'application de la science et de la technologie à des organismes vivants, de même qu'à ses composantes, produits et modélisations, pour modifier des matériaux vivants ou non-vivants aux fins de la production de connaissances, de biens et de services » [2]. Cette définition générale, délibérément large, est donc affinée en indiquant le type de technique employée parmi : ADN/ARN, protéines et autres molécules, culture et ingénierie des cellules et des tissus, techniques biotechnologiques des procédés, vecteurs de gènes et d'ARN, bioinformatique et enfin, nanobiotechnologie. À travers ces différentes définitions, nous pouvons cerner et comprendre les enjeux des biotechnologies qui ont, en somme, vocation à améliorer le vivant par le vivant. Il incombe ensuite à chacun de cerner la catégorie de biotechnologie mise en jeu afin de préciser le domaine d'étude. Dans notre cas, nous nous intéresserons notamment à la production de molécules d'intérêt à l'aide de cultures cellulaires pour la fabrication de vaccins.

1.2 Production mondiale de vaccins

Qualifié d'oligopolistique par les analystes financiers, près de 95 % du marché des vaccins est couvert par quelques entreprises (voir Figure 1.1) [3]. Dans la fin des années 1990, ce marché enregistrait environ 3 milliards d'euros de vente (résultats IMS Health™). En 2006, le marché des vaccins représentait 9,5 milliards d'euros, soit 1,7 % du marché global des entreprises pharmaceutiques. À cette époque, les différents organismes d'étude prévoyaient une croissance explosive des ventes de vaccins, tablant entre 16 et 24 milliards d'euros de vente en 2011. Cette forte croissance des ventes serait notamment due à l'augmentation de la demande des pays de la « vieille Europe ». Finalement, les ventes de vaccins avaient été évaluées à 23 milliards d'euros en 2011, puis à 28 milliards en 2014, présentant une croissance annuelle moyenne de 15 %, avoisinant même 24 % entre 2011 et 2014 (résultats Alcimed). Finalement, en 2016, le chiffre d'affaires de l'industrie du vaccin devrait atteindre les 42 milliards d'euros contre 20 milliards attendus pour la même année en 2012.

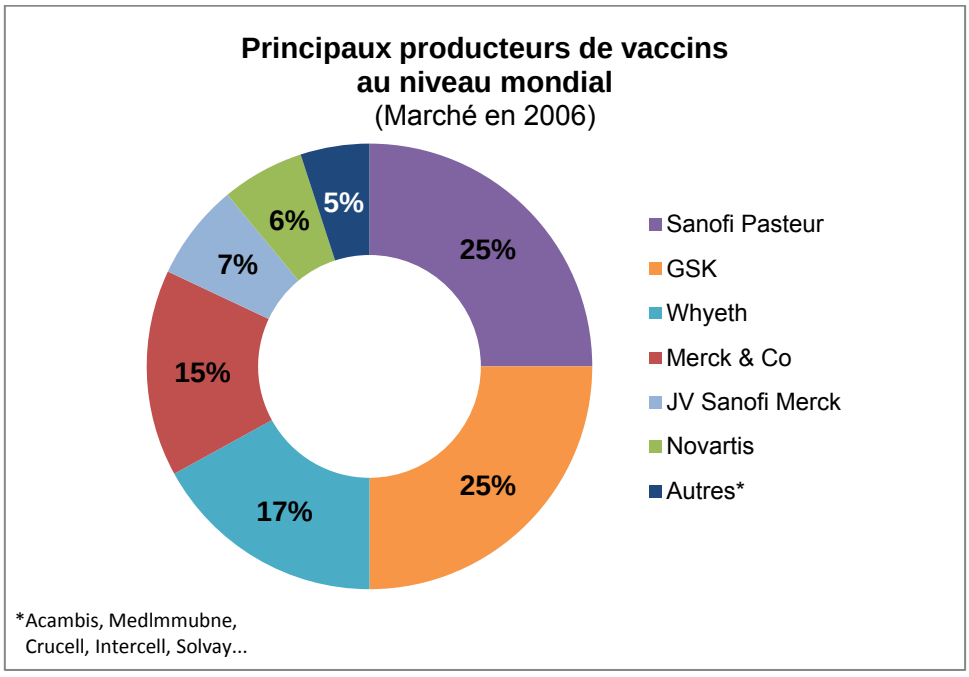

Figure 1.1 Répartition de la production mondiale de vaccins [3].

Ces résultats s'expliquent par la forte appétence du secteur pharmaceutique à l'innovation, améliorant d'une part les vaccins déjà présents sur le marché, tout en développant d'autre part de nouveaux produits innovants capables de répondre à différentes pathologies, qu'elles soient prévisibles ou non (pandémies mondiales ou menaces nées du bioterrorisme). Ainsi, près 20 % du chiffre d'affaires des laboratoires de vaccins sont réinvestis en Recherche et Développement pour concevoir et améliorer les vaccins (données 2014).

2 Fabrication d'un vaccin

Sans tenir compte des phases exploratoires et précliniques, la fabrication d'un vaccin peut prendre entre 6 et 22 mois. La confection du vaccin suit un procédé en deux phases : la première constitue la fabrication biologique et biochimique permettant de produire le système actif du vaccin. La seconde phase est la fabrication pharmaceutique, représentant essentiellement la formulation et le conditionnement du vaccin.

Tout d'abord, il convient de sélectionner les produits actifs du futur vaccin à fabriquer, principalement des virus, des bactéries ou des anticorps, au sein des banques disponibles selon les effets escomptés. Ces produits sont alors cultivés de façon maîtrisée (temps, température, pression, pureté, numération…) et continuellement soumis à des contrôles de qualité. Il faut noter également que la production d'anticorps ou de virus nécessite la culture préalable de cellules soumises à des règles spécifiques de qualité (contrôles des banques cellulaires pour vérifier la qualité des cellules, leur stérilité, leur absence de contamination…). Ceci fait donc de la culture cellulaire une étape initiale déterminante dans la production des produits actifs du vaccin. Une fois cette production terminée, ceux-ci sont récoltés et purifiés, toujours sous contrôle perpétuel afin de vérifier que les produits obtenus soient conformes aux attentes. Pour certains types de vaccins, une étape supplémentaire consiste parfois à inactiver le virus récolté. Ensuite, il convient d'assembler le nombre de valences du produit. D'abord, la valence antigénique est déterminée. Pour rappel, l'antigène est une molécule reconnue spécifiquement par le système immunitaire qui induit la production d'anticorps. Cette valence antigénique est donc le nombre d'anticorps capables de neutraliser l'antigène en question. Ensuite, la valence vaccinale est déterminée. Celle-ci traduit le nombre de maladies contre lesquelles le vaccin est censé lutter.

Suite à cette préparation biologique et biochimique, le vaccin peut ensuite être formulé et conditionné. Il s'agit d'abord de développer la formule du vaccin, à savoir les produits qui seront ajoutés tels que les adjuvants, les stabilisants ou les conservateurs. Le vaccin ainsi formulé est finalement réparti de façon homogène en différentes doses puis parfois lyophilisé. La lyophilisation permet notamment de retirer l'eau contenue dans les produits afin d'assurer une meilleure conservation des vaccins et de stabiliser les plus fragiles. Les produits peuvent alors être répartis en différents lots et distribués.

Les vaccins ainsi distribués sont alors disponibles sur le marché et utilisables. Puisqu'il s'agit de produits liés à la santé des usagers, le contrôle qualité est très strict. Ainsi, le processus précédemment décrit est perpétuellement contrôlé afin d'assurer une bonne traçabilité des produits et permettre aux services de santé de remonter jusqu'aux sources des problèmes en cas d'incident.

3 Cultures cellulaires à visée pharmaceutique

Étape cruciale pour la fabrication de vaccins, nous nous intéresserons en particulier à la culture de cellules au cours de ce manuscrit. Qu'elles soient productrices d'anticorps ou support pour la production de virus, les cellules jouent un rôle majeur dans le procédé de fabrication des vaccins.

3.1 Cultures de cellules pour la production d'anticorps

Que ce soit pour la production de virus ou d'anticorps, la culture de cellules est nécessaire pour la production des particules souhaités. Il existe deux types de cultures différentes. La première est la culture dite adhérente. Celle-ci vise à cultiver les cellules sur un support (fond d'erlenmeyer ou boite de Pétri) jusqu'à atteindre un point de confluence, traduisant l'état de culture dans lequel il n'existe plus aucun interstice entre les cellules. Une nouvelle couche de cellules peut alors être formée et ainsi de suite. L'inconvénient principal de cette technique est induit par l'utilisation éventuelle de sérums qui favorisent l'adhérence des cellules. Ces sérums proviennent d'animaux (exemple du Sérum de Veau Fœtal, SVF, le plus utilisé en culture de cellules) induisant ainsi une variabilité biologique difficile à maitriser au sein du procédé. De ce fait, le risque sanitaire est accru et la mise sur le marché de vaccins réalisés à partir de cultures cellulaires adhérentes est plus difficile.

C'est pourquoi un second type de culture est aujourd'hui favorisé pour les productions de vaccins : la culture en suspension. Les cellules sont introduites au sein d'un milieu de culture (ensemencement) contenu dans un bioréacteur. Le contenu et le type de milieu de culture varient selon la lignée cellulaire employée. Pendant la culture, plusieurs paramètres sont contrôlés et régulés afin d'assurer la stabilité du système tout au long du bioprocédé pour favoriser le développement cellulaire, et donc la production de produits d'intérêt.

Ainsi, pour la production d'anticorps, il convient donc d'optimiser le développement cellulaire tout au long du bioprocédé. Le nombre de cellules vivantes croît de façon exponentielle jusqu'à atteindre un certain palier (voir Figure 1.2). Cette phase stationnaire, aussi appelée « phase plateau », est représentative d'un équilibre au sein de la culture entre le développement des cellules et le début de la phase de « sénescence ». Celle-ci intervient suite à la phase plateau et traduit la dégradation du système biologique entraînant progressivement la mort cellulaire (Figure 1.2). C'est donc avant cette phase que la production d'anticorps est la plus importante. La récolte de protéines lors de productions d'anticorps est donc optimale durant la phase de sénescence puisque la concentration en produit d'intérêt tend à se stabiliser. Il convient donc d'arrêter la culture le plus rapidement possible dans un intérêt principalement économique.

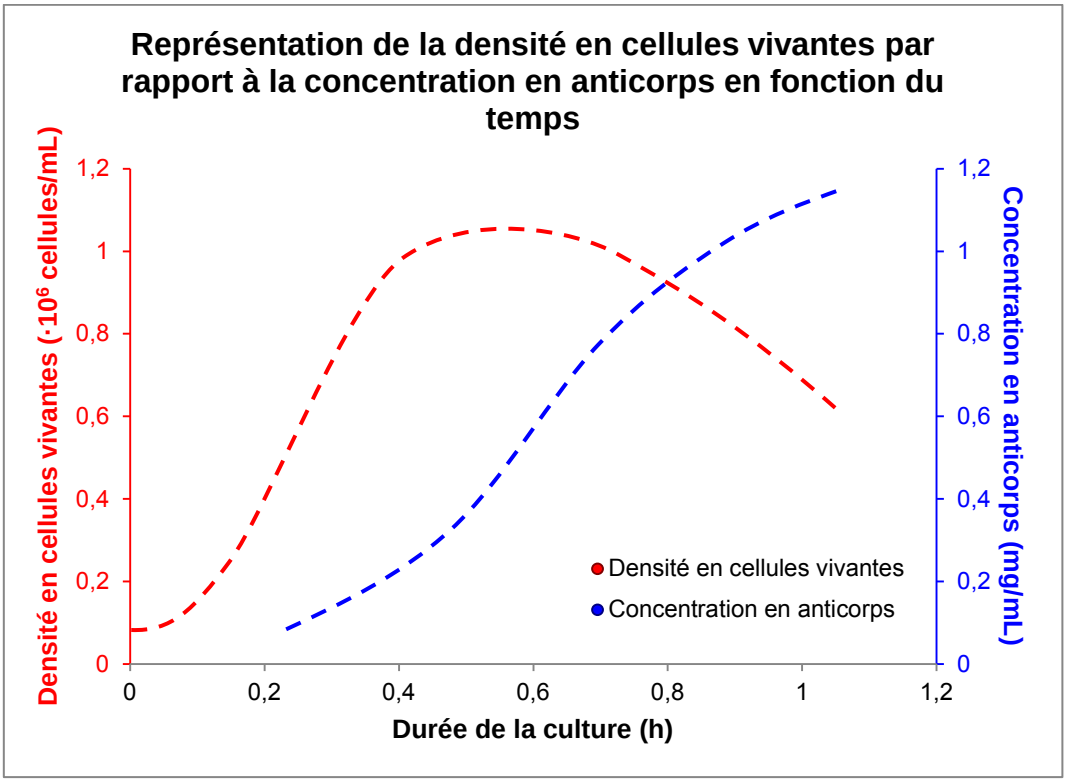

Figure 1.2 Représentation de l'évolution de la densité en cellules vivantes (VCD pour *Viable Cell Density*) et de la concentration en anticorps au cours de la culture cellulaire. Les mesures ont été relevées sur une culture de cellules CHO (*Chinese Hamster Ovary*).

3.2 Cultures pour la production de virus

En ce qui concerne la culture de virus, une culture cellulaire préliminaire est nécessaire. En effet, les cellules servent de support à la prolifération du virus. Ainsi, cette culture cellulaire préliminaire est moins longue que pour la production d'anticorps. L'intérêt majeur est d'obtenir un mélange dont la viabilité (pourcentage de cellules vivantes par rapport au nombre de cellules total) est maximale tout en ayant une forte densité en cellules vivantes. De ce fait, l'infection virale est idéalement réalisée sur une culture cellulaire pré-sénescente, pendant la phase de plateau.

Le virus se présente sous forme de « fragment d'ADN emballé » : le génome viral composé de molécules d'acide nucléique est enveloppé par des molécules protéiques (capside). Le virus est introduit dans le milieu de culture (inoculation virale) et va se développer au sein du milieu en utilisant les cellules comme support (voir Figure 1.3). Une fois la capside greffée à la cellule (étape 1), le patrimoine génétique du virus est libéré. Le virus va alors proliférer dans la cellule en multipliant son génome (étape 2) jusqu'à l'éclatement de cette-dernière (appelé aussi lyse cellulaire, étape 3).

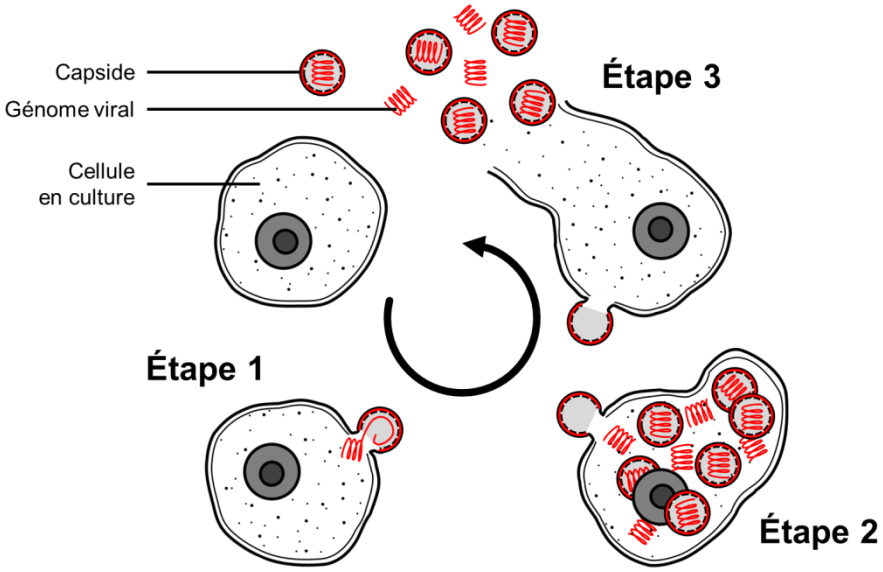

Figure 1.3 Schéma de l'infection virale des cellules par un virus encapsidé.

Les virus nouvellement formés vont alors infecter de nouvelles cellules en culture et ainsi de suite. La concentration en virus évolue de façon exponentielle en fonction du temps jusqu'à ce que le nombre de cellules hôtes ne soit plus suffisant pour permettre la reproduction. La reproduction virale peut alors être stoppée et la récolte peut avoir lieu.

3.3 Lignées cellulaires mises en jeu

Au cours de ce manuscrit, nous travaillerons sur plusieurs types de cellule permettant de cultiver des anticorps ou des virus selon la lignée cellulaire engagée.

3.3.1 Cellules *Chinese Hamster Ovary* (CHO)

Ces cellules mammifères étudiées au cours de ces travaux sont extraites d'ovaires de hamster chinois (ou *Chinese Hamster Ovary*, CHO). Elles sont devenues un véritable standard pour la production de la majorité des protéines thérapeutiques. Ces cellules sont généralement cultivées en ajoutant de temps en temps des nutriments et du milieu dans le bioréacteur (*feeds*), aussi appelé culture en mode *fed-batch*. Les cellules CHO présentent un métabolisme complexe, représenté en Figure 1.4.

Plusieurs réactions caractéristiques interviennent au sein des voies métaboliques des cellules mammifères [4]. La première met en jeu la consommation du glucose par la cellule : la glycolyse [5]. Le glucose assimilé est dégradé en pyruvate tout en produisant de l'énergie au sein de la cellule. Le pyruvate ainsi formé est utilisé dans plusieurs voies. D'une part, en présence de l'enzyme *lactate deshydrogenase*, le pyruvate peut être transformé en lactate et rejeté par la cellule au sein du milieu de culture (glycolyse anaérobique) [5]. Le lactate pourra être consommé par la cellule par la suite afin de produire à nouveau du pyruvate. D'autre

part, cette molécule est assimilée par les mitochondries présentes dans les cellules pour amorcer le cycle de Krebs (ou *TCA cycle* pour *TriCarboxylic Acid*) [6]. Les mitochondries vont ainsi produire de l'adénosine triphosphate (ATP), molécule essentielle fournissant l'énergie nécessaire aux réactions chimiques du métabolisme cellulaire.

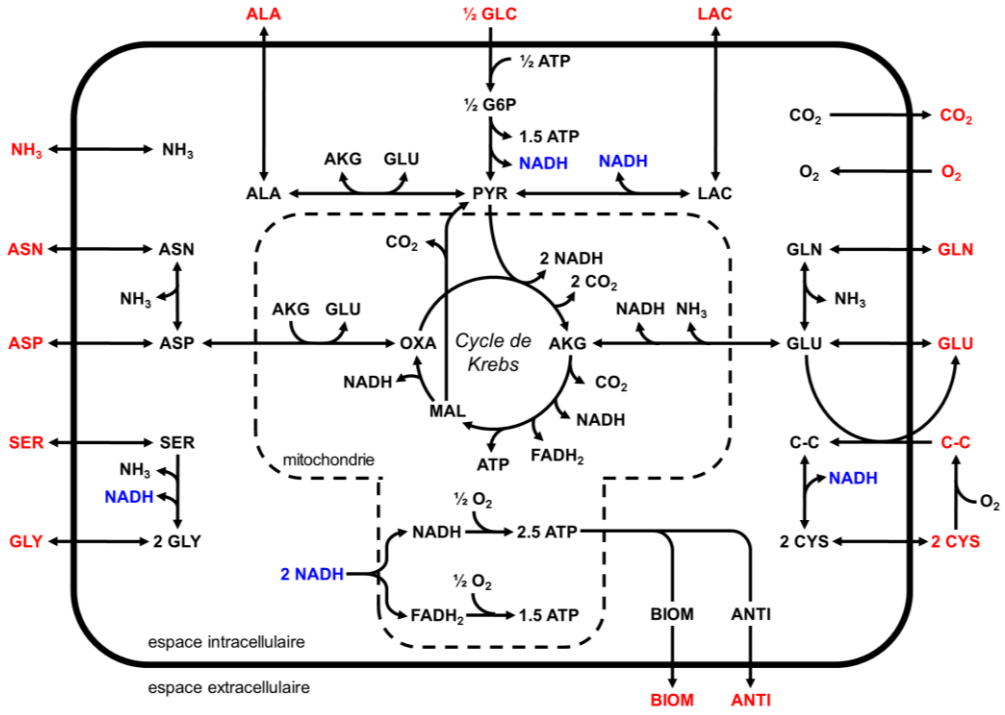

Figure 1.4 Schéma du métabolisme des cellules mammifères. Les métabolites extracellulaires sont représentés en rouge tandis que les métabolites intracellulaires sont représentés en noir, tout comme les molécules NADH mitochondriales. Les molécules NADH cytosoliques (présents dans le liquide cytoplasmique des cellules) sont représentées en bleu [4]. AKG : α-cétoglutarate, ALA : Alanine, ANTI : *Antibody concentration*, ASN : Aspargine, ASP : Aspartate, ATP : Adénosine triphosphate, BIOM : Biomasse, CYS : Cystéine, $FADH_2$: Flavine adénine dinucléotide, G6P : Glucose-6-phosphate, GLC : Glucose, GLN : Glutamine, GLU : Glutamate, GLY : Glycine, MAL : Malate, NADH : Nicotinamide adénine dinucléotide, NH_3 : Ammoniac, OXA : Oxaloacétate, PYR : Pyruvate, SER : Sérine.

Il convient d'ajouter qu'en plus de la consommation du glucose, les cellules vont consommer en partie la glutamine présente dans le milieu de culture pour entraîner la glutaminolyse : les molécules de glutamine sont transformées en glutamate (en présence de *glutaminase*) [7] puis en α-cétoglutarate (dans les mitochondries, en présence de *glutamate deshydrogenase*) [8]. Cette dernière molécule intervient également au sein du cycle de Krebs pour la production d'ATP afin de fournir de l'énergie à l'ensemble de la cellule. Il convient de noter toutefois que les molécules de glutamine et de glutamate sont principalement produites par les cellules au cours de la culture (voir Figure 1.5) dans le cas de cellules mammifères. Les mitochondries vont alors jouer un rôle déterminant dans la production de molécules d'intérêt dans le sens où elles vont apporter à la cellule l'énergie suffisante pour produire des anticorps et favoriser la multiplication cellulaire (voir Figure 1.4).

Dans notre cas, les cellules CHO seront cultivées en mode *fed-batch* pour la production exclusive d'anticorps recombinants. Il faut toutefois noter que nous utiliserons d'autres lignées cellulaires pour étudier le comportement cellulaire lors de la production de virus.

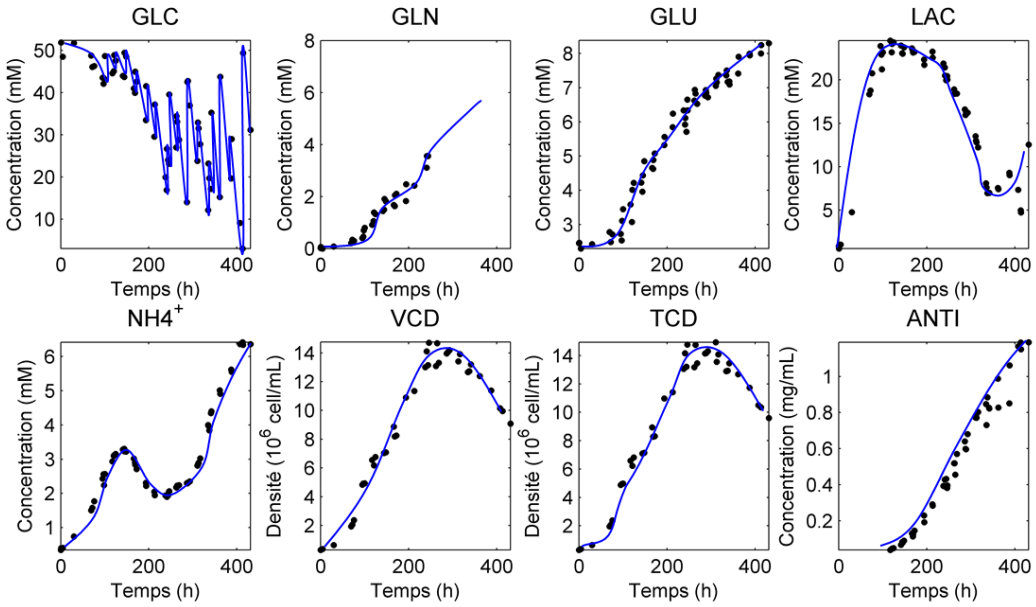

Figure 1.5 Évolution des paramètres biochimiques d'une culture de cellules CHO de plus de 400 h. Les paramètres glucose (GLC), glutamine (GLN), glutamate (GLU), lactate (LAC) et ammonium (NH_4^+) sont mesurés en mM. Les densités totales (TCD) et en cellules vivantes (VCD) sont en millions de cellules par mL. La concentration en anticorps (ANTI) est mesurée en mg/mL. Les variations de glucose sont dues à des ajouts manuels effectués pendant la culture (*feeds*).

3.3.2 Cellules HeLa

Les cellules cancéreuses mises en jeu au cours de ce travail de recherche sont des cellules HeLa. Cette dénomination est directement issue de l'origine de cette lignée cellulaire. Ces cellules, dites « immortelles » [9], proviennent d'un prélèvement effectué sur la tumeur d'une patiente atteinte d'un cancer du col de l'utérus : Henrietta Lacks (He-La). Le comportement de ces cellules est très similaire à celui des cellules mammifères [10]. En effet, le cycle métabolique met en jeu les mêmes voies, notamment la glycolyse, le cycle de Krebs ainsi que la glutaminolyse. La différence majeure de ce type de cellule provient de la forte consommation en glutamine [11] par rapport aux cellules mammifères (dont CHO). En effet, en suivant les paramètres d'une culture cellulaire HeLa (Figure 1.6), nous pouvons facilement observer la consommation du métabolite, contrairement aux tendances observées Figure 1.5 pour la lignée CHO.

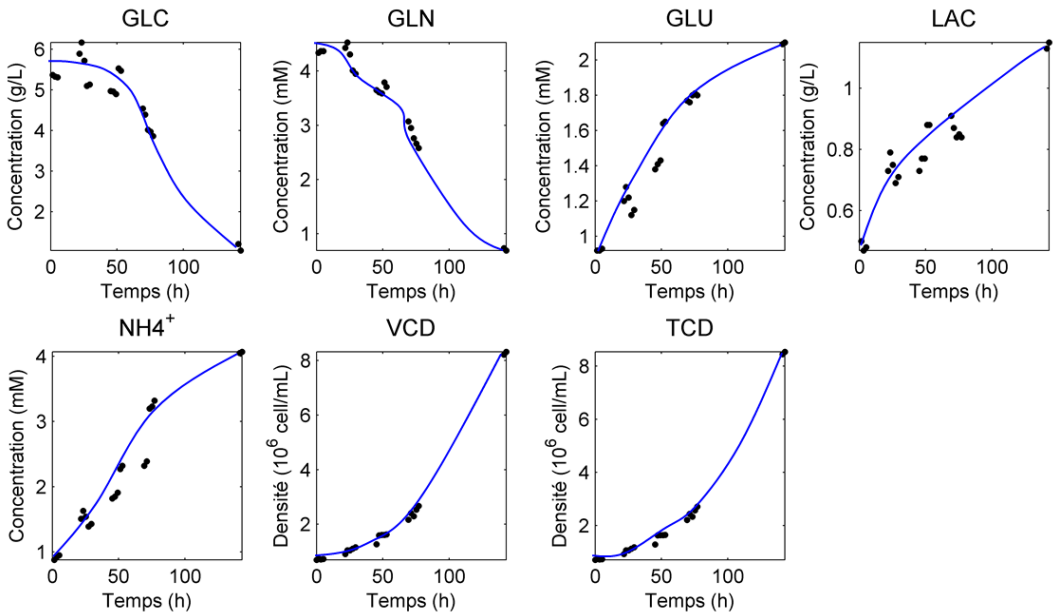

Figure 1.6 Évolution des paramètres biochimiques d'une culture de cellules HeLa de près de 150 h. Les concentrations en glucose (GLC) et lactate (LAC) sont mesurées en g/L. Les paramètres glutamine (GLN), glutamate (GLU) et ammonium ($NH4^+$) sont mesurés en mM. Les densités totales en cellules (TCD) et en cellules vivantes (VCD) sont en millions de cellules par mL.

Les cellules HeLa sont également utilisées pour la culture de virus. Ainsi, suite à la phase de croissance cellulaire présentée, les cultures sont infectées afin d'engager le processus de prolifération virale (Figure 1.3). La culture est arrêtée avant la phase de sénescence afin de recueillir un maximum de cellules vivantes qui serviront de support au développement du virus.

Lors de la phase de croissance de la concentration en virus, il n'y a pas de changement dans l'évolution des métabolites par rapport à la phase de multiplication cellulaire (Figure 1.7), seuls les niveaux de concentration sont différents. En revanche, les densités cellulaires diffèrent : les évolutions sont moins importantes notamment à cause du développement du virus au sein du bioréacteur. Ainsi, le virus se développe jusqu'à ce que le nombre de cellules vivantes ne soit plus suffisant pour servir d'hôte au virus.

3.3.3 Cellules de *Spodoptera frugiperda*

Un autre type de cellule employé pour la culture de virus est la cellule d'insecte. Dans notre cas, il s'agira d'une lignée dénommée Sf9, clone de la lignée Sf21 (plus précisément IPLB-SF21-AE) issue des ovaires de *Spodoptera frugiperda*, ou « légionnaire d'automne », une variété de papillon de nuit. Les voies métaboliques des cellules Sf9 sont similaires à celles des autres lignées présentées ci-dessus (Figure 1.8) [12].

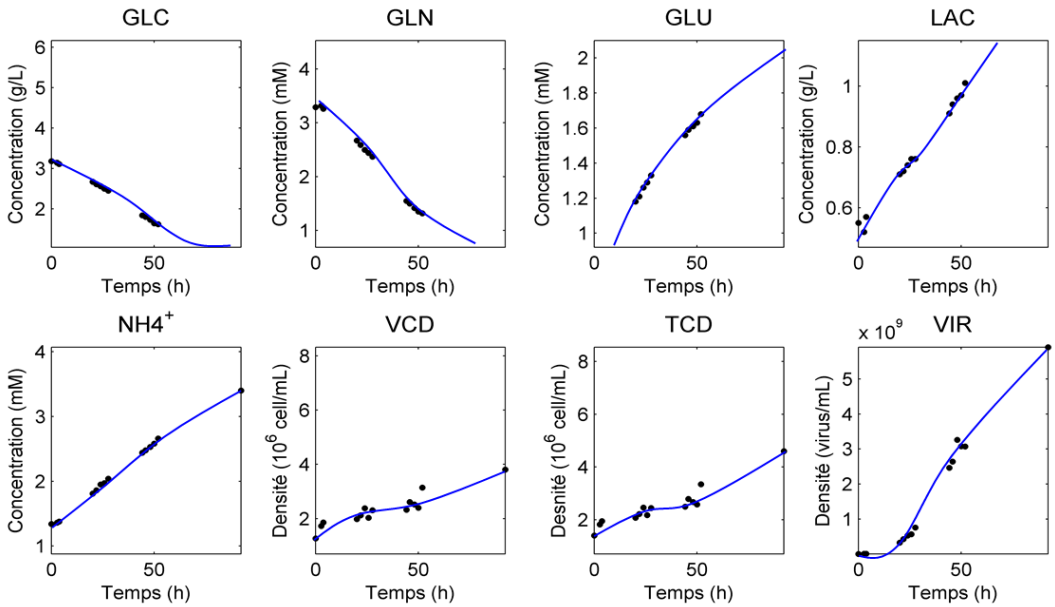

Figure 1.7 Évolution des paramètres biochimiques d'une culture virale ayant pour support les cellules HeLa. Les concentrations en glucose (GLC) et en lactate (LAC) sont mesurées en g/L. Les paramètres glutamine (GLN), glutamate (GLU) et ammonium ($NH4^+$) sont mesurés en mM. Les densités TCD et VCD sont en millions de cellules par mL. La concentration en virus totale (VIR) est indiquée en nombre virus par unité de volume (virus/mL).

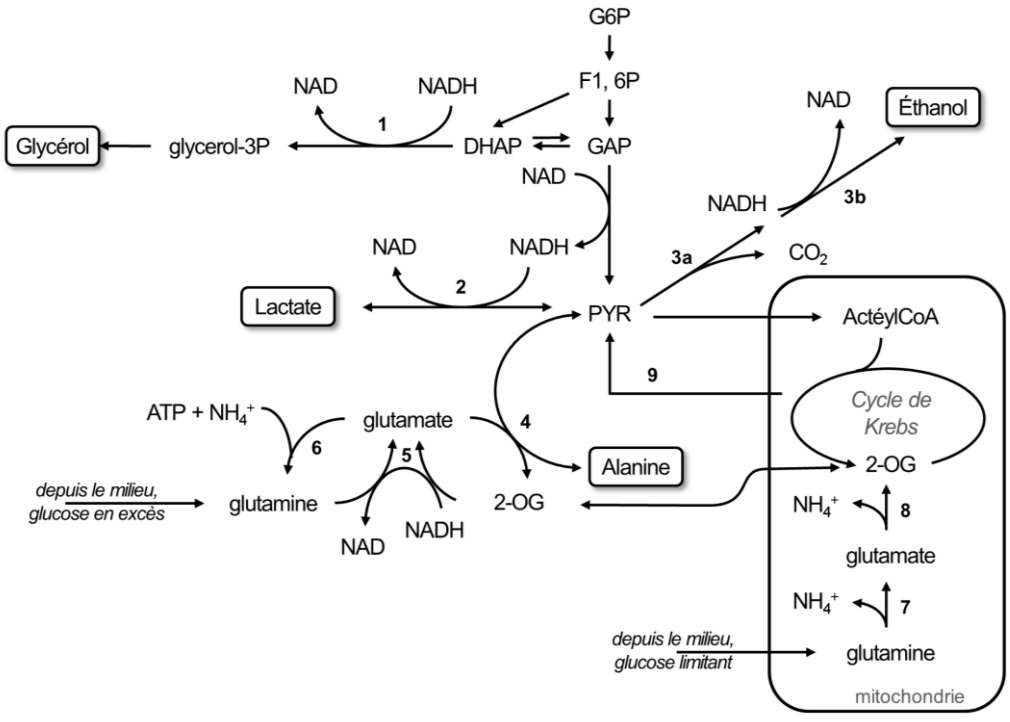

Figure 1.8 Schéma des voies métaboliques des cellules Sf9. Le glucose-6-phosphate (G6P) amorçant la formation de glycéraldéhyde phosphate (GAP) est issu de la consommation du glucose à travers la glycolyse réalisée dans les cellules [12]. 2-OG : 2-oxoglutarate, ATP : Adénosine triphosphate, DHAP : Dihydroxyacétone phosphate, F1, 6P : Fructose-bis-phosphate, G6P : Glucose-6-phosphate, GAP : Glycéraldéhyde phosphate, NAD : Nicotinamide adénine dinucléotide, NADH : Nicotinamide adénine dinucléotide (réducteur), NH_4^+ : Ammonium, PYR : Pyruvate, 1 : DHAP déshydrogénase, 2 : Lactate déshydrogénase, 3a : Pyruvate décarboxylase, 3b : Éthanol déshydrogénase, 4 : Glutamate:Pyruvate transaminase, 5 : Glutamate synthase (NADH-GOGAT), 6 : Glutamine synthétase, 7 : Glutaminase, 8 : Glutamate déshydrogénase, 9 : Flux du cycle de Krebs vers la glycolyse.

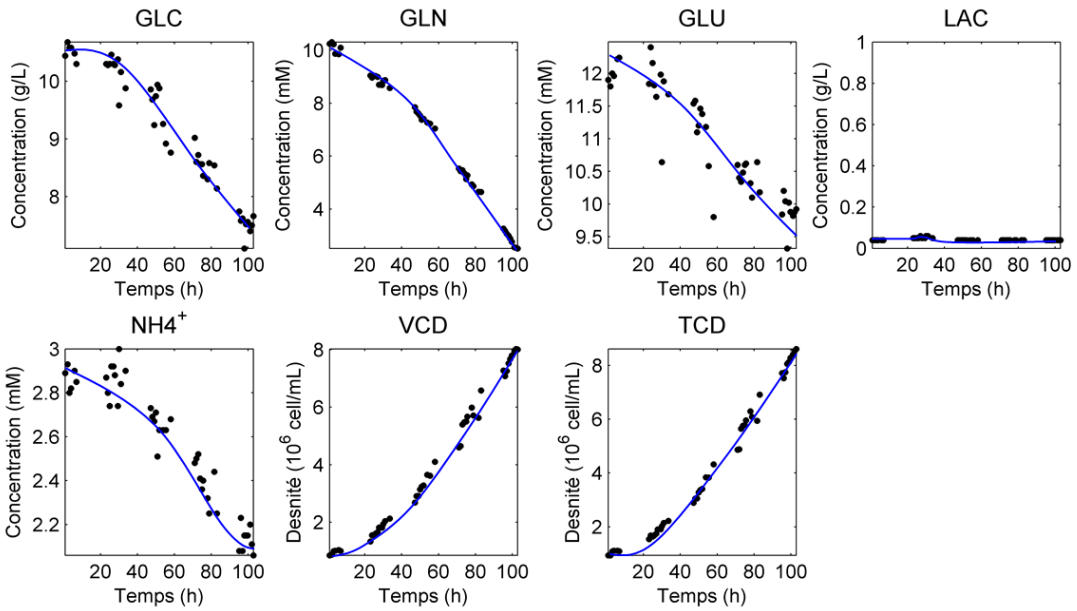

Figure 1.9 Évolution des paramètres biochimiques d'une culture de cellules Sf9 de plus de 100 h. Les concentrations en glucose (GLC) et en lactate (LAC) sont mesurées en g/L. Les paramètres glutamine (GLN), glutamate (GLU) et ammonium ($NH4^+$) sont mesurés en mM. Les densités TCD et VCD sont en millions de cellules.

Lors de la culture de cellules Sf9, celles-ci consomment non seulement du glucose pour amorcer la glycolyse, mais aussi de la glutamine, tout comme les cellules HeLa présentées précédemment. Les profils d'évolution des paramètres biochimiques des cellules Sf9 lors de la phase de croissance cellulaire permettent également de mettre en avant le faible rejet de ces cellules (Figure 1.9). En effet, la concentration en lactate reste en dessous du seuil de détection (quasi-nul) tout au long de la culture tandis que les niveaux de glutamate et d'ammonium tendent à diminuer. Ces deux derniers métabolites sont principalement consommés par la cellule pour la formation de glutamine.

Pendant la culture virale, les cellules Sf9 continuent à se développer jusqu'à atteindre un certain palier stoppant le développement cellulaire (à près de 100 h, Figure 1.10). La densité en cellules vivantes chute, induisant une forte mort cellulaire due à la prolifération du virus. Le nombre total de cellules tend donc à stagner, tout comme la concentration totale en virus. La glutamine n'est plus consommée, ce qui induit la présence de glutamate et d'ammonium en plus forte concentration au sein du milieu de culture [13].

Il convient donc de réaliser la récolte du virus avant la phase de stagnation de la concentration virale afin de limiter la durée du bioprocédé.

3.3.4 Cellules *Human Embryonic Kidney* (HEK)

Les dernières cellules mises en jeu au cours de ces travaux sont issues des reins d'embryons humains, aussi appelées *Human Embryonic Kidney* 293 (HEK-293). Cette lignée

cellulaire HEK-293 provient plus précisément de la cellule HEK transfectée par de l'ADN d'adénovirus pour la rendre immortelle. La désignation 293 pour cette lignée de cellules HEK résulte dans les faits de la numérotation de F. L. Graham pour ses expériences : c'est à sa 293ème expérience que ce dernier considéra établie la lignée cellulaire [14].

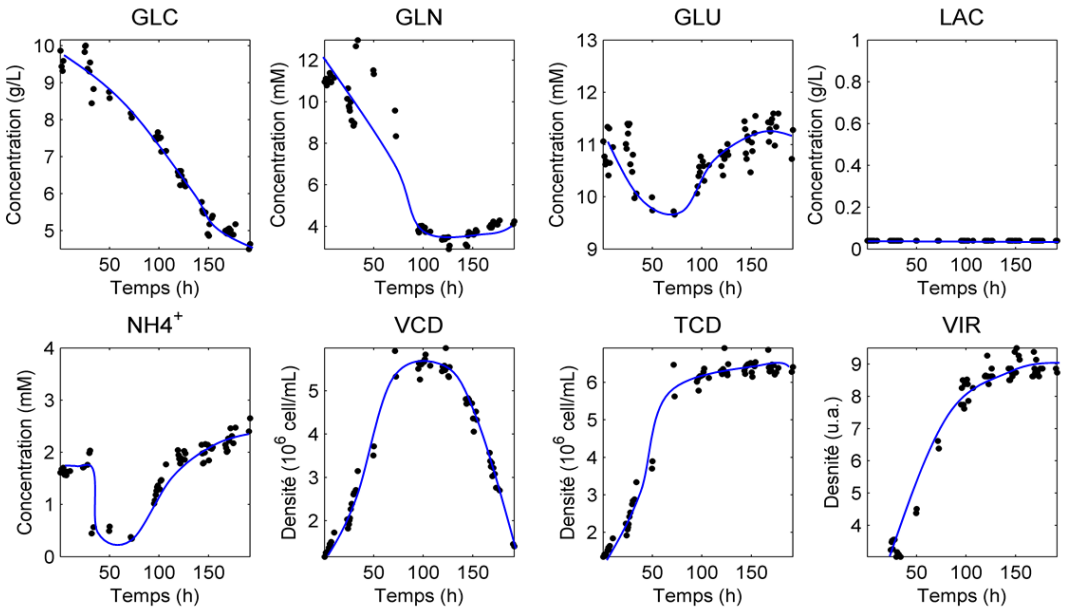

Figure 1.10 Évolution des paramètres biochimiques d'une culture virale ayant pour support les cellules Sf9. Les concentrations en glucose (GLC) et en lactate (LAC) sont mesurées en g/L. Les paramètres glutamine (GLN), glutamate (GLU) et ammonium ($NH4^+$) sont mesurés en mM. Les densités TCD et VCD sont en millions de cellules. La concentration en virus totale (VIR) est indiquée en unité logarithmique.

La cellule HEK-293 présente les mêmes caractéristiques métaboliques que les cellules CHO ou HeLa, à savoir deux principales voies de consommation du carbone : la glycolyse et la glutaminolyse. Ainsi, les évolutions des paramètres biochimiques de ce type de cellule sont similaires aux cellules présentées précédemment (Figure 1.11). Il convient de remarquer la décroissance de la concentration en glutamate pour les cultures HEK-293 présentées. Ceci provient de la substitution de la glutamine par le glutamate pour la dégradation en ammonium lorsque le niveau en glutamine diminue [15].

La lignée HEK-293 est aujourd'hui très employée pour la production de nombreux vecteurs viraux tels que les adénovirus ou virus adéno-associés, les rétrovirus, lentivirus, réovirus, *influenza* (grippe) et autres [15]. La culture de cellules HEK-293 est stoppée avant la phase de plateau de densité en cellules vivantes et la culture est infectée. Suite à l'infection, les évolutions des concentrations des biomarqueurs ne varient pas avant la consommation totale du glucose (Figure 1.12). Une fois le glucose totalement assimilé, la consommation du lactate est observée dans le cas de plusieurs lignées cellulaires [16], dont HEK-293.

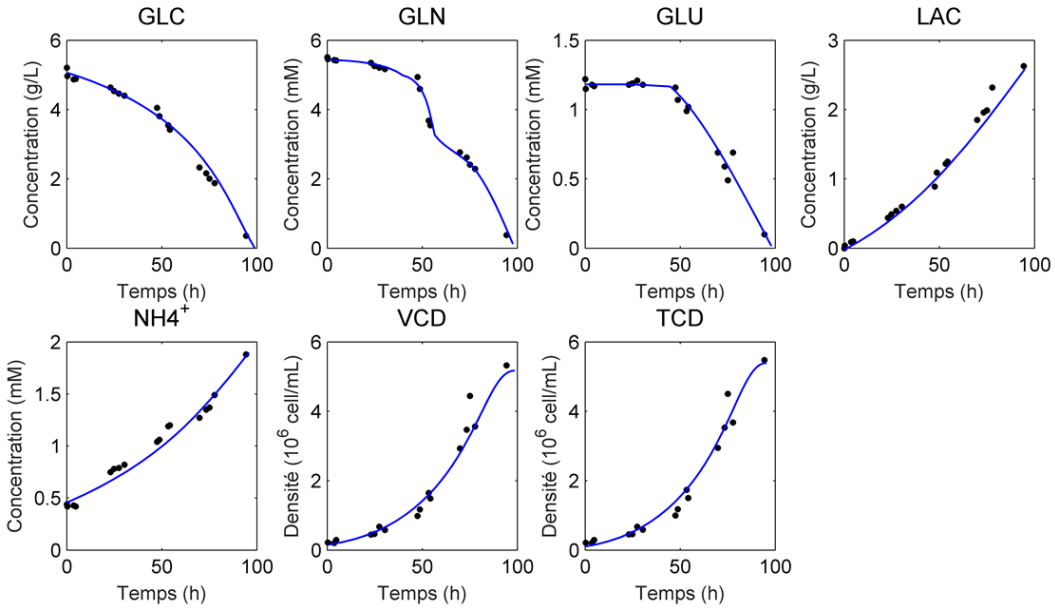

Figure 1.11 Évolution des paramètres biochimiques d'une culture de cellules HEK-293 de près de 100 h. Les concentrations en glucose (GLC) et lactate (LAC) sont mesurées en g/L. Les paramètres glutamine (GLN), glutamate (GLU) et ammonium ($NH4^+$) sont mesurés en mM. Les densités en cellule totale TCD et VCD sont en millions de cellules par mL.

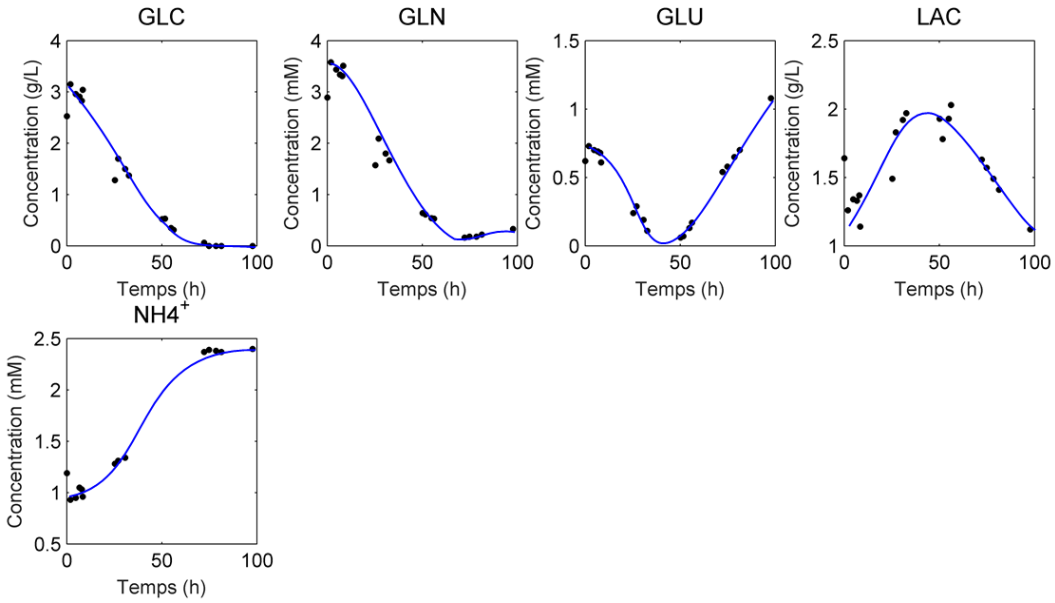

Figure 1.12 Évolution des paramètres biochimiques d'une culture virale ayant pour support les cellules HEK-293. Les concentrations en glucose (GLC) et en lactate (LAC) sont mesurées en g/L. Les paramètres glutamine (GLN), glutamate (GLU) et ammonium ($NH4^+$) sont mesurés en mM. Les densités cellulaires ne sont pas représentées dans ce cas-ci, la méthode de référence manquant de fiabilité.

4 Suivi et contrôle des cultures

Nous avons pu montrer que la culture cellulaire est une étape déterminante dans la conception des vaccins, notamment pour la culture des germes actifs (virus, anticorps). En ce sens, plusieurs paramètres sont contrôlés, pour certains en temps réel, pour d'autres *a posteriori*.

4.1 Régulation des paramètres physiques

Les paramètres physiques contrôlés sont ceux permettant de maintenir un environnement favorable au développement des cellules. En effet, le développement cellulaire est soumis à un certain nombre de paramètres qui doivent être contrôlés. Pour cela, la plupart des procédés sont équipés d'un système appelé *Set Point Control* (SPC, Figure 1.13).

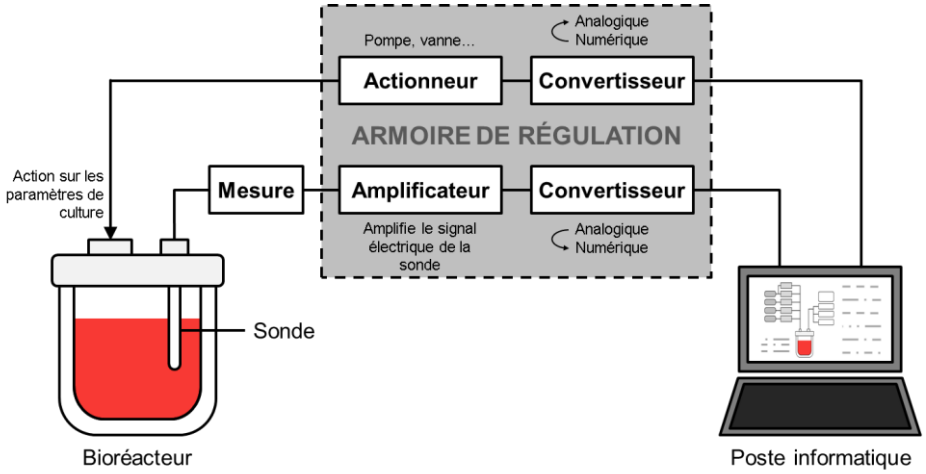

Figure 1.13 Représentation du système de régulation des paramètres : *Set Point Control*.

Celui-ci met en jeu un dispositif de gestion à distance contrôlé via un poste informatique. Ainsi, l'utilisateur peut actionner les différentes commandes permettant de réguler les paramètres suivis depuis le poste à distance. De plus, équipé d'une armoire de régulation, le système devient autonome, ce qui confère au poste informatique un rôle purement de supervision. De ce fait, les risques dus à une panne informatique sont exclus. Les différents paramètres continuent d'être contrôlés par l'armoire de régulation.

4.1.1 Contrôler la température

Un des premiers facteurs influençant la culture cellulaire est la température. En effet, de nombreuses études mettent en avant l'effet de ce paramètre sur différentes lignées cellulaires [17-19]. Il convient donc de contrôler et de réguler ce paramètre. Les technologies actuelles mettent en jeu des thermomètres à résistance de platine, incluant un dispositif de

résistances (pont de Wheatstone) permettant de mesurer la température à travers des changements de résistivité. Dans l'industrie, nous retrouvons couramment les sondes de type Pt-100, caractérisant un thermomètre à résistance de platine possédant une résistance de 100 Ω à 0 °C (selon l'échelle internationale de température de 1990, EIT-90). Pendant la culture, si la température mesurée s'écarte substantiellement de la température cible (quelques dixièmes de degrés Celsius), le régulateur introduira préférentiellement de l'eau de refroidissement ou de la vapeur dans l'enveloppe thermique du bioréacteur afin de diminuer ou augmenter la température du milieu (Figure 1.14).

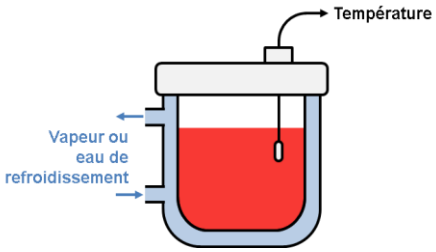

Figure 1.14 Schéma du dispositif de contrôle de la température.

4.1.2 Contrôler le pH

Un autre paramètre influençant la croissance cellulaire d'une culture est le pH [20-22]. Celui-ci est généralement mesuré via la différence de potentiel existante entre deux électrodes : une électrode de mesure et une électrode de référence. La sonde pH, contenant généralement les deux électrodes, est reliée à un pH-mètre indiquant le paramètre en temps réel. De plus, grâce au système de contrôle SPC, le pH est régulé en continu afin de favoriser le développement cellulaire (Figure 1.15). Ainsi, si le milieu devient trop acide, le dispositif introduit une solution basique, telle que KOH ou NaOH. Si le pH est trop basique, du CO_2 gazeux ou une solution acide, généralement H_3PO_4, est alors injecté (à noter qu'HCl n'est pas utilisé car cet acide attaque l'inox, matériau utilisé pour fabriquer les bioréacteurs à gros volumes).

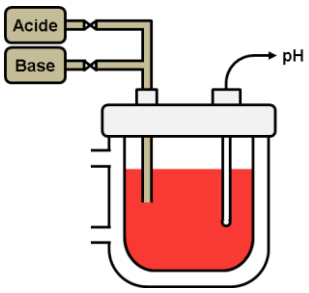

Figure 1.15 Schéma du dispositif de contrôle du pH.

4.1.3 Fournir de l'oxygène

Paramètre également influent sur la culture de cellules [22,23], l'apport en oxygène est essentiel à la respiration cellulaire. L'injection d'oxygène sous forme gazeuse (O_2) dans le milieu s'effectue à l'aide d'un système de bullage (Figure 1.16). Cependant, la difficulté réside dans le contrôle et la solubilisation du dioxygène. En effet, à 30 °C et pression atmosphérique, la solubilité du dioxygène dans l'eau est de 7,55 mg/L et dépend de la

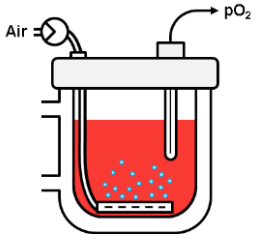

Figure 1.16 Schéma du dispositif d'apport en oxygène (bullage).

pression partielle du dioxygène (loi de Henry), de la température (loi empirique de Truesdale et Downing [24]) ainsi que de la présence ou non de soluté dans le milieu (sels, tensio-actifs, substances organiques). Cependant, il est difficile d'agir sur ces différents paramètres pour augmenter la solubilité du dioxygène dans l'eau sans pour autant dégrader la culture. En effet, en augmentant la pression en gaz dans le milieu, bien que les cellules soient généralement peu sensibles à ces variations, non seulement la solubilité du dioxygène augmente (loi de Henry), mais aussi celle du dioxyde de carbone. En ce qui concerne la température, celle-ci est régulée à un certain seuil et ne varie donc pas. La solubilité n'est donc pas augmentée via l'augmentation de la température [24]. Quant aux différents solutés qui peuvent être mis en solution, les produits organiques n'augmentent que faiblement la solubilité, les sels dissous baissent la solubilité mais sont importants pour la croissance cellulaire et les tensio-actifs créent de la mousse à la surface du milieu. Par conséquent, un autre moyen que l'augmentation de la solubilité du dioxygène est utilisé. Celui-ci met en jeu le coefficient volumique de transfert d'oxygène $k_L \cdot a$ (produit de la constante de transfert du dioxygène en phase liquide k_L et de la surface spécifique d'échange a) [25]. La constante de transfert étant par définition constante, c'est en augmentant la surface d'échange que le dioxygène se solubilisera plus rapidement, soit en augmentant l'agitation (cisaillement des bulles de gaz), soit en augmentant l'aération (plus de bulles), soit en changeant le design du bioréacteur (très peu appliqué dans l'industrie). La quantité d'oxygène dissous est ensuite mesurée à l'aide d'un capteur à oxygène dissout (DO pour *Dissovled Oxygen*), qui indiquera la pO_2 du milieu de culture cellulaire. Le niveau de dioxygène dans le milieu est stabilisé afin de conserver une croissance cellulaire correcte et optimale.

4.1.4 Assurer un mélange homogène

Afin d'assurer l'homogénéité du milieu, un système d'agitation est monté sur les bioréacteurs de culture cellulaire. Nous pouvons retrouver des travaux comparatifs des systèmes d'agitation pour les cuves dès les années 1950 [26]. De nos jours, le système d'agitation le plus utilisé en fermentation repose sur la turbine à disque, aussi appelée turbine de Rushton [27], à 4 ou 6 pâles (Figure 1.17). Ainsi, équipé d'un moteur, le système permet d'entraîner une agitation radiale, qui présente une force de cisaillement plus importante que l'agitation axiale, ce qui améliore la surface spécifique d'échange a et favorise la solubilisation du dioxygène dans le milieu. Enfin, les systèmes actuels sont équipés de contre-pâles disposées autour du bioréacteur afin de casser l'effet vortex

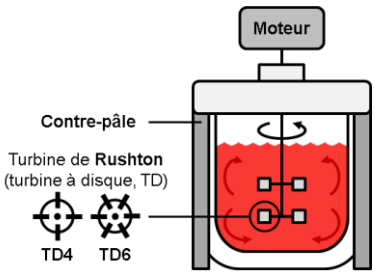

Figure 1.17 Schéma du dispositif d'agitation radiale.

engendré par l'agitation. Concernant les systèmes d'agitation utilisés en cultures de cellules animales, ce sont généralement des hélices marines qui sont employées, permettant d'assurer un mélange homogène tout en limitant le cisaillement des cellules en culture.

4.1.5 Contenir la mousse

Les cultures cellulaires produisent de la mousse à la surface du milieu qui peut engendrer un dépôt de biomasse sur les parois du bioréacteur. La régulation de cette mousse est généralement faite à l'aide d'un capteur de mousse relié à de l'anti-mousse : si le capteur détecte la présence de mousse, un agent démoussant (huile de silicone) est injecté dans le milieu, ce qui permet de contenir ou d'écraser la mousse (Figure 1.18). D'autres systèmes existent mettant en jeu des briseurs de mousse présents directement à la surface du milieu, tels que des disques de mousse [28].

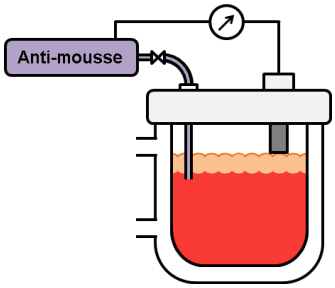

Figure 1.18 Schéma du dispositif anti-mousse.

4.1.6 Maintenir l'environnement stérile

En plus des paramètres précédents à réguler, il est important de maintenir un environnement stérile afin d'éviter une quelconque contamination extérieure de la culture cellulaire. Plusieurs points sont à prendre en compte pour assurer cette stérilité. Le premier point porte sur la qualité des matériaux pour les parois des bioréacteurs. Pour les faibles volumes, le verre est privilégié car facile à nettoyer et peu adhérent. Pour les plus gros volumes, l'inox est préféré car plus résistant. Le deuxième point à assurer pour maintenir le maintien de la stérilité est la résistance du bioréacteur à la pression, croissante lors des phases de stérilisation. Le système doit être capable de supporter des pressions allant de 3 bar à 6 bar (pression épreuve) avant la première utilisation. Le troisième point à contrôler est d'assurer l'imperméabilité du système motorisé pour l'agitation, souvent à l'aide d'un système de garniture d'étanchéité composé de bagues autour de l'axe d'agitation. Enfin, le dernier point à gérer est le traitement de l'air qui vient dans le bioréacteur. Des filtres sont placés aux entrées et sorties d'air du système afin de capter les micro-organismes qui pourraient contaminer la culture. Il s'agit de filtres en profondeur à l'aide de bouchons de coton secs (l'eau risque de créer un film qui laisserait passer les micro-organismes) ou de filtres écran se présentant sous forme de parois microporeuses (diamètres des pores inférieurs à 0,1 µm). Les différents filtres ont ainsi pour objectif d'éviter de laisser entrer les différents micro-organismes dans le bioréacteur en cas de dépression, mais aussi d'éviter de laisser sortir les produits contenus dans le bioréacteur.

4.1.7 Autres capteurs

Aujourd'hui, d'autres capteurs sont installés sur les bioréacteurs afin d'avoir une vue d'ensemble de tous les paramètres suivis pour la régulation de la culture cellulaire. Ainsi, les bioréacteurs peuvent également être équipés de sondes CO_2 ou d'analyseurs de gaz, de chromatographes (gazeux ou liquide) en ligne, ou encore de capteurs indirects. À titre d'exemple, il existe des capteurs calculant le quotient respiratoire de la culture. Dans le cas d'une culture de levure (*Saccharomyces cerevisiae*) en mode *fed-batch*, un capteur indirect permet de suivre le quotient respiratoire (QR) des micro-organismes. Le paramètre QR est maintenu préférentiellement à 1. Au dessus de ce seuil, les cellules produisent de l'éthanol tandis qu'en dessous, la culture est en manque de glucose pour nourrir les cellules. Ainsi, en reliant ce capteur à l'armoire de régulation utilisée dans un système SPC, l'ajout de glucose dans la culture pourra être contrôlé en fonction des mesures du capteur mesurant le quotient respiratoire (Figure 1.13).

4.2 Suivi biochimique des cultures

Nous venons de voir qu'il existe un certain nombre de paramètres physiques suivis et régulés en temps réel afin de maintenir le développement cellulaire dans un environnement favorable. Il est toutefois nécessaire de contrôler également le cycle métabolique des cellules en culture. En effet, bien que l'environnement soit favorable, il est important de vérifier si les cellules présentent un comportement cohérent et consomment normalement les nutriments apportés au sein du milieu de culture. À travers les différentes lignées cellulaires présentées précédemment, plusieurs paramètres biochimiques ressortent majoritairement des voies métaboliques. Il s'agit du glucose, consommé par les cellules, de la glutamine, du glutamate, du lactate et de l'ammonium dont les profils dépendent de la lignée cellulaire mise en jeu, ainsi que des densités TCD et VCD permettant de déterminer la viabilité au sein du bioréacteur. Tous ces paramètres sont représentés sur les Figures 1.5-1.7 et 1.9-1.12.

Contrairement aux paramètres physiques présentés dans la partie 4.1 (Régulation des paramètres physiques), les biomarqueurs ne sont pas régulés en temps réels au cours de la culture. Afin de déterminer les teneurs et niveaux des différents paramètres, des prélèvements sont effectués directement depuis le bioréacteur. Les échantillons ainsi prélevés sont analysés à l'aide de méthodes de références dépendantes du biomarqueur dosé. Cependant, le nombre de prélèvements reste faible, d'un part à cause des risques de contamination du système de culture (interruption de l'herméticité du système lors du prélèvement), d'autre part à cause du coût et du temps des analyses qui empêche parfois d'éventuelles rétroactions, et donc, la régulation en temps réel.

4.2.1 Méthodes de référence pour l'évaluation des teneurs en métabolites

Dans l'industrie biotechnologique, les principaux métabolites sont dosés à l'aide d'appareils automatiques tels que le BioProfile® FLEX développé par la société Nova Biomedical [29], le RX Daytona$^+$ de la société Randox [30] ou encore l'analyseur Cedex Bio de la société Roche CustomBiotech [31]. Il s'agit généralement d'appareils constitués de plusieurs types d'analyseurs permettant ainsi de mesurer, à partir d'un seul échantillon, une large variété de paramètres. En prenant l'exemple du Bioprofile® FLEX, cet outil regroupe huit différents analyseurs (analyseur de gaz du sang, analyseur à électrolytes, analyseur glucose/lactate, kit de test pour les ions ammonium, analyseur glutamate/glutamine, osmomètre) en un seul et unique instrument.

Figure 1.19 Analyseurs automatiques (de gauche à droite) Bioprofile® FLEX (Nova Biomedical), Cedex Bio (Roche CustomBiotech) et RX Daytona$^+$ (Randox) [29-31].

Tout d'abord, plusieurs électrodes potentiométriques permettent de mesurer les ions chargés et donc de mesurer les concentrations en ions ammonium ($NH4^+$), sodium (Na^+), potassium (K^+), mais aussi le pH et la pCO_2. Chaque électrode est équipée d'une membrane sélective selon l'ion analysé. Ainsi, le potentiel mesuré sur la membrane permet de remonter jusqu'aux différents niveaux des paramètres via l'équation de Nernst simplifiée :

$$E_m = E_0 + 2{,}303 \cdot \frac{RT}{nF} \cdot \log_{10} a_0 \qquad (1.1)$$

avec E_m le potentiel mesuré (en volt, V), E_0 le potentiel standard du couple redox de l'ion mis en jeu, R la constante des gaz parfaits, T la température en K, n le nombre d'électrons transférés lors des demi-équations redox, F la constante de Faraday et a_0 l'activité chimique de l'ion d'intérêt. Ainsi, l'activité chimique déterminée permet d'accéder au paramètre souhaité. Les incertitudes de mesure sont estimées à ± 1,5 % pour les ions Na^+, ± 3,0 % pour les ions K^+ et ± 5,0 % pour les ions ammonium à l'aide de ce type d'analyseur.

L'analyseur est également équipé d'électrodes ampérométriques pour la mesure des concentrations en glucose, lactate, glutamine et glutamate, mais aussi pour la détermination de la pO_2 dans l'échantillon. Pour mesurer la pO_2, le principe est basé sur une simple

électrode à oxygène : une cathode de platine à potentiel constant est recouverte d'une membrane perméable à l'oxygène. Puisque l'échantillon analysé diffuse de l'oxygène, celui-ci est réduit à la cathode en passant à travers la membrane. Les variations de courant sont alors directement proportionnelles à la concentration en oxygène dans l'échantillon. Pour les quatre métabolites mesurés, le principe est quasi-similaire. Chacun dispose de sa propre électrode constituée d'une cathode de platine recouverte d'une membrane perméable à l'oxygène et contenant des enzymes. L'oxygène diffusé oxydera les analytes contenus dans la solution, en présence des enzymes, pour produire du peroxyde d'hydrogène (H_2O_2) qui sera oxydé à une anode de platine placée à potentiel constant (voir les équations chimiques ci-dessous, oxydation du glucose à la membrane, équation (1.2), et réaction chimique à l'anode, équation (1.3)). Les variations de courant observées sont alors proportionnelles aux concentrations des molécules dans l'échantillon.

$$\text{Glucose} + O_2 \xrightarrow[\text{oxidase}]{\text{Glucose}} \text{Acide glutonique} + H_2O_2 \tag{1.2}$$

$$H_2O_2 \xrightarrow[0,7\ V]{} 2H^+ + O_2 + 2e^- \tag{1.3}$$

Les mesures effectuées pour les différents métabolites présentent une incertitude estimée à ± 5,0 %. L'incertitude déterminée pour la pO_2 est également estimée à ± 5,0 % de la mesure. Bien entendu, il est également possible de trouver des appareils permettant de mesurer indépendamment chaque métabolite (kit de test pour l'ammoniaque de la société R-Biopharm [32] ou analyseur glucose et lactate 2300 STAT Plus™ de la société YSI [33]).

4.2.2 Techniques de comptage cellulaire

Les mesures des paramètres de densité cellulaire (densité cellulaire totale, TCD, et densité en cellules vivantes, VCD) se font séparément à l'aide de compteurs cellulaires. À titre d'exemple, nous pouvons citer le NucleoCounter® SCC-100™ de la société Chemometec [34] ou encore les séries Vi-CELL® XR de la société Beckman Coulter [35].

Figure 1.20 Compteurs cellulaires (de gauche à droite) NucleoCounter® SCC-100™ (Chemometec) et Vi-CELL® XR (Beckman Coulter) [34,35].

Le principe du NucleoCounter® repose sur la succession de plusieurs étapes avant le comptage de cellules. Dans un premier temps, l'échantillon à analyser contenant les cellules est mélangé à un réactif spécifique contenu dans l'appareil provoquant la lyse des cellules (rupture de la membrane cellulaire) et libérant leurs contenus, dont l'ADN. Les réactifs en question sont généralement des détergents tels que le Triton X-100 dans le cas du NucleoCounter®. Cette lyse préalable des cellules est nécessaire dans le sens où les cellules résistent complètement à l'introduction de substances étrangères. La désintégration membranaire permet dont de libérer le contenu des cellules dans la solution. Par la suite, en ajoutant de l'iodure de propidium (toujours dans le cas du NucleoCounter®), celui-ci viendra s'intercaler entre les deux branches de la chaîne hélicoïdale des molécules d'ADN libérées par les cellules. Enfin, à l'aide d'un système d'imagerie fluorescente, il est possible de détecter à 600 nm les noyaux cellulaires fluorescents colorés par l'iodure de propidium et les compter en traitant les images acquises. Il est important de noter que les incertitudes de mesure pour les différents appareils dépendent du nombre de cellules présentes dans l'échantillon à doser. À titre d'exemple pour le NucleoCounter® SCC-100™, l'écart-type (pour une seule mesure) du comptage cellulaire est de 14,0 % pour 50 000 cellules, 5,0 % pour 400 000 cellules et de 2,2 % pour 2 000 000 de cellules. Toutefois, le fournisseur indique que la précision de la mesure peut être réduite d'un facteur $\sqrt{n}$ en répétant n fois la mesure (comptage).

D'autres techniques existent également pour le comptage cellulaire. Ainsi, la technique de cytométrie en flux (CMF) [36] permet également de déterminer, entre autres, la quantité de cellules dans un échantillon donné. Cette technique permet notamment la caractérisation individuelle, quantitative et qualitative de particules en suspension dans un liquide (Figure 1.21). Les cellules sont alignées à l'aide d'un système de centrage hydrodynamique afin d'être excitées les unes après les autres par un faisceau lumineux (laser).

Ensuite, plusieurs phénomènes physiques résultant de l'excitation des cellules sont mesurés. D'une part, la diffusion des cellules est mesurée dans deux directions : en face de la source et à 90°. La diffusion axiale mesurée en face (mesure FSC pour *Forward Scatter*) permet d'obtenir des informations corrélées à la taille de la cellule mesurée ou à la viabilité cellulaire. La diffusion latérale (mesure SSC pour *Side Scatter*) procure des informations quant à la structure intracellulaire de la cellule. Ainsi, grâce à la mesure SSC, il est possible d'intégrer un système de tri des cellules dans le cas de populations hétérogènes (Figure 1.21).

D'autre part, la lumière absorbée par la cellule permet de remonter au diamètre de cette dernière (en supposant la cellule sphérique) et à l'indice d'absorption des constituants

cellulaires. Enfin, bien que les cellules émettent spontanément de la fluorescence, un fluorochrome est souvent ajouté dans le milieu afin de contrôler préférentiellement l'émission fluorescente. Ainsi, en connaissant le produit ajouté, il est possible de mesurer le taux d'ADN, d'ARN ou de protéines dans la cellule.

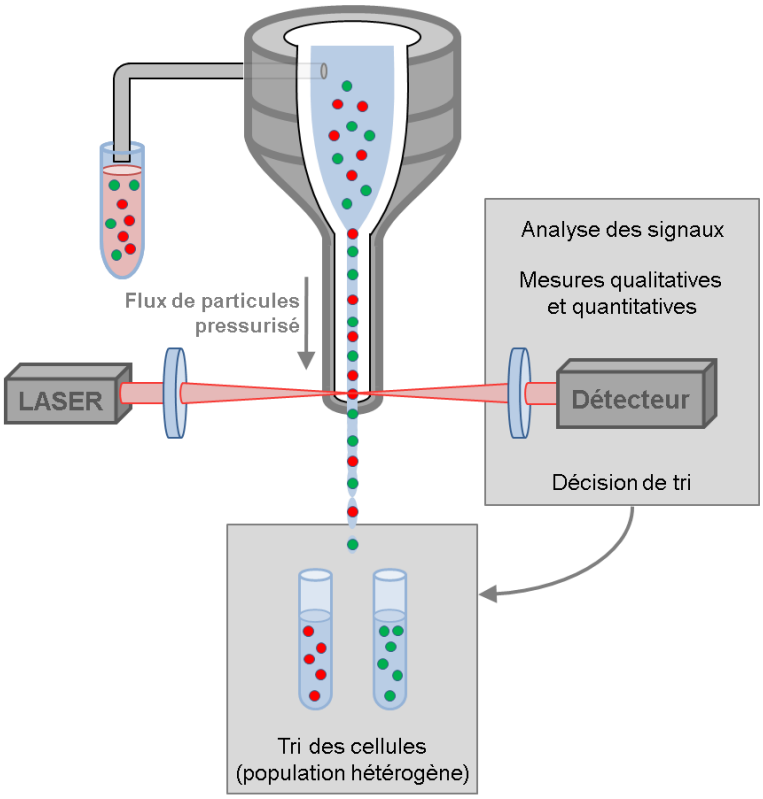

Figure 1.21 Représentation simplifiée d'un cytomètre en flux.

Quant à la détermination du taux de cellules mortes dans un échantillon, celle-ci s'effectue généralement à l'aide d'un test d'exclusion au bleu de trypan [37]. Ce produit est utilisé pour sa faculté à colorer les cellules mortes tandis qu'il est directement relargué par les cellules vivantes. Ainsi, la coloration bleue des cellules mortes permet de les caractériser à l'aide d'un microscope optique pour le comptage. Cependant, la molécule du bleu de trypan étant toxique, elle finit par tuer les cellules vivantes restantes et ainsi recouvrir l'ensemble de la population exposée au colorant. Ainsi, certains appareils ont été développés afin de réaliser ce test de manière directe et automatique pour la détermination des densités cellulaires et de la viabilité d'un échantillon de culture (cas de l'analyseur Vi-CELL®, Figure 1.20).

4.2.3 Mesure de la concentration en anticorps

Afin de déterminer la concentration en anticorps dans les prélèvements effectués sur le contenu du bioréacteur, les techniques chromatographiques sont généralement employées. À titre d'exemple, certains équipements ont été spécialement conçus pour la mesure de

grandes molécules telles que les protéines, les peptides ou les acides nucléiques. C'est notamment le cas du chromatographe liquide haute performance (UPLC pour *Ultra Performance Liquid Chromatography*) ACQUITY® UPLC H-Class Bio développé par la société Waters [38].

L'UPLC est simplement une amélioration des techniques de chromatographie liquide haute performance (HPLC pour *High Performance Liquid Chromatography* ou encore *High Pressure Liquid Chromatography*) : les différents composés présents dans les échantillons sont séparés dans une colonne présentant une phase mobile et une phase stationnaire. La phase mobile permet d'éluer la solution tout au long du passage dans une colonne contenant la phase stationnaire, généralement constituée de silice, silice greffé ou particules polymériques. Les molécules seront alors plus ou moins retenues selon le mode de séparation choisi (adsorption, échange d'ion, exclusion stérique, interactions hydrophiles…).

La notion de pression entre en ligne de compte à partir du moment où il y a des frottements des composés à séparer sur les parois de la colonne et dans la phase stationnaire. Selon l'équation de Van Deemter, équation (1.4), plus les diamètres des particules composant la phase stationnaire sont petits, plus il y aura de molécules à pouvoir circuler dans la colonne mais plus l'écoulement de la solution sera long [39].

$$H = 2\gamma \cdot \frac{D_I}{u} + 2\lambda \cdot d_p + C \cdot \frac{u}{D_I} \tag{1.4}$$

avec H la hauteur équivalente à un plateau théorique (en m), γ le facteur de labyrinthe, D_I la diffusivité moléculaire dans la phase mobile (en $m^2.s^{-1}$), u la vélocité interstitielle (en $m.s^{-1}$), d_p le diamètre des particules (en m), λ un facteur du coefficient de diffusivité de Eddy et C un facteur dépendant des volumes fractionnels des différentes phases, du facteur de distribution et du rapport entre les diffusivités moléculaires de chaque phase. Ainsi, afin de maintenir la vitesse d'écoulement, il est nécessaire d'augmenter la pression du système lorsque la granulométrie de la phase stationnaire décroit. L'élution du composé est alors plus efficace.

C'est en travaillant sur des particules à granulométrie sensiblement plus basses que l'UPLC a permis d'améliorer les techniques d'HPLC, de 3 à 5 µm à moins de 2 µm. De surcroit, l'évolution des technologies a également permis de travailler à des pressions plus élevées pour le maintien de la vitesse d'écoulement malgré des particules plus fines. De ce fait, l'UPLC permet finalement d'obtenir des chromatogrammes mieux résolus plus rapidement. En appliquant des techniques de régression simple sur les signaux des chromatogrammes obtenus, il est possible de remonter aux teneurs des différents

composés. Un système tel que ACQUITY® UPLC H-Class Bio permet notamment d'atteindre la teneur en anticorps avec une incertitude de ± 15,0 % sur la mesure.

4.2.4 Mesure de la concentration en virus

En ce qui concerne la détermination de la teneur en virus dans un échantillon, la technique la plus employée est la technique de « dosage d'immuno-absorption par enzyme liée » ou ELISA (pour *Enzyme-linked immunosorbent assay*). Cette technique fut proposée en 1971 par Eva Engvall et Peter Perlmann [40] pour remplacer les techniques de détermination de titre antigéniques par radioimmuno-absorption (technique RIST pour *Radioimmunosorbent test*) [41] mettant en jeu des marqueurs radioactif directement liés aux antigènes et aux anticorps insolubles. Les liaisons marquées étaient ensuite compétitivement inhibées par des antigènes non-marqués en solution standard ou par des échantillons inconnus.

Par la suite, il a été démontré que marquer des antigènes avec des enzymes adaptées plutôt que des isotopes offrait certains avantages (préparations plus stables sur la durée, équipement plus simple) [40]. C'est sur ce principe que la technique ELISA a été mise au point. Ainsi, à l'aide de substances révélatrices adaptées, il est possible de quantifier le taux d'antigènes d'un échantillon à doser. Une technique très utilisée pour la détermination de la concentration en virus est le test ELISA en sandwich (ou DAS ELISA pour *Double Antibody Sandwich* ELISA, Figure 1.22).

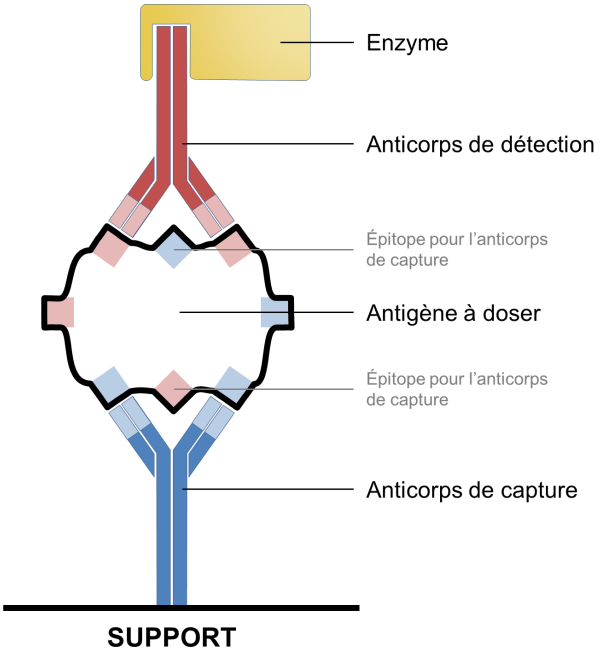

Figure 1.22 Représentation du test DAS ELISA pour la détermination de la teneur d'un antigène dans un échantillon.

Des anticorps de capture spécifiques à l'antigène sont fixés sur le support, généralement des barrettes de microtitration composées de puits. L'échantillon à analyser possédant les antigènes est ensuite déposé. Après une certaine période d'incubation, une solution contenant les anticorps de détection marqués par une enzyme sont fixés. Enfin, en déposant une solution révélatrice dans les puits, les anticorps marqués sont colorés. À l'aide d'un spectrophotomètre il est possible de déterminer l'intensité de la coloration et donc de déterminer la concentration en antigène contenu dans l'échantillon analysé. Dans notre cas, les déterminations obtenues suite au test ELISA varient entre ± 3 % et ± 5 % de la mesure.

Bien entendu, la méthode ELISA peut aussi être appliquée à la détermination du titre en anticorps dans la solution. Cependant, elle nécessite l'utilisation d'antigènes spécifiques aux anticorps dosés pour la réalisation du test. Il en est de même pour l'utilisation de la méthode DAS ELISA et la détermination du titre en virus qui requiert l'utilisation de deux anticorps monoclonaux différents reconnaissant des épitopes de l'antigène différents.

Ainsi, nous pouvons montrer qu'il existe une grande variété de méthodes permettant de mesurer les teneurs des différents biomarqueurs présentés jusqu'ici. Cependant, il est important de noter que toutes ces méthodes sont réalisées à partir de prélèvements effectués sur le bioréacteur de culture. Les échantillons sont alors mesurés par les méthodes de référence hors ligne, *off-line*, en dehors du système de culture.

4.3 Enjeux du *Process Analytical Technology*

Nous avons pu voir dans la partie précédente (4.2 Suivi biochimique) que les déterminations des teneurs des différents métabolites, densités cellulaires ou encore concentrations en molécules d'intérêt sont réalisées à partir de méthodes de référence *off-line*. La mise en place ainsi que les temps d'analyse de ces techniques ne permettent pas de réaliser un suivi en continu des différents biomarqueurs. Cependant, la dynamique actuelle tend vers un contrôle permanent et une connaissance parfaite de l'état d'un système à n'importe quel moment.

C'est l'équipe du centre de recherche et d'évaluation des médicaments (CDER pour *Center for Drug Evaluation and Research*) de la FDA qui propose pour la première fois l'initiative *Process Analytical Technology* (PAT) avec pour objectif de moderniser les contrôles de procédés et les tests dans les entreprises biopharmaceutiques [42]. Avec l'approbation du conseil scientifique de la FDA, le comité consultatif des sciences pharmaceutiques (ACPS pour *Advisory Committee for Pharmaceutical Science*) fut formé en novembre 2001. Il est constitué de plusieurs groupes de travail associant des représentants de la FDA, ainsi que des experts industriels et académiques [43]. Leurs travaux traitaient de

nombreux sujets telles que les opportunités pour l'amélioration de la fabrication de produits pharmaceutiques, les barrières à l'innovation du milieu, les solutions à apporter pour éliminer ces barrières, qu'elles soient réelles ou simplement perçues par les entreprises.

Ainsi, en septembre 2004, la FDA dévoila les résultats des travaux sous forme de guide pour les industries, cadrant et structurant la direction à prendre pour l'innovation en terme de production, de développement et d'assurance qualité dans le domaine pharmaceutique [44]. En encourageant les entreprises à implémenter et à développer de nouveaux outils, plus efficaces et innovants, la FDA compte lever les perpétuelles hésitations des industries pharmaceutiques à introduire de nouvelles technologies pour les contrôles qualité. En effet, le système de régulation des produits biopharmaceutiques est extrêmement rigide et peu favorable à l'importation de nouveaux systèmes. C'est notamment le cas de nombreux procédés, jugés immuables aujourd'hui après avoir été soumis à un grand nombre de procédures réglementaires.

L'objectif final des PAT est donc d'amener les entreprises à mieux connaitre leurs procédés pour mieux les contrôler, ce qui est cohérent face au système de contrôle qualité actuel : « *quality cannot be tested into products; it should be built-in or should be by design* » [44], qui peut être traduit par « la qualité ne peut être testée au sein même des produits, elle doit être intégrée ou présente par nature », aujourd'hui connu sous le terme *Quality by Design* (QbD). Cette idée fut théorisée dans le rapport tripartite du Conseil International à l'Harmonisation (ICH pour *International Council for Harmonisation*), le rapport ICH Q8(R2), résultant de la conférence internationale sur l'harmonisation des exigences techniques pour l'enregistrement des médicaments à usage public (*International conference on harmonisation of technical requirements for registration of pharmaceutcials for human use*), réunissant les membres des délégations de l'Union Européenne, des États-Unis et du Japon [45]. Ainsi, dans cet élan, les équipes de Recherche et Développement, les ingénieurs procédé, les chercheurs ou encore les services réglementaires mettent en place de plus en plus systématiquement le QbD pour les différents procédés de fabrication de produits pharmaceutiques, introduisant de plus en plus de techniques nouvelles.

Chapitre 2
Outils spectroscopiques et chimiométriques au service des biotechnologies

1 Spectroscopie Raman

Les techniques spectroscopiques sont des outils de caractérisation de la matière qui étudient les interactions entre cette matière et les ondes électromagnétiques. Ainsi, il existe une large gamme de techniques spectroscopiques permettant de nombreuses applications. Certaines spectroscopies mettent en jeu les phénomènes liés à l'absorption et l'émission d'un rayonnement électromagnétique sur un échantillon donné. Ainsi, les spectroscopies RMN (Résonnance Magnétique Nucléaire) et RPE (Résonnance Paramagnétique Électronique) permettent respectivement d'observer les transitions de spin pour les noyaux atomiques possédant un spin nucléaire et d'étudier la présence d'électrons non-appariés. Ensuite, en exposant un échantillon à des micro-ondes (longueurs d'onde supérieures à 100 µm), la spectroscopie rotationnelle permet d'étudier les mouvements de rotation de petites molécules et ainsi évaluer leurs structures. En exposant un système à des longueurs d'ondes présentes dans l'ultraviolet (UV) et le visible, spectroscopie UV-visible (longueurs d'onde entre 200–400 nm pour l'UV et 400–750 nm pour le visible), il est possible d'accéder à des informations sur la variation de la distribution électronique de l'échantillon. C'est également le cas de la spectroscopie basée sur l'absorption de rayons X (longueurs d'onde entre 10 nm et 100 pm). La spectroscopie de fluorescence permet quant à elle d'étudier certains types de composés, notamment des molécules cycliques rigides insaturées, en mesurant l'intensité de l'émission fluorescente d'un échantillon exposé à un rayonnement lumineux excitateur (habituellement dans l'UV). Il existe également des spectroscopies vibrationnelles permettant d'accéder aux informations structurales de la matière à travers les vibrations des liaisons des matériaux. Les spectroscopies moyen-infrarouge (MIR, *Mid-Infrared*) et proche-infrarouge (PIR ou NIR pour *Near Infrared*) étudient l'absorbance d'un échantillon exposé à des sources émettant dans le domaine infrarouge (de 750 nm à 2,5 µm pour le NIR et de 2,5 µm à 50 µm pour le MIR). Au cours des travaux de recherche présentés dans ce manuscrit, nous utiliserons un autre type de spectroscopie vibrationnelle

issue de l'émission lumineuse de la matière suite à une certaine excitation, à savoir, la spectroscopie Raman.

1.1 Phénomène de diffusion Raman

C'est en 1928 que la diffusion Raman est rapportée pour la première fois. En effet, C.V. Râman et K.S. Krishnan, physiciens indiens, relatent l'observation d'un nouveau type de radiation de second ordre dans le journal *Nature* [46]. La même année, les physiciens soviétiques L. Mandelstam et G. Landsberg exposent leurs travaux dans *Naturwissenschaften*, sur un nouveau phénomène de diffusion de la lumière dans les cristaux [47]. Ces derniers soulignent que leurs travaux ainsi que le phénomène de C.V. Râman doivent certainement être liés, sans pour autant pouvoir le certifier par manque de description du dit phénomène. Finalement, C.V. Râman rapportera des travaux plus détaillés sur cette nouvelle radiation [48,49] et se verra remettre le prix Nobel en 1930 pour ses travaux sur la diffusion de la lumière et la découverte de l'effet qui portera son nom.

Cette nouvelle radiation rapportée par les différents groupes de recherche provient directement de la diffusion inélastique de la matière suite à l'exposition à un rayonnement électromagnétique. En effet, lorsque la matière, caractérisée par des niveaux d'énergie vibrationnelle, est exposée à un rayonnement lumineux, celle-ci peux présenter différents comportements (Figure 2.1). Si la fréquence de vibration de l'excitation lumineuse ν est égale à la fréquence caractéristique du niveau d'énergie de la molécule constituant la matière observée ν_m, alors un phénomène d'absorption est observé. C'est sur ce principe que repose notamment la spectroscopie infrarouge. En revanche, si la fréquence ν est très supérieure à ν_m, alors des phénomènes de diffusion sont observés.

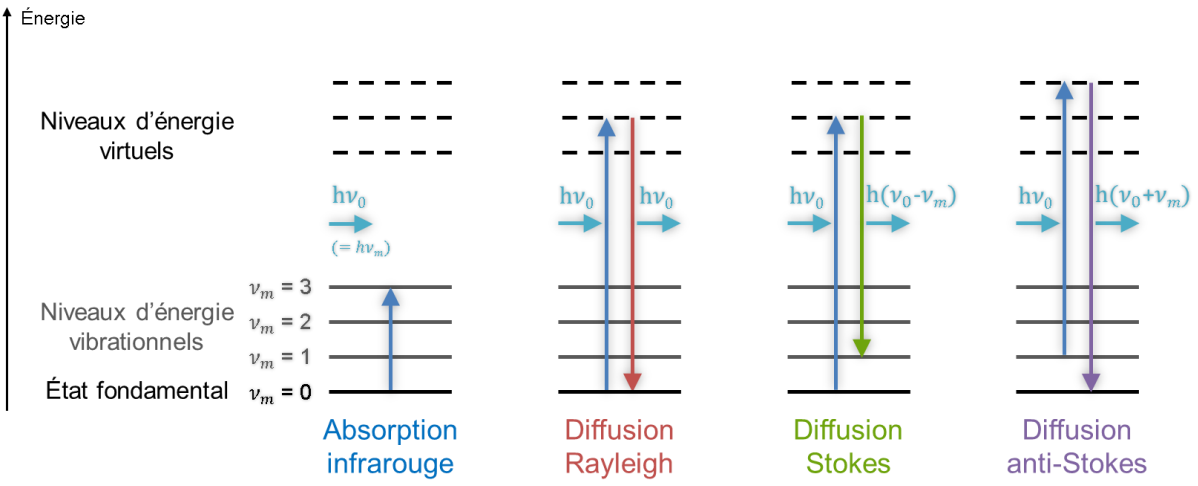

Figure 2.1 Représentation des phénomènes liés à l'excitation d'une molécule sous un faisceau lumineux. ν_0 et ν_m sont des fréquences, h la constante de Planck.

En bombardant les molécules de photons d'énergie hv_0, la plupart seront transmis, réfléchis ou absorbés, tandis qu'une fraction plus faible sera diffusée. La diffusion communément observée est la diffusion Rayleigh (Figure 2.1) : les photons émis sont de même fréquence ν_0 que les photons de la source excitatrice (monochromatique en général, soit une radiation électromagnétique dont la longueur d'onde est unique et connue avec précision). Toutefois, il existe également un phénomène très faible, intervenant sur un photon tous les millions (soit 0,0001 % d'occurrence), appelé rayonnement inélastique Stokes ou anti-Stokes (Figure 2.1) en hommage à G.G. Stokes, physicien irlandais qui décrivit le phénomène en 1852 dans la fluorescence du CaF_2. Les photons émis par ces deux diffusions présentent des fréquences différentes, soit plus faibles ($\nu_d = \nu_0 - \nu_m$, diffusion Stokes), soit plus importantes ($\nu_d = \nu_0 + \nu_m$, diffusion anti-Stokes), toutes deux décalées symétriquement par rapport à la diffusion Rayleigh (diffusion à la longueur d'onde excitatrice), Figure 2.2. Ainsi, la mesure de l'intensité des rayonnements inélastiques (Stokes ou anti-Stokes) permet d'atteindre les informations structurales du composé soumis au rayonnement excitateur sous la forme d'un spectre sur lequel la fréquence des bandes est reliée aux énergies de vibration des liaisons atomiques de la matière.

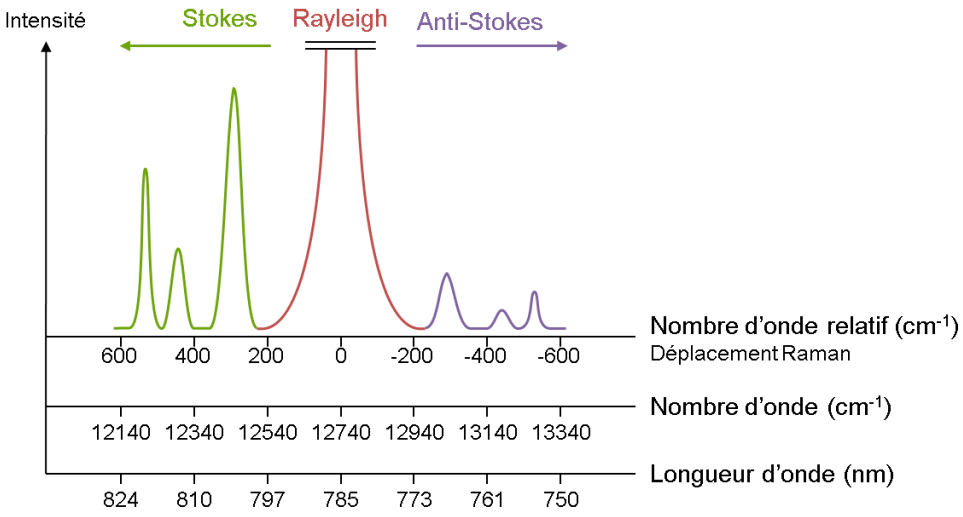

Figure 2.2 Représentation d'un spectre de diffusion. La source est un faisceau monochromatique de longueur d'onde 785 nm.

Pour la mesure des spectres Raman, l'intensité des radiations est proportionnelle au nombre de molécules diffusantes en fonction du déplacement Raman exprimé non pas en fréquence, mais en nombre d'onde $\bar{\nu}$ (cm^{-1}). Cette grandeur physique est calculée en fonction de la fréquence suivant l'équation (2.1)

$$\bar{\nu} = \frac{\nu}{c} = \frac{1}{\lambda} \qquad (2.1)$$

où ν est la fréquence des rayons diffusés, c est la célérité de la lumière et λ la longueur d'onde des rayons diffusés. Ainsi, le nombre d'onde relatif ($\Delta\bar{\nu}$) utilisé pour la représentation du déplacement Raman sur les spectres est l'écart entre le nombre d'onde de la raie diffusée ($\bar{\nu}_d$) et le nombre d'onde de la radiation excitatrice ($\bar{\nu}_0$). De ce fait, la nature du spectre Raman est indépendante de la longueur d'onde ou de la fréquence de la source excitatrice mise en jeu, ce qui permet de travailler sur une gamme étendue de sources lumineuses selon le type et la nature du composé observé.

Bien que les raies Stokes et anti-Stokes soient réparties symétriquement autour de la diffusion Rayleigh (déplacement Raman nul, nombre d'onde relatif égal à 0), la distribution de Maxwell-Boltzmann, équation (2.2), permet de montrer, dans les gammes énergétiques mises en jeu, que les photons diffusés par les radiations Stokes sont statistiquement plus occurrents que ceux provenant des radiations anti-Stokes. En effet, à température ambiante T donnée, avec k la constante de Boltzmann, le nombre d'entités N_1 situées à un état énergétique E_1, est inférieure au nombre d'entités N_0 situées dans l'état énergétique E_0 plus faible que E_1 : selon l'équation (2.2), puisque $E_1 > E_0$, la valeur du terme exponentiel est comprise entre 0 et 1 et donc $N_1 < N_0$.

$$N_1 = N_0 \cdot \exp\left(-\frac{E_1 - E_0}{k \cdot T}\right) \tag{2.2}$$

Puisque pour l'apparition d'une raie anti-Stokes, la molécule se trouve nécessairement dans un état excité (Figure 2.1), le nombre de photons provenant de cette diffusion sera donc moins important que ceux provenant de raies Stokes. C'est pourquoi, les méthodes de mesure Raman conventionnelles sont centrées sur la mesure des signaux provenant des diffusions Stokes.

1.2 Système de mesure Raman

Il existe un grand nombre d'applications du principe de diffusion Raman et ainsi, autant d'appareillages associés. Dans ce travail de recherche, nous nous limiterons à l'utilisation de la spectrométrie Raman conventionnelle, mettant en jeu une source excitatrice, des fibres optiques et un monochromateur permettant de traiter le signal de diffusion Raman et un détecteur pour l'acquisition des spectres.

1.2.1 Source excitatrice

Afin d'observer convenablement le phénomène de diffusion Raman, il est nécessaire d'avoir une source excitatrice suffisamment intense. Ainsi, les lasers, délivrant un faisceau monochromatique intense, sont généralement utilisés en tant que sources dans les

dispositifs actuels. Ce sont les lasers à gaz qui sont majoritairement utilisés pour le grand nombre de longueurs d'onde disponibles. Les premiers datent des années 1960, durant lesquels A. Javan développe le premier laser Hélium-Néon [50], permettant de libérer un faisceau de 15 mW à 1153 nm. Aujourd'hui, la transition la plus utilisée est celle à 633 nm pour des puissances allant de 1 à 100 mW. D'autres lasers à gaz mettent en jeu des particules ionisées tels que Kr^+ ou Ag^+, permettant d'obtenir différentes longueurs d'onde selon le type de gaz mis en jeu. Ces lasers permettent notamment d'obtenir des faisceaux puissants de l'ordre de la vingtaine de Watts. Cependant, la puissante alimentation de ces sources requiert un refroidissement continu généralement effectué à l'aide d'eau ou d'air, ce qui rend leur utilisation contraignante.

Les sources dites solides tendent aujourd'hui à prendre une place importante dans la fabrication de spectromètres. Bien qu'elles n'offrent pas une gamme de longueurs d'onde aussi riche que les lasers à gaz, ces sources présentent de meilleurs rendements. Le laser Nd-YAG [51] (acronyme de *Neodymium-doped Yttrium Aluminium Garnet* signifiant grenat d'yttrium-aluminium dopé au néodyme) permet de travailler sur les longueurs d'onde 1064 nm (infrarouge) et 532 nm (visible, vert). Les lasers titane:saphir ont été présentés plus tard, en 1982 [52], et comportent un cristal de saphir dopé aux ions titane pour milieu amplificateur. Généralement doublés d'une autre source, ces lasers sont réglables entre 650 nm et 1100 nm et permettent de générer des impulsions de l'ordre de quelques femtosecondes, ce qui les rend très utilisés en recherche, notamment pour des expériences spectroscopiques résolues en temps. Toutefois dans l'industrie, les dispositifs Raman présentent généralement des sources solides mettant en jeu des diodes laser émettant à une seule longueur d'onde. Au cours de ces travaux de recherche, c'est une diode laser émettant à 785 nm, pour une puissance maximale de 400 mW, qui est mise en jeu. En effet, en travaillant à cette longueur d'onde, la potentielle fluorescence des échantillons[1] est grandement réduite, tout en procurant un signal Raman suffisamment intense pour être mesuré. Cependant, l'émission des diodes laser est moins directionnelle que les autres sources, ce qui nécessitera la présence d'une lentille ou d'autres composants optiques pour focaliser le faisceau émis. Dans ces travaux de recherche, nous acquerrons le signal Raman à l'aide de sondes à immersion mettant en jeu des fibres optiques pour guider le signal excitateur jusqu'à l'échantillon analysé (section 1.2.3 Sonde à immersion).

[1] Les cultures de cellules présentent une forte tendance à la fluorescence à cause du matériel biologique principalement constitué de composants organiques.

1.2.2 Spectromètre Raman

Suite à l'excitation de l'échantillon, le signal diffusé par l'échantillon est conduit à travers le spectromètre jusqu'au détecteur pour l'acquisition des spectres. Le dispositif du spectromètre dépend principalement du type de détecteur employé. Lorsque le détecteur est monocanal, à savoir un détecteur acquérant chaque fréquence spectrale séparément, le spectromètre est constitué d'un réseau de diffraction rotatif permettant de faire défiler les faisceaux préférentiellement face au détecteur. Chaque élément du spectre est alors acquis indépendamment. Toutefois, ce sont principalement des détecteurs multicanaux qui sont actuellement utilisés dans les dispositifs Raman, permettant ainsi d'acquérir toutes les fréquences du signal diffusé simultanément. Il s'agit de détecteurs type CCD (*Charge-Coupled Device*) [53,54] ou barrettes de photodiodes. Les spectromètres comportent alors un réseau de diffraction fixe précédé d'un système de traitement du faisceau diffusé par l'échantillon, ce qui est notamment le cas des microspectromètres utilisant un microscope (Figure 2.3).

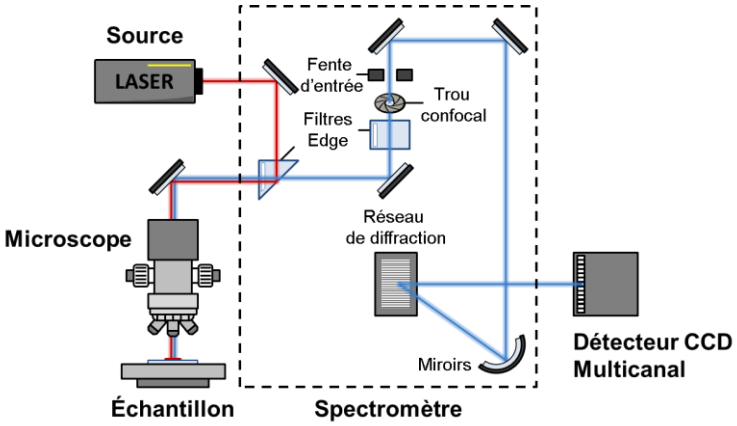

Figure 2.3 Représentation schématique d'un microspectromètre Raman.

Suite à l'excitation de l'échantillon, les rayons diffusés sont orientés d'abord vers un ou plusieurs filtres coupe-bandes pour bloquer les signaux de diffusion Rayleigh. Il peut s'agir de filtres dits « Notch » (provenant de l'anglais, *to notch* : entailler, enchocher) ou bien de filtres « Edge », séparant plus efficacement la bande Rayleigh. Le faisceau passe ensuite généralement par un trou confocal permettant de gérer le volume de travail en faisant varier le diamètre du trou. Un jeu de miroirs entraîne ensuite le faisceau sur un réseau de diffraction ayant pour but de séparer les rayons de différentes fréquences issus de la diffusion Raman. Ces rayons diffractés sont orientés vers le détecteur multicanal pour la mesure de l'intensité des différentes fréquences. Un traitement informatique des informations acquises par le détecteur permet alors de représenter le spectre Raman sous forme d'intensité des signaux en fonction des fréquences relatives à la source excitatrice.

En plus du traitement des rayons diffusés par l'échantillon, l'acquisition d'un spectre Raman dépend d'autres paramètres, dont notamment le temps d'acquisition composé du temps d'exposition multiplié par le nombre d'accumulations (nous parlerons également de *scans*) effectuées. Le temps d'exposition est la durée pendant laquelle le détecteur sera exposé à la diffusion Raman. Plus le détecteur est exposé, plus l'intensité spectrale sera élevée car plus de photons auront été détectés. Toutefois, il faut noter que les détecteurs CCD présentent un niveau de saturation correspondant au maximum de photons pouvant venir frapper la région photoactive du détecteur. La qualité spectrale peut également être améliorée en accumulant le nombre de *scans* Raman. En effet, les spectres possèdent un certain niveau de bruit, phénomène aléatoire lors de l'acquisition. En moyennant plusieurs *scans*, il est alors possible de réduire ce phénomène aléatoire. Numériquement parlant, le rapport signal sur bruit (ou SNR pour *Signal to Noise Ratio*), mettant en jeu l'intensité des signaux et le niveau de bruit, est augmenté d'un facteur $\sqrt{n}$, avec n le nombre de scans moyennés pour obtenir le spectre final. Le temps d'acquisition total d'un spectre Raman est donc la composition du temps d'exposition et du nombre de répétitions, ou *scans,* effectué.

Dans le cas d'un mircospectromètre Raman (Figure 2.3), l'échantillon à mesurer est disposé sur une lame. Le faisceau incident est focalisé sur l'échantillon à l'aide d'un microscope optique, permettant ainsi de travailler à échelle microscopique avec une résolution spatiale de l'ordre du micromètre. Équipée d'une platine XY mobile, il est alors possible de quadriller l'échantillon et d'acquérir un spectre Raman par pixel afin de cartographier celui-ci. Des images de bactéries, de cellules ou encore des agrandissements d'échantillons macroscopiques sont alors accessibles. De plus, sans prendre en compte la platine, il est tout à fait possible de disposer des échantillons liquides sous le microscope pour simplement acquérir des spectres Raman des liquides étudiés. Néanmoins, dans le cas des cultures cellulaires présentées dans le Chapitre 1, un autre type d'appareil de mesure a été développé.

1.2.3 Sonde à immersion

Dans le cadre des cultures cellulaires, il est possible de prélever un échantillon du bioréacteur afin de pouvoir réaliser un spectre Raman représentatif de la culture. Cependant, le fait de réaliser un prélèvement expose le système de culture à l'environnement extérieur, ce qui accroit le risque de contamination de la culture par un agent externe. De plus, suite aux recommandations PAT de la FDA (Chapitre 1, section 4.3 Enjeux du *Process Analytical Technology*), il est nécessaire pour les entreprises pharmaceutiques d'accéder aux informations métaboliques des cultures cellulaires en temps réel. En effet, en réalisant des prélèvements du bioréacteur, il se peut toujours que ces-derniers ne soient pas

représentatifs de la réelle nature de l'ensemble de la culture. C'est pourquoi plusieurs travaux ont été menés afin de réaliser un nouveau type d'appareil de mesure : une sonde à immersion qui, reliée au spectromètre, permettrait d'acquérir des spectres *in situ* du bioréacteur de culture.

Les développements de techniques utilisant des fibres optiques pour les mesures Raman remontent à 1973 [55]. Ceci inspira par la suite la création des premières sondes de mesure basées sur les fibres optiques [56] qui permettaient de conduire la diffusion des échantillons jusqu'au spectromètre. L'avantage de cette technique était surtout de pouvoir acquérir des spectres dans différents environnements, qu'ils soient liquides ou gazeux. Les sondes furent ensuite adaptées pour les mesures *in situ* [57] dans des liquides ou des gaz notamment en incorporant une cellule de mesure à la sonde. Ces travaux furent entrepris en particulier pour les analyses et mesures en ligne de contaminants chimiques dans les eaux de surface, dont le suivi *in situ* et en temps réel est difficile à réaliser par le biais de méthodes externes [57]. Nous pouvons souligner ici le parallèle qui existe entre cette application et le suivi métabolique en ligne de cultures cellulaires.

Dans le début des années 1990, les sondes sont améliorées afin de pouvoir réaliser des mesures plus performantes de la diffusion Raman [58]. En effet, en introduisant différents composants optiques, ces sondes permettaient notamment de soustraire en partie la fluorescence induite et la diffusion Rayleigh. Les sondes furent continuellement améliorées par la suite, que ce soit le gainage de la tête de sonde [59] ou encore l'amélioration des composants optiques pour le traitement des faisceaux lumineux [60]. Le montage actuel des sondes à immersion est inspiré de système Raman pour l'analyse de gaz dans le sang [61]. Ce design, représenté Figure 2.4, permet de les insérer à travers de faibles diamètres tout en conservant un signal Raman suffisant et minimisant les effets parasites [62], ce qui est idéal pour disposer ces sondes sur les platines des bioréacteurs de culture de cellules.

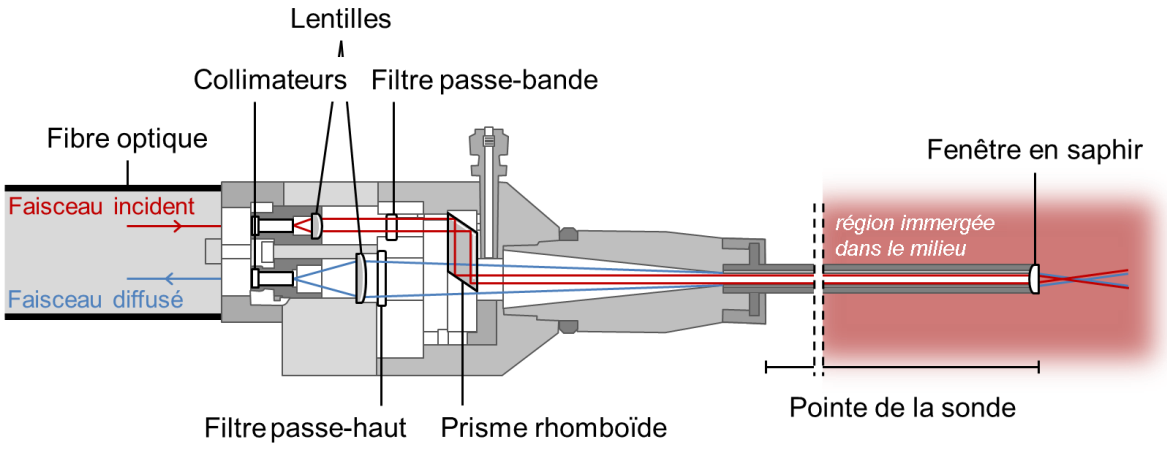

Figure 2.4 Représentation d'un exemple de montage de la tête de sonde à immersion Raman [62].

Le faisceau incident provenant de la source excitatrice traverse une fibre optique reliée au spectromètre puis est collimaté dans la tête de sonde. Un premier filtre passe-bande est employé afin de ne laisser passer que la longueur d'onde de la source excitatrice, ce qui permet notamment d'extraire toutes les sources lumineuses parasites. Toujours dans la tête de la sonde, un prisme rhomboïde est utilisé afin d'aligner le faisceau incident face à la « pointe » de la sonde, immergé dans le milieu. À noter qu'un jeu de miroir peut aussi être employé pour le même effet. La longueur de la pointe reliée à la tête de la sonde dépend du système sur lequel la sonde est disposée. À titre d'exemple, la société Kaiser Optical Systems, Inc. développe des sondes Raman disposables sur différents types de bioréacteurs de culture dont les pointes peuvent mesurer 20 cm (pour des bioréacteurs de quelques litres) ou 42 cm (pour de plus gros volumes, centaines de litres). Au bout de cette pointe, une fenêtre de saphir permet de focaliser le faisceau à quelques millimètres de distance dans le milieu immersif. La diffusion de l'échantillon remonte ensuite la pointe jusqu'à la tête de sonde. Un filtre passe-haut permet alors de conserver uniquement les longueurs d'onde diffusées au-delà d'un certain seuil. Ceci permet notamment de pouvoir extraire les signaux diffusés de même longueur d'onde que la source, à savoir la diffusion Rayleigh. Le faisceau converge finalement vers un collimateur, relié à la fibre optique qui dirigera le signal Raman vers le spectromètre.

Ainsi, l'utilisation de ces sondes permet d'acquérir les signaux Raman *in situ* de différents milieux. Cependant, les spectres Raman des milieux sont parfois complexes et difficiles à interpréter. Par exemple, dans le cas des cultures de cellules, un grand nombre de produits sont présents dans le milieu de culture. Le suivi des métabolites sur la seule base des spectres est alors extrêmement difficile, si ce n'est impossible. Différents outils de traitement des signaux existent pour exploiter au maximum les informations présentes dans des bases de données multivariées, comprenant notamment les jeux de données spectrales. Ainsi, en combinant ces outils statistiques et les données spectroscopiques, il est possible d'analyser et de traiter de manière performante les spectres Raman générés.

2 Outils chimiométriques

La chimiométrie est née de la nécessité d'expliquer de manière simple et efficace les phénomènes complexes en chimie. De ce fait, les chimiométriciens avaient pour but de dégager les informations les plus pertinentes d'un jeu de données complexes, de les traiter et les représenter. Ainsi, la chimiométrie est aujourd'hui définie telle que la discipline de chimie mettant en jeu les mathématiques, les statistiques et la logique formelle pour concevoir de manière optimale les procédures expérimentales, extraire un maximum d'information chimique pertinente à partir des analyses de données chimiques pour ainsi

connaitre et comprendre les systèmes chimiques étudiés [63]. Cette définition s'appuie notamment sur une « arche de connaissances » [64] permettant d'accéder à la connaissance d'un système chimique grâce aux informations recueillies lors de l'expérience (analyse), puis de prévoir de nouvelles expériences en fonction des hypothèses énoncées grâce aux connaissances acquises (synthèse).

Historiquement, l'essor de la chimiométrie est fortement lié à l'utilisation d'ordinateurs dans les investigations scientifiques, début des années 1970. Un groupe de recherche publie alors une série d'articles dans le journal *Analytical Chemistry* mettant en jeu des « systèmes d'apprentissage informatisés » (*computerized learning machines*) afin de résoudre des problèmes en chimie analytique tels que la détermination de formules chimiques à partir de spectres de masse faiblement résolus [65], l'évaluation d'un modèle de classification [66], le développement d'un système de classification à plusieurs catégories [67] ou encore l'étude et l'interprétation de spectres infrarouge [68]. Le terme « chimiométrie » (*chemometrics* en anglais) a été utilisé pour la première fois en 1974 par Svante Wold, par analogie aux termes biométrie (mesure et indentification de caractéristiques biologiques) ou économétrie (estimation et évaluation de modèles économiques tels que la croissance ou la monnaie). Ainsi, le terme chimiométrie permet de mettre l'accent sur l'utilisation de modèles mathématiques, tout comme les autres disciplines « métriques ».

Toutefois, il est important de noter que la chimiométrie n'est pas toujours utilisée dans un unique but de recherche et d'apprentissage sur les systèmes chimiques. En effet, elle permet également de définir plus efficacement les procédés et d'extraire des informations essentielles à partir d'informations chimiques. C'est pourquoi cette discipline est de plus en plus utilisée dans l'industrie afin de mettre en place des systèmes de contrôle qualité plus pertinents. C'est notamment le cas du secteur pharmaceutique depuis l'initiative PAT de la FDA. De plus, les techniques d'analyse permettant aujourd'hui d'acquérir des volumes importants de données, la chimiométrie est également employée afin de représenter ces données de manière simple et efficace.

2.1 Analyse en composantes principales

L'analyse en composantes principales (ACP ou PCA de l'anglais *Principal Component Analysis*), est une technique reconnue comme étant la base de la chimiométrie. Celle-ci permet, à partir d'un jeu de données contenu dans un espace de grande dimension, de condenser l'information en quelques composantes, facilitant ainsi l'interprétation et la représentation de ces données [69]. L'ACP fut initialement formulée par K. Pearson, mathématicien britannique, au début du XX^{e} siècle [70]. Celui-ci présentait alors une interprétation géométrique pour retrouver les droites ou plans permettant de décrire le plus

précisément possible un nuage de points dans un espace. La présentation de l'algorithme d'estimations non-linéaires itératives par les moindres carrés partiels, ou NIPALS (*Non-linear Iterative Partial Least Squares*) par S. Wold en 1966 [71] et les développements de H. Hotelling dans les années 1930 [72] menèrent l'analyse ACP à son stade actuel. C'est autour des années 1960 que la technique ACP fut utilisée en chimie pour la première fois, puis très étendue au domaine après les années 1970 [73].

En considérant que les données soient présentées sous forme de spectres, les différents échantillons peuvent être perçus comme des individus présents dans un espace à n dimensions, n étant alors le nombre de variables des spectres (longueurs d'onde ou déplacements Raman par exemple). L'ACP permet alors de réduire cette dimensionnalité en calculant un nouveau repère à quelques composantes principales (PC pour *Principal Component*) basé sur le maximum de variabilité de l'ensemble initial. L'écriture matricielle, équation (2.3), du développement ACP permet de résumer cette technique.

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{t}} + \mathbf{E} \tag{2.3}$$

Soit les m spectres à n dimensions représentés sous la forme d'une matrice $\mathbf{X}$, le développement ACP permet de représenter les m individus projetés dans un nouvel espace à k dimensions. Ainsi, $\mathbf{P}$ est la matrice contenant l'ensemble des k vecteurs à n dimensions constituant la nouvelle base orthonormée, aussi appelés « *loadings* ». La matrice $\mathbf{T}$ contient les coordonnées factorielles des m individus dans le nouveau repère à k composantes, aussi appelées « *scores* ». Enfin, la matrice $\mathbf{E}$, de mêmes dimensions que $\mathbf{X}$, représente la matrice des résidus obtenus.

Dans la pratique, les composantes sont calculées les unes après les autres à l'aide de l'algorithme NIPALS. L'équation (2.3) peut alors être développée pour les k composantes du système.

$$\mathbf{X} = \mathbf{t_1}\,\mathbf{p_1}^{\mathrm{t}} + \mathbf{t_2}\,\mathbf{p_2}^{\mathrm{t}} + \cdots + \mathbf{t_k}\,\mathbf{p_k}^{\mathrm{t}} + \mathbf{E} \tag{2.4}$$

Les paramètres de la première composante sont définis en prenant en compte la totalité du jeu de données $\mathbf{X}$. Le vecteur des *scores* $\mathbf{t_1}$ est initialisé arbitrairement dans un premier temps. Le vecteur des *loadings* $\mathbf{p_1}$ de la première composante est alors obtenu en régressant la matrice de données $\mathbf{X}$ sur le vecteur des *scores* $\mathbf{t_1}$. Le vecteur $\mathbf{p_1}$ normalisé peut être obtenu suivant l'équation (2.5).

$$\mathbf{p_1} = \frac{\mathbf{X}^{\mathrm{t}}\mathbf{t_1}}{\|\mathbf{X}^{\mathrm{t}}\mathbf{t_1}\|} \tag{2.5}$$

Il est alors possible de recalculer $\mathbf{t_1}$ par projection orthogonale de la matrice $\mathbf{X}$ sur le vecteur $\mathbf{p_1}$, équation (2.6).

$$\mathbf{t_1} = \mathbf{X}\,\mathbf{p_1} \tag{2.6}$$

En reprenant l'équation (2.5), le vecteur des *loadings* $\mathbf{p_1}$ peut être estimé à nouveau et ainsi de suite jusqu'à converger vers une direction fixe ($\mathbf{t_1}$, $\mathbf{p_1}$). Ce premier couple de vecteurs définit la première composante principale de l'ACP. La composante principale suivante est alors estimée suivant une étape de déflation. En effet, l'analyse ACP est une technique dite imbriquée (en anglais *nested*) dans le sens où quelque soit le nombre de composantes qui sera calculé ultérieurement, les paramètres de la première composante principale ne varieront pas. Les paramètres de la deuxième PC sont donc estimés sur les résidus $\mathbf{E}$ obtenus en excluant de l'ensemble $\mathbf{X}$ la partie calculée par la première composante principale :

$$\mathbf{E} = \mathbf{X} - \mathbf{t_1}\,\mathbf{p_1}^{\mathrm{t}} \tag{2.7}$$

Cette dernière équation traduit également le fait que les résidus sont orthogonaux à la direction qui explique le nuage de points définie par les vecteurs des *scores* $\mathbf{t_1}$ et des *loadings* $\mathbf{p_1}$. De ce fait, la composante principale suivante, définie pour expliquer le maximum de variabilité des résidus $\mathbf{E}$, sera nécessairement orthogonale à la première PC. Les k composantes principales de l'ACP sont ainsi déterminées successivement pour retranscrire le maximum d'information du jeu de données initial $\mathbf{X}$ en un minimum k de composantes orthogonales, tel qu'indiqué équation (2.4).

En reprenant le cas des m spectres à n variables, l'ACP permet alors de condenser l'information sur k composantes ($k \ll n$). La dimensionnalité ainsi réduite, il est possible de représenter les spectres de manière plus simple. Par exemple, si nous choisissons un nombre de composantes k égal à 2, il est alors possible de représenter les échantillons sous forme de points dans un graphique en prenant pour coordonnées les *scores* de la première et la seconde composante principale. Ceci permet *in fine* d'avoir une meilleure vue d'ensemble des données pour des analyses plus pertinentes.

2.2 Régression par les moindres carrés partiels

La régression par les moindres carrés partiels, ou régression PLS (pour *Partial Least Squares* ou *Projection to Latent Structures*) est une méthode de régression multiple alliant la décomposition en variables latentes (ou composantes) de l'ACP ainsi que les principes de régression bilinéaire. Cette approche découle directement de l'application de l'algorithme

NIPALS [71] développé en 1966. Dans les années 1970, H. Wold propose l'utilisation de l'algorithme NIPALS pour l'élaboration de modèles de régression basés sur des variables latentes, prémices de la PLS [74,75]. Ainsi, dans les années 1980, l'algorithme PLS est couplé à l'algorithme NIPALS par S. Wold et H. Wold (fils et père) et H. Martens [76] afin d'appliquer cette méthode à des jeux de données dont le nombre de variables dépasse le nombre d'échantillons, tels que des jeux de données spectrales.

2.2.1 Calcul d'un modèle PLS

Bien que pour l'ACP, l'algorithme NIPALS soit très repandu, un autre algorithme vit le jour en 1993 afin d'améliorer les calculs pour les modèles PLS. S. de Jong propose alors l'algorithme SIMPLS (pour *Straightforward Implementation of the PLS method* ou *Statistically Inspired Modification of the PLS method*). L'idée de la méthode de régression PLS est d'expliquer le maximum de variabilité de $\mathbf{X}$ (comme l'ACP) tout en maximisant la corrélation entre les réponses et les données d'entrée à travers un modèle de prédiction tel que $\mathbf{Y} = \mathbf{X}\,\mathbf{B} + \mathbf{E}$ avec $\mathbf{Y}$ la matrice des réponses (de dimension $m \times h$, h le nombre de réponses différentes), $\mathbf{X}$ la matrice contenant les données de calibration (dans notre cas m spectres à n variables), $\mathbf{B}$ la matrice contenant les coefficients de régression du modèle PLS et $\mathbf{E}$ les résidus. L'objectif pour retrouver *in fine* ces coefficients avec l'algorithme SIMPLS est d'extraire successivement les facteurs orthogonaux de $\mathbf{X}$, équation (2.8), déterminés de manière à maximiser leur covariance avec les facteurs correspondant de $\mathbf{Y}$, équation (2.9).

$$\mathbf{t_a} = \mathbf{X_0}\,\mathbf{r_a} \tag{2.8}$$

$$\mathbf{u_a} = \mathbf{Y_0}\,\mathbf{q_a} \tag{2.9}$$

$\mathbf{t_a}$ représente les *scores* sur la matrice des données centrées $\mathbf{X_0}$ pour la composante a ($a = 1, 2, \cdots, A$), $\mathbf{r_a}$ les poids correspondants. De façon similaire, $\mathbf{u_a}$ traduit les *scores* sur la matrice des réponses centrées $\mathbf{Y_0}$ pour la composante a, $\mathbf{q_a}$ contenant alors les poids correspondants. La méthode de régression PLS à travers SIMPLS tend finalement à maximiser la covariance entre les *scores* $\mathbf{t_a}$ et $\mathbf{u_a}$ sous certaines contraintes [77] :

(1) Maximisation de la covariance : $\mathbf{u_a}^{\mathrm{t}}\,\mathbf{t_a} = \mathbf{q_a}^{\mathrm{t}}\left(\mathbf{Y_0}^{\mathrm{t}}\,\mathbf{X_0}\right)\mathbf{r_a} = \max!$

(2) Normalisation des poids $\mathbf{r_a}$: $\mathbf{r_a}^{\mathrm{t}}\,\mathbf{r_a} = 1$

(3) Normalisation des poids $\mathbf{q_a}$: $\mathbf{q_a}^{\mathrm{t}}\,\mathbf{q_a} = 1$

(4) Orthogonalité des *scores* $\mathbf{t}$: $\mathbf{t_b}^{\mathrm{t}}\,\mathbf{t_a} = 0$ $(a > b)$

Le concept de calcul de l'algorithme SIMPLS pour réaliser la régression PLS met en jeu le produit croisé $\mathbf{S}$ des deux matrices tel que :

$$\mathbf{S} = \mathbf{X_0}^{\mathrm{t}}\,\mathbf{Y_0} \tag{2.10}$$

avec $\mathbf{X_0}$ la matrice des données centrées et $\mathbf{Y_0}$ les réponses centrées également. Ensuite, selon la variable latente calculée, une décomposition en valeur singulière (SVD pour *Singular Value Decomposition*) est effectuée sur $\mathbf{S}$ pour la première variable latente et sur $\mathbf{S} - \mathbf{P}(\mathbf{P}^{\mathrm{t}}\,\mathbf{P})^{-1}\,\mathbf{P}^{\mathrm{t}}\,\mathbf{S}$ pour les suivantes, $\mathbf{P}$ étant la matrice contenant les *loadings* (équation (2.12) ci-dessous). Ainsi, pour toute variable latente a, le vecteur des poids $\mathbf{r_a}$ est obtenu à partir du premier vecteur singulier obtenu depuis la SVD. Les *scores* peuvent donc ensuite être calculés suivant l'équation (2.8). Enfin, depuis les *scores*, le vecteur des *loadings* $\mathbf{p_a}$ est déterminé suivant l'équation (2.11).

$$\mathbf{p_a} = \frac{\mathbf{X_0}^{\mathrm{t}}\,\mathbf{t_a}}{(\mathbf{t_a}^{\mathrm{t}}\,\mathbf{t_a})} \tag{2.11}$$

La matrice des poids $\mathbf{R}$, la matrice des *scores* $\mathbf{T}$ et la matrice des *loadings* $\mathbf{P}$ sont finalement obtenues étape par étape à chaque variable latente en concaténant respectivement les vecteurs $\mathbf{r_a}$, $\mathbf{t_a}$ et $\mathbf{p_a}$. L'avantage de l'algorithme SIMPLS sur l'algorithme NIPALS intervient notamment lors de l'étape de déflation. En effet, l'algorithme NIPALS réalise ces étapes sur les matrices de résidus successives de $\mathbf{X_0}$ et $\mathbf{Y_0}$. Dans le cas de SIMPLS, les étapes de déflation sont opérées sur la matrice $\mathbf{S}$, ce qui permet d'extraire uniquement la paire de vecteurs singuliers d'intérêt, à savoir celle qui correspond à la valeur singulière la plus élevée (égale au maximum de covariance). Dans l'ACP, la deuxième composante était calculée à partir de la matrice $\mathbf{E}$ des résidus, déflation de la matrice de données $\mathbf{X}$ initiale. Ici, la matrice obtenue pour la composante $a+1$ est calculée suivant l'équation (2.12) :

$$\mathbf{S_{a+1}} = \mathbf{S_a} - \mathbf{P_a}\left(\mathbf{P_a}^{\mathrm{t}}\,\mathbf{P_a}\right)^{-1}\,\mathbf{P_a}^{\mathrm{t}}\,\mathbf{S_a} \tag{2.12}$$

Finalement, les coefficients de régression $\mathbf{B_{PLS}}$ du modèle PLS sont calculés à partir des matrices déterminées précédemment :

$$\begin{gathered}\mathbf{B_{PLS}} = \mathbf{R}(\mathbf{R}^{\mathrm{t}}\,\mathbf{S}) = \mathbf{R}(\mathbf{T}^{\mathrm{t}}\,\mathbf{Y}) = \mathbf{R}\,\mathbf{Q}^{\mathrm{t}} \\ \mathbf{Q} \equiv \mathbf{Y}^{\mathrm{t}}\,\mathbf{T}\end{gathered} \tag{2.13}$$

De ce fait, pour tout nouveau spectre $\mathbf{x}$, les réponses pourront être prédites suivant $\widehat{\mathbf{Y}} = \mathbf{x}\,\mathbf{B_{PLS}}$. Dans la pratique, il faut toutefois noter que l'utilisation de la méthode PLS pour plusieurs réponses (h réponses supérieur à 1), approche PLS2, procure généralement des

modèles dont les capacités prédictives sont moins performantes que la construction de plusieurs modèles PLS1 (un modèle PLS par réponse, modèles où h est 1). En effet, puisque toutes les réponses sont prises en compte dans un modèle PLS2, il est nécessaire d'optimiser les modèles de chacune des réponses de façon équivalente, contrairement à un modèle PLS1 concentré sur une seule réponse. De manière générale, la PLS1 est privilégiée, l'approche PLS2 étant plutôt utilisée pour la régression de réponses fortement corrélées.

2.2.2 Évaluation d'un modèle PLS

Les capacités prédictives d'un modèle PLS sont généralement évaluées à l'aide des erreurs de prédiction du modèle d'après le critère RMSE (*Root Mean Square Error*), mettant en relation les prédictions du modèle $\hat{y}_i$ par rapport aux réponses mesurées y_i, équation (2.14).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \qquad (2.14)$$

Considérant alors un ensemble d'étalonnage permettant de construire le modèle et un ensemble test pour valider celui-ci, nous pouvons distinguer plusieurs types d'erreurs.

La première est l'erreur RMSEC (*Root Mean Square Error of Calibration*) : le modèle est appliqué sur la totalité des échantillons de calibration. Ces prédictions seront ensuite comparées aux mesures correspondantes à l'aide de l'équation (2.14). Cette erreur est toutefois très peu utilisée pour caractériser les modèles puisqu'elle ne représente pas de réelles capacités de prédiction.

La validation croisée (ou *cross-validation*) a été développée afin de pouvoir palier le manque de représentativité de l'erreur RMSEC tout en travaillant seulement sur l'ensemble de calibration. En effet, pour optimiser un modèle, déterminer le nombre de variables latentes à sélectionner ou encore par manque de données, la validation croisée permet d'évaluer un modèle de régression PLS sans pour autant disposer de jeu test. Il s'agit d'extraire une partie (ou segment) de l'ensemble d'apprentissage et construire un sous-modèle basé sur le reste. Le segment est alors prédit par le sous-modèle établi. Cette opération est répétée plusieurs fois en faisant varier les éléments du segment. L'erreur RMSECV (*Root Mean Square Error of Cross-Validation*) est finalement obtenue en moyennant les erreurs déterminées pour chaque segment. La nature de la validation croisée employée dépend notamment du nombre d'éléments contenus dans le segment ainsi que leur mode de sélection. Ainsi, si les segments sont constitués d'un seul échantillon, la

validation croisée est appelée « *Leave One Out* ». En sélectionnant de façon régulière k éléments à inclure dans le segment, nous parlerons alors de « k*-fold cross-validation* ». Ainsi, selon les données, l'une ou l'autre des validations croisées est choisie préférentiellement.

La validation croisée s'appuie uniquement sur l'ensemble d'étalonnage. L'utilisation d'un jeu de données n'ayant pas été employé pour la construction du modèle de régression, jeu test, est donc nécessaire pour l'évaluation des performances de ce modèle. Ainsi, en appliquant le modèle sur les échantillons de l'ensemble test, l'erreur RMSEP (*Root Mean Square Error of Prediction*) est déterminée. Cette dernière erreur est nécessaire dans le sens où elle permet d'estimer une incertitude empirique de prédiction du modèle PLS à plus ou moins deux fois l'erreur RMSEP [78].

Nous pouvons également noter que dans certains cas, ce n'est pas l'erreur RMSE qui est utilisée pour représenter les erreurs des modèles mais plutôt l'erreur SE (pour *Standard Error*), calculée suivant l'équation (2.15). De la même façon que pour l'erreur RMSE, l'erreur SE se décline en SEC (*Standard Error of Calibration*), SECV (*Standard Error of Cross-Validation*) et SEP (*Standard Error of Prediction*).

$$SE = \sqrt{\frac{1}{n}\sum_{i=1}^{m}(y_i - \hat{y}_i)} \tag{2.15}$$

Aujourd'hui, les modèles sont développés afin d'obtenir une erreur de prédiction la plus faible possible. Par conséquent, de nos jours, en plus d'améliorer les techniques de calcul des modèles de régression PLS, plusieurs travaux mettent en avant les bénéfices que peuvent apporter des traitements spectraux et sélections de variables sur les capacités prédictives des modèles développés [79-81].

2.3 Techniques de corrections spectrales

2.3.1 Corrections de ligne de base

Les techniques de correction de ligne de base sont utilisées lorsque les spectres acquis présentent une déviation de cette base, traduisant la présence d'un fond présent sur les spectres qui peut numériquement être supprimé (Figure 2.5). Ainsi, deux techniques existent principalement : la technique « *detrend* » (signifiant littéralement « enlevé la tendance ») et la correction par les moindre carrés pondérés (WLS pour *Weighted Least Squares*).

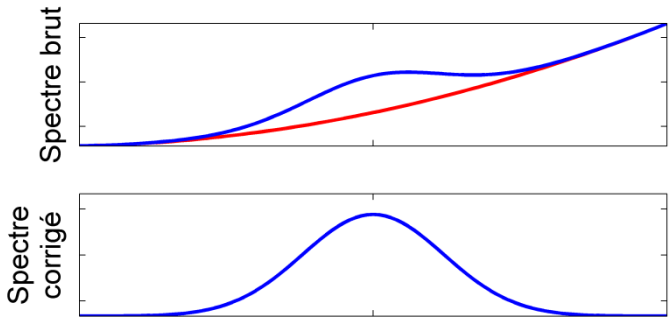

Figure 2.5 Représentation d'une correction de ligne de base. En bleu, le spectre brut (au dessus) puis corrigé (en dessous), en rouge, la ligne de base soustraite.

Les deux techniques fonctionnent sur le même principe. Il s'agit d'ajuster un polynôme de degré k puis le soustraire au spectre. La différence entre les deux techniques provient du système d'ajustement du polynôme. Dans le cas de la correction WLS, un algorithme itératif basé sur les moindres carrés est employé en affectant un poids à chaque point du spectre. Finalement, le spectre prétraité est calculé tel que

$$\mathbf{x}_{\mathbf{i,corrigé}} = \mathbf{x}_{\mathbf{i,brut}} - \mathbf{d} \tag{2.16}$$

avec $\mathbf{x}_{\mathbf{i,corrigé}}$ le spectre corrigé, $\mathbf{x}_{\mathbf{i,brut}}$ le spectre brut et $\mathbf{d}$ le polynôme calculé selon la technique de correction de ligne de base employée. Cependant, il faut noter que la difficulté à déterminer le degré du polynôme de correction est un inconvénient majeur de ces techniques de correction. C'est pourquoi les dérivations sont privilégiées pour la correction de ligne de base.

2.3.2 Dérivées spectrales

De manière générale, les spectres acquis mettent en avant la corrélation entre la nature physico-chimique des procédés observés et les intensités des bandes ou pics mesurés. Les interprétations et les calculs sont alors directement opérés sur les spectres. Néanmoins, l'utilisation des spectres dérivés s'est avérée plus intéressante dans l'analyse des données tant ces spectres améliorent l'interprétation des positions ou des largeurs de bandes. Ainsi, la dérivée première permet de mettre en avant les largeurs de bandes puisque les maxima et minima du spectre dérivé représentent les intensités à mi-hauteur des bandes du spectre brut (Figure 2.6 ci-dessous). Les dérivées secondes quand à elles permettent de mettre l'accent sur les positions des bandes (Figure 2.6).

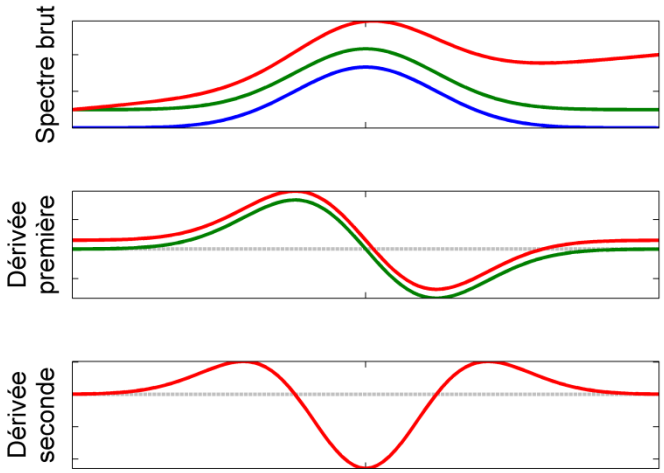

Figure 2.6 Représentation des effets de la dérivation, en particulier sur différents effets. En rouge, un effet multiplicatif sur la ligne de base, en vert un effet additif, en bleu un spectre sans déviation de ligne de base et en pointillés noir, les abscisses nulles.

De plus, les techniques de dérivation sont particulièrement utiles pour éliminer les effets de ligne de base sur les spectres acquis. La Figure 2.6 reprend les différents effets de ligne de base (additifs et multiplicatifs) sur les signaux acquis ainsi que les conséquences de la dérivation.

Il existe aujourd'hui plusieurs moyens de calculer la dérivée d'un spectre. La première méthode calcule simplement la dérivée en un point considérant les valeurs à des distances spécifiques en fréquence (ou *gap*). Ainsi, la dérivée peut s'écrire sous la forme :

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}x_i} = \frac{y_{i+gap/2} - y_{i-gap/2}}{x_{i+gap/2} - x_{i-gap/2}} \tag{2.17}$$

avec y_i l'intensité à la variable i, x_i la fréquence ou longueur d'onde pour cette même variable et enfin gap un écart déterminée. De cette façon il est possible de déterminer la dérivée par simple différence entre deux points. Une application particulière de cette dérivée est la dérivée dite de « simple différence » qui prend un *gap* de 2 points. Ainsi, les dérivées sont calculées sur les points adjacents des valeurs. Par la suite, les dérivées d'ordre deux voire plus sont réalisées sur les dérivées de l'ordre précédent : les dérivées secondes sont issues de dérivations des spectres de dérivée première, et ainsi de suite.

À l'origine, l'intégration d'un *gap* dans la formule permettait de réduire le bruit qui était observé sur les spectres dérivés par rapport aux résultats prenant en compte seulement les points adjacents [82]. Cependant, l'étroitesse des bandes spectrales, notamment en spectroscopie Raman, empêche l'application de distances de calcul élevées et donc la réduction du bruit. C'est pourquoi la méthode de dérivation Savitzky-Golay a été proposée

[83]. Cet algorithme calcule la dérivée au point i en commençant par ajuster un polynôme de degré k sur une fenêtre spectrale de f points ($f \geq k+1$) centrée autour de la variable i. L'équation du polynôme déterminée, celui-ci est alors dérivé selon l'ordre indiqué par l'utilisateur. La fenêtre f balaye ensuite la totalité du spectre afin de générer le spectre dérivé.

L'avantage de cette technique réside dans le lissage opéré par l'algorithme de Savitzky-Golay sur le spectre. Ceci permet donc d'atténuer l'augmentation du niveau de bruit engendré par la dérivation du spectre. Nous pouvons également noter que l'algorithme Savitzky-Golay peut également être employé comme technique de lissage uniquement en fixant l'ordre de dérivée à zéro. L'étape de dérivation n'entre alors pas en compte après avoir ajusté le polynôme sur la fenêtre spectrale.

De manière générale, les dérivées sont également employées pour la correction de déplacements de ligne de base ou de déplacements multiplicatifs du fond spectral. D'autres techniques plus spécifiques permettent également de corriger spécifiquement ces différents déplacements.

2.3.3 Techniques de normalisation

Les techniques de normalisation sont utilisées afin de minimiser la variabilité qui existe au sein d'un lot de spectres. Plusieurs méthodes ont été mises au point pour parvenir à normaliser les spectres. Une des techniques les plus simples consiste à considérer l'aire sous les spectres. En divisant chaque spectre par la somme des valeurs absolues des intensités de chaque variable, les aires sous chacun des spectres sont ramenées à 1. Cette méthode a rapidement été remplacée par deux autres techniques, très utilisées aujourd'hui pour normaliser les signaux : les normalisations *Multiplicative Scatter Correction* (MSC) et *Standard Normal Variate* (SNV).

La technique de normalisation MSC corrige chaque spectre de l'ensemble de données par un spectre de référence, généralement le spectre moyen du lot. Un modèle linéaire est ajusté entre le spectre à corriger $\mathbf{x_i}$ et le spectre de référence $\mathbf{x_{ref}}$ tel que

$$\mathbf{x_i} = \mathbf{a_i} + \mathbf{b_i}\,\mathbf{x_{ref}} + \mathbf{e_i} \tag{2.18}$$

où $\mathbf{a_i}$ et $\mathbf{b_i}$ sont les vecteurs de coefficients de régression calculés à partir du spectre $\mathbf{x_i}$ à corriger et du spectre de référence $\mathbf{x_{ref}}$. Ainsi, les spectres sont normalisés suivant l'équation (2.19) en utilisant les coefficients calculés équation (2.18), considérant $\mathbf{x_{i,MSC}}$ comme le spectre normalisé par la méthode MSC.

$$\mathbf{x_{i,MSC}} = \frac{\mathbf{x_i} - \mathbf{a_i}}{\mathbf{b_i}} \tag{2.19}$$

La méthode de normalisation SNV s'appuie non pas sur des coefficients déterminés à partir d'un spectre de référence mais spécifiquement sur la moyenne ainsi que l'écart-type calculés pour l'ensemble des éléments du spectre à normaliser. Ainsi, pour chaque spectre à corriger $\mathbf{x_i}$, la normalisation SNV applique l'équation (2.20), telle que

$$\mathbf{x_{i,SNV}} = \frac{\mathbf{x_i} - \mu}{\sigma} \tag{2.20}$$

en prenant μ la moyenne de l'ensemble des éléments du spectre à normaliser, σ l'écart-type à la moyenne des mêmes éléments et $\mathbf{x_{i,SNV}}$ le spectre normalisé par la méthode SNV.

2.3.4 Méthodes de mise à l'échelle et de centrage

Les techniques de mise à l'échelle (*scaling*) et de centrage (*centering*) s'appliquent, non pas à un seul spectre, mais à l'entièreté du jeu de données à traiter. Les prétraitements dits de « *scaling* » effectuent une division de chaque variable spectrale par une valeur donnée. Un *scaling* très utilisé consiste à diviser chaque variable par son écart-type, calculé à partir des valeurs des éléments de chaque spectre à la variable donnée. Ainsi, chaque élément spectral dispose du même impact sur la variabilité générale, traduit par un écart-type unique égal à 1. Cette technique n'est que très rarement utilisée en spectroscopie puisqu'elle entraîne la surévaluation des régions spectrales sans signaux, zones de bruit n'ayant pas autant d'impact que les bandes spectrales lors d'analyses multivariées notamment.

Les méthodes de centrage ou *centering* consistent à centrer les données autour d'un spectre de référence déterminé. De manière générale, le centrage par la moyenne (*Mean Center*) est la technique la plus employée. Le spectre moyen de l'ensemble est calculé en prenant les moyennes individuelles des éléments de chaque variable spectrale. Le spectre ainsi obtenu est ensuite soustrait à chaque spectre de l'ensemble de données. Ce prétraitement est réalisé de manière systématique pour les calculs d'ACP ou de PLS car cela permet de placer le nouvel espace déterminé par les composantes principales ou les variables latentes au centre de l'espace des variables initial (composé de chaque variable spectrale).

Un prétraitement regroupant les deux types de prétraitement est le prétraitement « *autoscale* » qui réalise les *scaling* et *centering* que nous avons pris en exemple (*scaling* par l'écart-type et *mean centering*). Ceci étant, la technique dite *autoscale* est très peu utilisée en spectroscopie vibrationnelle pour les mêmes raisons que les *scaling*. L'*autoscale* est plus répandue sur les analyses multivariées mettant en jeu différents paramètres ayant

des variabilités et des moyennes sensiblement différentes. En appliquant cette correction, toutes les variables sont mises au même niveau.

3 État de l'art des suivis métaboliques de cultures cellulaires

Depuis plusieurs années déjà, la spectroscopie et la chimiométrie sont employées afin d'améliorer les connaissances et les techniques de suivi des bioprocédés. L'initiative PAT de la FDA, lancée dans les années 2000, n'a fait qu'accroitre l'utilisation de ces techniques et l'application des principes d'analyse multivariée pour le suivi de différents paramètres.

3.1 Suivi spectroscopique des cultures

De nos jours, les techniques les plus employées pour le suivi spectroscopique de plusieurs paramètres de bioprocédés sont basées sur les spectroscopies NIR et Raman. Cependant, il faut tout de même noter plusieurs travaux mettant en jeu d'autres spectroscopies.

3.1.1 Spectroscopie proche infrarouge (NIR)

En 1985, J. C. Alberti [84] présenta ses travaux sur le suivi de métabolites contenus dans un mélange de culture à l'aide de la spectroscopie proche-infrarouge. Ces travaux, portant sur les cellules type *Saccharomyces cerevisiae* (levure de boulanger) mettaient en avant la possibilité de suivre les concentrations des métabolites (glucose, éthanol et glycérol) dans le temps. Les qualités non-destructives de la spectroscopie proche-infrarouge étaient mises en avant, mais la nécessité de réaliser des prélèvements de culture pour l'analyse spectroscopique était un frein à l'utilisation de cette technique. C'est pourquoi en 1990, A. G. Cavinato [85] proposa un système permettant d'acquérir les spectres d'une culture en reliant une fibre optique directement sur la paroi en verre d'un bioréacteur de culture. Ces travaux furent également appliqués à une culture de levure de boulanger pour le suivi de la production d'éthanol des cellules à l'aide d'un modèle de régression multilinéaire. Les propriétés non-invasives de la spectroscopie proche-infrarouge avaient une nouvelle fois été mises en avant pour le suivi de paramètres. L'étude avait ensuite été étendue à plus de paramètres, notamment le suivi de la concentration en glucose et la densité cellulaire [86]. L'application de la spectroscopie NIR fut ensuite étendue à d'autres types de cellules telles que *Lactobacillus casei* (contrôle de la consommation en glucose et production d'acide lactique et de biomasse) [87], *Phaffia rhodozyma* (suivi de la consommation en glucose et production d'acide lactique et de biomasse) [88] ou encore Sf9 (suivi des concentrations en

glucose et glutamine) [89]. Bien que les fibres optiques aient été utilisées précédemment, ces derniers travaux font seulement état de la faisabilité de suivre les différents paramètres métaboliques et donc ne mettent pas en jeu les acquisitions *in situ* ou en ligne au moyen de fibres optiques. En 1999, la spectroscopie NIR est proposée pour suivre les concentrations en protéines et en ARN ainsi que la densité cellulaire sur un flux de débris cellulaires suite à la centrifugation de cellules *Saccharomyces cerevisiae* [90] (il s'agit ici de la phase *downstream* du bioprocédé de culture).

Dans les années 2000 et l'essor de la sonde à immersion pour l'acquisition de spectres *in situ*, de nombreuses études mettent en jeu ces systèmes pour les suivis métaboliques. Ainsi, suite aux travaux précédemment réalisés, les modèles nouvellement développés permettent de suivre plusieurs métabolites simultanément à l'aide de modèles de régression multivariés, généralement calculés à l'aide de la méthode PLS [91-98]. La faisabilité du suivi en ligne des métabolites n'est alors plus à démontrer.

Par la suite, les travaux de recherche autour des suivis métaboliques tendront à améliorer les modèles de régression déjà établis. C'est généralement autour de la cellule CHO, la plus utilisée dans le domaine pharmaceutique pour la production d'anticorps [99], que les investigations porteront pour l'amélioration des outils de suivis spectroscopiques. Ainsi, en 2013, M. Clavaud [100] mit en avant la nécessité de considérer un certain nombre de paramètres influençant la variabilité de l'instrumentation NIR dans la construction des modèles de régression. Il démontra alors la capacité des modèles chimiométriques à supporter certains changements de paramètres (différent type de bioréacteur, culture atypique, etc.). En 2014, B. Kozma [101] présente des travaux allant dans la même direction en induisant délibérément des variations dans les systèmes de culture afin d'imiter différents effets du milieu de culture. Toujours en 2014, M. Milligan [102] propose des modèles de régression « semi-synthétiques » mettant en jeu des spectres acquis *in situ* ainsi que des spectres acquis sur les prélèvements pour mesure de référence afin d'agrandir l'ensemble d'apprentissage.

3.1.2 Spectroscopie moyen infrarouge (MIR)

Bien que la spectroscopie proche-infrarouge ait un puissant potentiel pour le suivi métabolique des cultures cellulaires, d'autres techniques spectroscopiques ont été mises en jeu pour réaliser les suivis de cultures cellulaires. Le potentiel de la spectroscopie moyen-infrarouge (MIR) a ainsi été comparé à la spectroscopie NIR pour la prédiction des concentrations en acide lactique et en glucose des cellules *Lactobacillus casei*, ainsi que la densité cellulaire [103]. La spectroscopie MIR s'est révélée plus performante pour cette

application. Notons qu'il s'agissait uniquement de spectres acquis sur des prélèvements de culture (*off-line*).

Dans le même temps, le développement de sondes à immersion ou de systèmes en ligne pour la mesure de spectres MIR, basées sur les principes des appareils de mesure à réflexion totale atténuée (ATR pour *Attenuated Total Reflectance*), permet dans les années 2000 de réaliser des travaux sur le suivi métabolique en temps réel des concentrations pour différents types de cellule [104-113]. En 2013, M. Sandor [114] confronte les deux techniques de suivi en ligne, NIR et MIR, dans l'étude des concentrations métaboliques et densités cellulaires de cultures CHO. Ces travaux montrent finalement que les erreurs de prédiction obtenues pour les modèles de régression PLS basés sur les concentrations métaboliques (telles que celle du glucose) et les spectres MIR sont trois fois inférieures que pour les spectres NIR. À l'inverse, la spectroscopie NIR est plus performante pour le suivi des densités cellulaires puisque les sondes ATR-MIR présentent une faible profondeur de pénétration dans le milieu de culture. Ces-dernières sont donc incapables d'acquérir des spectres permettant de mesurer directement la densité cellulaire. Les modèles pour les densités cellulaires s'appuient donc sur la corrélation entre la densité et les concentrations en glucose et lactate.

Les deux techniques spectroscopiques, MIR et NIR, permettent donc de mesurer préférentiellement certains paramètres des cultures cellulaires. Cependant, la fragilité du dispositif de mesure en ligne MIR, à savoir les sondes à immersion ayant un dispositif ATR, empêche l'utilisation des systèmes moyen-infrarouge pour le contrôle des systèmes de production, ce qui contraste avec la robustesse des sondes à immersion NIR [114].

Bien que les résultats précédents prouvent que la spectroscopie infrarouge est en mesure de réaliser un suivi métabolique des cultures de cellules, celle-ci présente un inconvénient principal lié à l'eau. En effet, les bandes caractéristiques de la molécule d'eau sont très intenses en IR, ce qui peut poser problème vis-à-vis du milieu de culture cellulaire aqueux. De ce fait, d'autres types de spectroscopie furent engagés telle que la spectroscopie de fluorescence. Un dispositif de mesure de spectres de fluorescence fut développé en 1998 [115,116] et largement utilisé pour différents suivis de culture [117-123]. Néanmoins, c'est la spectroscopie Raman qui présenta les investigations les plus poussées, proposant des sondes à immersion performantes et permettant facilement de s'affranchir des bandes dues aux molécules d'eau.

3.2 Évolution du suivi par spectroscopie Raman

Dès 1987, T. B. Shope [124] soulignait les avantages de la spectroscopie Raman par rapport à la spectroscopie infrarouge par rapport aux signaux dus à l'eau. Ces travaux mettaient en application les capacités de la spectroscopie Raman, couplée à une cellule de mesure ATR, à suivre des produits tels que l'éthanol, le méthanol et l'acétone dans les fermentations en se basant sur quelques bandes Raman caractéristiques. Dans cette dynamique, C. Gomy propose en 1988 [125,126] une méthode de suivi des concentrations en éthanol, glucose et fructose lors d'une fermentation alcoolique à l'aide de la spectrométrie Raman associée aux fibres optiques : une fibre conduisant le signal excitateur vers l'échantillon tandis que deux fibres disposées à 90° de la première (afin de minimiser la mesure de la source) permettaient de collecter le signal. Bien que la fluorescence se révèle être un obstacle pour les mesures Raman sur du matériel biologique, ces différents travaux permettent alors de montrer que cette spectroscopie, couplée à des outils statistiques, permet de suivre les concentrations de différents métabolites lors de procédés de fermentation.

Par la suite, C. H. Spiegelman [127] développa en 1998 un modèle de régression PLS sur la base de spectres Raman acquis sur des solutions diluées de glucose. Ceci inspira A. D. Shaw en 1999 [128] pour le suivi *at-line* des concentrations en glucose et en éthanol d'une culture de levure. Le principe repose sur un canal de dérivation permettant, à l'aide d'une pompe, de faire circuler une partie de la culture dans un conduit annexe composé de tubes en quartz de 3 mm de diamètre. Ainsi, il était possible de réaliser des spectres Raman en continu en concentrant le faisceau laser excitateur directement sur le contenant des tubes en quartz. Lors de ces travaux, une source laser NIR à 780 nm fut proposée afin de diminuer l'effet fluorescent apporté par la majeure partie du matériel biologique. De plus, ce type de source présente une sensibilité plus importante qu'une source à 1064 nm par exemple car l'intensité Raman est inversement proportionnelle à la longueur d'onde de la source à la puissance 4.

En 2003, C. Cannizzaro [129] publia des travaux sur le suivi en ligne de caroténoïdes dans les cultures de cellules *Phaffia rhodozyma*. À la différence des travaux précédents, les spectres Raman étaient ici acquis *in situ* à l'aide d'une sonde à immersion de 12,5 mm insérée dans le bioréacteur de culture. La sonde était équipée d'un obturateur en téflon qui permettait de bloquer le faisceau lumineux provenant de la source et de calibrer le détecteur CCD. Une source à 785 nm était employée afin de minimiser la fluorescence des échantillons [128] et un filtre disposé avant le détecteur permettait de bloquer la diffusion Rayleigh. Un an plus tard, H. L. T. Lee [130] proposait également d'utiliser une sonde à

immersion pour acquérir des spectres *in situ* de cultures de cellules *Escherichia coli*. Ici, la fibre optique de la partie immergée de la sonde était protégée soit par un tube en laiton (fileté à son extrémité afin de permettre un ajustement axial du point focal), soit par un tube d'aluminium anodisé et gainé par deux couches d'acétate de magnésium, une fenêtre de saphir scellée à l'extrémité. La tête de sonde présentait alors un design similaire à celui des travaux de A. S. Arnold de 1998 [131] et permettait d'émettre un excitation à 785 nm, procurant 50–60 mW en sortie à l'échantillon. Ces derniers travaux permirent de mettre en avant la possibilité de suivre en temps réel plusieurs paramètres métaboliques d'une culture de cellules à l'aide d'une sonde Raman immergée. Les travaux portant sur le suivi de cultures cellulaires à l'aide de la spectrométrie Raman mettraient alors en jeu quasi-systématiquement des sondes à immersion disposées sur les bioréacteurs de culture.

Une nouvelle fois, c'est autour de la cellule CHO que les développements furent ensuite les plus notables. Ainsi, en 2011, N. R. Abu-Absi [132] proposa une méthode de suivi de plusieurs paramètres métaboliques pour les cultures de cellules CHO. L'acquisition des spectres était réalisée à l'aide d'une sonde à immersion en acier inoxydable reliée au spectromètre Raman à l'aide d'une fibre optique. La source excitait à 785 nm pour une puissance résultante avoisinant 200 mW. Les cultures mises en jeu était opérées en mode *fed-batch*, présentant notamment des variations importantes des concentrations en glucose et glutamine. De plus, ces travaux proposaient de travailler sur plusieurs lots de culture (*batches*) afin d'intégrer plus de variabilité dans les modèles de régression PLS. La même année, J. Moretto [133] propose également un suivi de culture de cellules CHO mettant en jeu le même type de sonde. Plusieurs modèles de régression PLS étaient alors proposés pour les concentrations en glucose, glutamine, glutamate, lactate, ammonium ainsi que les densités totales et en cellules vivantes. J. Moretto souligne à son tour le fait que la spectroscopie Raman utilisant ces sondes à immersion est très adaptée au suivi métabolique en ligne de différents procédés cellulaires. De ce fait, les travaux de recherche à venir se concentreraient sur l'amélioration des modèles de régression en développant des méthodes de référence plus performantes pour la mesure des métabolites, mais aussi en agrandissant l'ensemble d'apprentissage pour des modèles plus robustes.

Les premières améliorations des modèles de régression PLS pour les métabolites des cultures CHO sont présentées en 2012 par J. Whelan [134]. Toujours en utilisant des sondes Raman en acier inoxydable et une source à 785 nm (pour une puissance d'environ 350 mW à l'échantillon), celle-ci proposait d'ajouter au sein de l'ensemble d'apprentissage des données acquises à partir de différentes tailles de lot (cultures en bioréacteurs de 3 L et 15 L). Les modèles de régression PLS alors proposés présentaient les erreurs SEP figurant Tableau 2.1.

Tableau 2.1 Erreurs et coefficients obtenus pour les modèles de régression de J. Whelan [134].

Métabolite (unité)	**N. LV**	**R^2**	**SEC**	**SEP**	**Gammes**
Glucose (mM)	5	0,942	1,11	2,09	5,13 – 22,33
Glutamine (mM)	7	0,953	0,12	0,22	1,57 – 4,03
Glutamate (mM)	4	0,921	0,14	0,17	0,34 – 2,03
Ammoniac (mM)	7	0,978	0,16	0,36	1,25 – 4,32
Lactate (mM)	3	0,947	10,56	11,49	1,6 – 155,8
TCD (10^6 cell/mL)	3	0,937	0,44	0,71	0,27 – 5,93
VCD (10^6 cell/mL)	9	0,997	0,81	0,90	0,26 – 4,96

La robustesse étant toujours le point de mire dans le développement des modèles de régression, B. Berry [135] proposa en 2014 un modèle prenant en compte des spectres acquis sur des cultures de cellules CHO dans différentes tailles de bioréacteurs. Ainsi, 20 cultures furent mises en jeu pour cette étude : 13 cultures sur bioréacteur de 5 L (petite échelle), 4 cultures sur bioréacteurs de 200 L (échelle pilote) et 3 cultures sur bioréacteurs de 2000 L (échelle production). Au cours de ces travaux, plusieurs modèles ont été développés afin d'intégrer progressivement l'augmentation du volume de culture. En reprenant les résultats obtenus par leurs travaux, Figure 2.7, nous pouvons voir l'impact de l'intégration des différentes tailles de bioréacteur dans les travaux.

Ceux-ci permettent de montrer que l'intégration de la variabilité due aux différentes tailles de bioréacteurs de culture, en ajoutant dans l'ensemble de calibration des lots de production, permet *in fine* d'améliorer la robustesse des modèles. Ces derniers seront alors plus à même de prédire les paramètres métaboliques dans tout type de volume de culture, permettant ainsi de déplacer plus facilement les modèles sur les lignes de production directement.

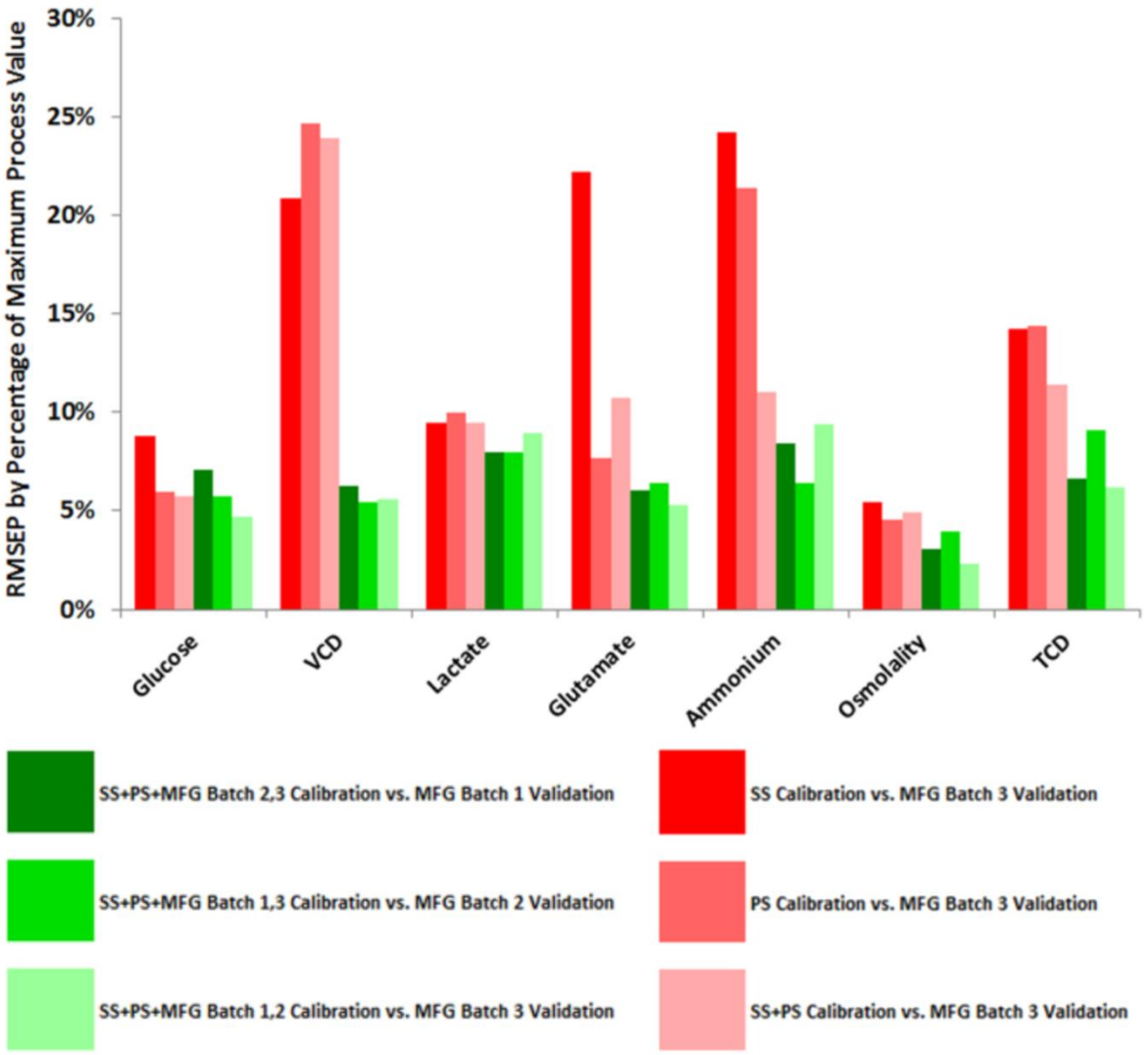

Figure 2.7 Résultats obtenus par B. Berry [135]. RMSEP (en pourcentage par rapport aux niveaux maximaux des gammes des métabolites) obtenus pour chaque métabolite et par combinaison des tailles de bioréacteur. Les lots SS (*Small Scale*) concernent les bioréacteurs de 5 L, les lots PS (*Pilot Scale*) ceux de 200 L et les MFG (*Manufacturing*) 2000 L.

Finalement, en 2015, H. Mehdizadeh [136] propose un modèle de régression PLS dit « générique » permettant de prendre en compte des variations de procédé ainsi que différentes tailles de culture. Ainsi, les modèles de régression développés pourraient être utilisés sur différents procédés de culture et différentes tailles de lot. Toutefois, il convient de noter que ces modèles sont développés pour les paramètres glucose, lactate et VCD, identifiés en tant que paramètres clés des cultures de cellules CHO. L'ensemble d'apprentissage pour le développement des modèles contenait en tout 7 lots de culture en bioréacteurs de 1 L à 3 L et 1 lot acquis sur une culture en bioréacteur de 500 L. Trois cultures différentes ont été utilisées pour valider les modèles de régression : un premier lot « normal » afin de pouvoir évaluer simplement le modèle (RMSEP1), un deuxième lot acquis dans un bioréacteur de 500 L afin de pouvoir évaluer la capacité du modèle à s'affranchir de la taille du bioréacteur de culture (semblable à [135] ; RMSEP2) et un troisième lot acquis

sur une lignée cellulaire CHO différente de celles employées en calibration pour évaluer le caractère « générique » du modèle (RMSEP3). Les résultats obtenus par M. Mehdizadeh sont récapitulés dans le Tableau 2.2.

Tableau 2.2 Erreurs et coefficients obtenus pour les modèles de régression de H. Mehdizabeh [136].

Métabolite (unité)	N. LV	R^2	RMSEP1	RMSEP2	RMSEP3	Gammes
Glucose (g/L)	6	0,938	0,43	0,28	0,44	0,00 – 8,45
Lactate (g/L)	5	0,954	0,26	0,072	0,41	0,00 – 3,95
VCD (10^5 cell/mL)	9	0,936	19,82	30,87	13,95	0 – 300

Les travaux de recherche montrent finalement qu'en intégrant la variabilité de différentes lignées d'un même type cellulaire et en ajoutant des données concernant des changements de volume de culture, il serait alors possible de déplacer les modèles de régression sur des sites de production dits BPF (pour Bonnes Pratiques de Fabrication, ou GMP en anglais, *Good Manufacturing Practice*) dans les entreprises pharmaceutiques. De ce fait, en disposant une sonde Raman sur un bioréacteur de culture de cellules CHO en production, il serait possible, à l'aide des derniers modèles de prédiction, de déterminer en temps réel les concentrations des différents paramètres métaboliques avec des incertitudes de mesure autour de 10 % des concentrations maximales des procédés.

Ainsi, ce type de procédé permettrait donc d'implémenter la philosophie QbD de la FDA au sein des unités de culture cellulaire pour la fabrication de vaccin. Néanmoins, il faut rappeler que la totalité des modèles de prédiction développés sont applicables uniquement à la cellule CHO. Il convient donc de rappeler que celles-ci ne sont pas les seules employées pour la culture de produits d'intérêt pour les vaccins. Il reste encore à montrer la possibilité de suivre les différents métabolites en temps réel sur les cultures d'autres types de cellules telles que Sf9, HeLa ou HEK.

De plus, avant d'être parfaitement capable de suivre les paramètres métaboliques des cultures de cellules en production, il reste plusieurs points à soulever. L'accès à des modèles de régression robustes passe par l'intégration de différentes variabilités. Mais les différents travaux prennent surtout en compte les différences entre les lignées de cellules CHO et les différences de volumes de culture. Cependant, en restant à petite échelle, il est important de noter l'existence d'une multitude de paramètres physiques à prendre en compte. Ainsi, des variations de température, de pH ou encore de vitesse d'agitation entre autres peuvent impacter la culture et donc modifier les signaux Raman acquis *in situ*. Il convient donc d'étudier ces paramètres avant d'exporter les modèles de prédiction sur les lignes de production.

Chapitre 3

Mise en place et paramétrage du suivi Raman pour différentes cultures cellulaires

Au cours de ce chapitre, nous détaillerons comment appliquer les principes de la spectroscopie Raman et de la chimiométrie au suivi métabolique de cultures de cellules et les paramètres inhérents à ce type de travail. Nous aurons également pour objectif d'optimiser les paramètres de l'acquisition Raman et du traitement des signaux mesurés.

1 Description du suivi spectroscopique d'une culture

Le montage utilisé au cours de ces travaux de recherche pour effectuer le suivi spectroscopique d'une culture cellulaire est représenté Figure 3.1.

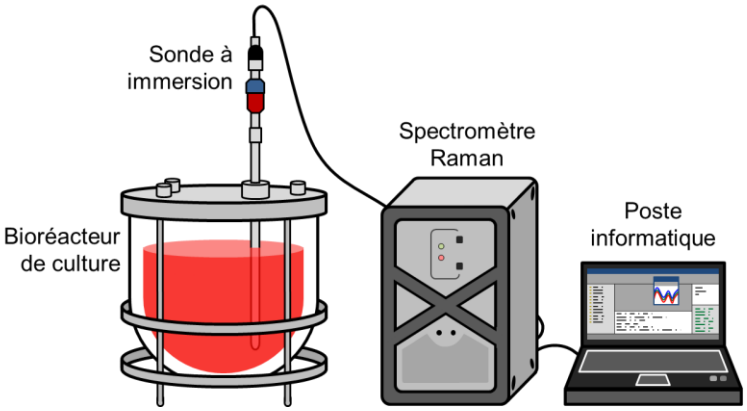

Figure 3.1 Représentation du montage pour le suivi Raman d'une culture cellulaire. La sonde à immersion reliée au spectromètre Raman est fixée à la platine du bioréacteur et plongée au sein du milieu de culture. Le poste informatique permet de régler les paramètres spécifiques à l'acquisition des spectres Raman.

1.1 Cultures de cellules : stratégies de *feed* et prélèvements

Les cultures étudiées ici sont généralement réalisées dans des bioréacteurs de 7 L (4 L de volume de travail). Selon le type de cellule employé, le procédé de culture évolue de différentes manières.

Pour les cellules CHO, un procédé *fed-batch* est mis en place afin de conserver un certain seuil de concentration en glucose. En effet, ces cellules consomment principalement le glucose du milieu pour entrainer la croissance cellulaire. Il convient donc de surveiller l'abondance du glucose dans les cultures de cellules CHO. Le procédé *fed-batch* mis en

place prévoit donc l'ajout de *feeds* dans le milieu tout au long de la culture cellulaire. Pour rappel, les tendances des concentrations des différents métabolites sont représentées Figure 1.5, page 31. La stratégie de *feed* des cultures CHO mises en jeu comprend deux types d'ajout. Le premier, est réalisé à partir du quatrième jour de culture (à hauteur de 1,6 % du volume du bioréacteur), date à partir de laquelle les cellules consomment de manière plus importante les ressources du milieu de culture, puis au septième jour (à hauteur de 2,4 % du volume du bioréacteur) afin de garantir un apport suffisant en acides aminés et en sels minéraux. Le deuxième type de *feed* employé a été mis en place afin de garantir une abondance minimale en glucose pour la culture. En effet, à partir du quatrième jour, lorsque la concentration atteint 3 g/L (16,65 mM), un ajout à hauteur de 6 g/L (33,40 mM) est réalisé. De cette façon, la concentration en glucose est continuellement maintenue au dessus d'un certain seuil afin qu'il n'y ait pas d'appauvrissement du milieu. Auquel cas, cela entrainerait une forte mortalité cellulaire et stopperait la production de particules d'intérêt, ici les anticorps recombinants. Afin d'avoir un minimum de connaissances d'un point de vue métabolique sur la culture, des prélèvements de 20–30 mL sont donc réalisés durant la culture. Cet échantillonnage a été renforcé pour ces travaux de recherche afin de disposer de suffisamment de mesures des paramètres métaboliques pour les développements futurs de modèles chimiométriques, passant de un à deux prélèvements par jour à trois ou quatre.

Les cultures de cellules HeLa employées au cours de ces travaux présentent des procédés de nutrition différents. En effet, les cellules étant moins consommatrices en glucose, les procédés ne nécessitent pas de *feeds* quasi-continus à l'instar des cultures CHO. Un seul *feed* est réalisé au bout du troisième jour de culture, essentiellement composé de glucose et de glutamine. En effet, les cellules HeLa sont plus consommatrices que productrices de glutamine (Figure 1.6, page 32) et nécessitent donc d'être réapprovisionnées au cours de la culture. Dans notre cas, les cellules HeLa sont employées en tant que supports pour la culture de virus. Il convient donc de stopper la croissance cellulaire le plus rapidement possible une fois que le maximum de cellules vivantes dans le milieu est atteint, sans entamer la phase de sénescence traduisant la mort biologique des cellules hôtes pour la culture virale. L'inoculation virale peut alors être réalisée. Pour cela, la culture cellulaire stoppée précédemment est diluée au quart dans du milieu de culture frais ajouté à du milieu infectieux. Les paramètres physiques de contrôle sont alors modifiés afin de favoriser la prolifération virale au sein de la culture. Tout comme les cultures de cellules CHO, le nombre de prélèvements pour les mesures de référence métaboliques des cultures de cellules HeLa a été augmenté pour les deux phases de culture (amplification cellulaire et culture virale) afin de pouvoir dégager un nombre d'échantillons suffisant pour les analyses chimiométriques.

Les cellules Sf9 sont également employées pour la production de virus dans ces travaux de recherche. Durant une première phase, les cellules sont d'abord cultivées, tout comme les cellules HeLa. Toutefois, la faible consommation des cellules Sf9 permet de se passer de *feed*. Ainsi, les concentrations des différents biomarqueurs sont suffisantes pour soutenir l'amplification cellulaire jusqu'à atteindre le seuil plateau de densité en cellules vivantes. Suite à la phase de multiplication cellulaire, une partie de la culture est infectée (un quart de la culture, tout comme les cellules HeLa) pour la production virale. Encore une fois, le nombre de prélèvements a été augmenté afin de disposer de suffisamment de mesures de référence pour les études chimiométriques.

1.2 Acquisitions spectrales et développements chimiométriques

Le matériel Raman mis en jeu au cours des travaux de recherche présentés ici est un spectromètre Kaiser Rxn2™ (Kaiser Optical Systems, Inc.) pouvant disposer de quatre sondes à immersion chacune reliée grâce à une fibre optique. Ces dernières sont plongées dans les bioréacteurs de culture pour l'acquisition des spectres *in situ*. Le bioréacteur est enveloppé sous une couche de papier aluminium afin de protéger le contenu du bioréacteur de toute perturbation extérieure (telle que la lumière provenant de l'éclairage du laboratoire), la mesure de la diffusion Raman s'effectuant dans le visible.

La source excitatrice du spectromètre employée ici est une diode laser émettant à 785 nm pour une puissance d'environ 400 mW délivrée à l'échantillon. La diffusion Raman n'est pas maximale à cette longueur d'onde, mais elle permet toutefois de limiter le phénomène de fluorescence. Muni d'un détecteur CCD, le spectromètre Raman permet d'acquérir les spectres sur une plage spectrale comprise entre 3425 cm^{-1} et 100 cm^{-1}, présentant une résolution de 1 cm^{-1}. Le logiciel d'acquisition iC Raman™ 4.1 (Kaiser Optical Systems, Inc.) est employé pour l'acquisition des spectres Raman. Celui-ci corrige automatiquement les raies cosmiques présentes sur les spectres, bandes très fines typiques d'appareils présentant un détecteur CCD, et effectue automatiquement la soustraction du spectre d'obscurité ou *dark spectrum*. Les temps d'acquisition représentant un intérêt particulier pour nos travaux, ceux-ci seront plus largement détaillés dans la partie 2 de ce chapitre (Détermination du temps d'acquisition Raman). Les spectres Raman sont acquis en continu tout au long de la culture, tels que représentés Figure 3.2.

Le logiciel iC Raman™ permet d'exporter les spectres au format SPC, extension développée par Galactic Industries, propriété de Thermo Fisher Scientific. Ce format permet de sauver un ou plusieurs spectres sous forme de bases de données comprenant non seulement les valeurs des intensités et déplacements Raman, mais aussi les informations d'utilisation telles que les temps, les dates ou les heures d'acquisition. Cette extension

permet notamment d'importer les spectres dans l'environnement MATLAB® (The MathWorks) à l'aide de la boite à outil nommée PLS_Toolbox (Eigenvector Research Inc.). En prenant en compte les dates et heures des prélèvements du bioréacteur pour les mesures de référence, il est possible de relier chacune des mesures à un spectre Raman. Ainsi, pour une culture cellulaire, nous disposons d'une part de m spectres Raman associés aux m mesures de référence et d'une série de spectres sans référence acquis en continu pendant toute la durée du bioprocédé.

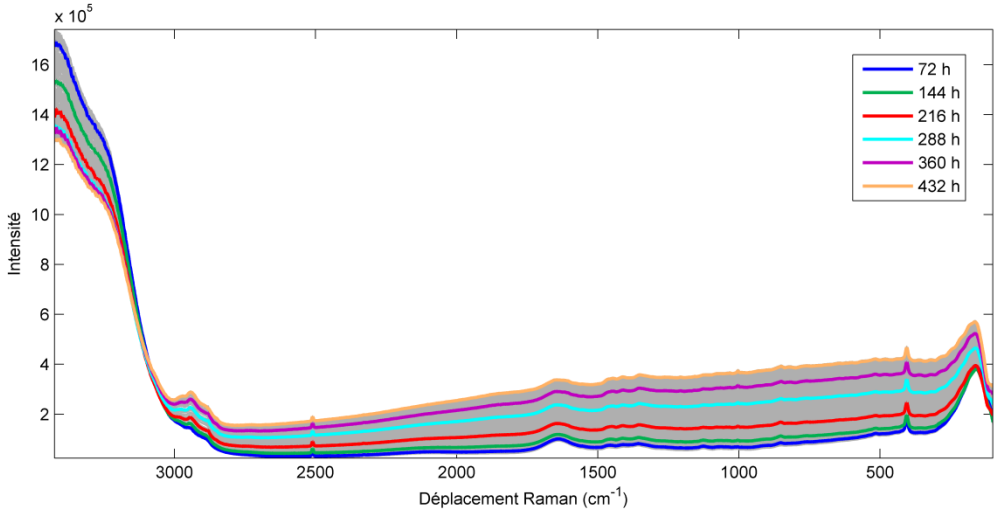

Figure 3.2 Spectres Raman acquis en continu au cours d'une culture de cellules CHO de près de 430 h. Les spectres sont acquis sur 5 min (10 × 30 s). En gris, les spectres acquis en continu.

Les prétraitements spectraux et calculs chimiométriques des travaux présentés sont effectués sous MATLAB®, version 7.7, appuyé de la PLS_Toolbox 7.0.3. Afin de réaliser les régressions PLS, plusieurs matrices sont construites. La première contient les informations sur les prélèvements de référence, matrice des réponses $\mathbf{y}$. Par exemple, pour le développement d'un modèle de régression PLS du glucose, cette matrice est un vecteur contenant les m concentrations obtenues par la méthode de référence sur les prélèvements du bioréacteur de culture. Ainsi ce vecteur sera associé à la matrice contenant les m spectres associés, matrice des données d'entrée $\mathbf{X}$, pour l'élaboration du modèle PLS suivant l'équation (2.13), page 65. Le modèle PLS peut alors être appliqué sur tous les spectres acquis en continu pendant la culture, ce qui permet, toujours dans le cas du glucose, de remonter à la concentration du biomarqueur tout au long du procédé.

D'un point de vue théorique, le développement d'un modèle de régression PLS pour le suivi s'appuie uniquement sur les spectres et les réponses associées. Cependant, dans la pratique, ce modèle est dépendant d'un très grand nombre de paramètres. D'une part, il est important de noter que l'erreur de prédiction du modèle PLS développé est irrémédiablement liée à celle de la méthode de référence employée pour la détermination des paramètres

métaboliques. En effet, puisque le modèle de régression s'appuie sur les concentrations apportées par les systèmes de référence, l'erreur des modèles sera dépendante de celle des appareils hors-ligne. D'autre part, les données d'entrée, à savoir les spectres, influent beaucoup sur la qualité d'un modèle de régression PLS. C'est pourquoi il est important de bénéficier de signaux Raman de qualité pour le développement des modèles, qu'ils soient traités ou non. Enfin, il est important de noter que la signification des modèles chimiométriques est irrémédiablement dépendante des informations d'entrée. C'est d'ailleurs le point d'orgue des derniers travaux [134-136] qui mettent en avant l'importance de l'ensemble d'apprentissage du modèle de régression ainsi que celle du lot de prédiction pour justifier de sa robustesse face aux différentes variations du bioprocédé (variations entre les cultures, différences de volume ou de lignée d'un même type cellulaire).

2 Détermination du temps d'acquisition Raman optimal

Comme il a été souligné dans la section 1.2.2 du Chapitre 2, le temps d'acquisition des spectres Raman joue un rôle majeur dans la qualité des spectres obtenus et donc sur les performances des modèles de régression PLS. Cependant, dans la littérature, la détermination de ce temps n'est pas particulièrement détaillée. De surcroit, pour des supports d'étude similaires, les temps d'acquisition peuvent varier d'un article à l'autre. Ainsi, en 2011, N. R. Abu-Absi [132] proposait un temps d'acquisition de 10 × 75 s, prenant en considération entre 40 % et 80 % de la gamme du convertisseur analogique-numérique du détecteur pour déterminer le temps d'exposition de chaque accumulation. Toujours en 2011, J. Moretto [133] proposait un temps d'acquisition total de 10 min (600 × 1 s). En 2015, B. Berry [135] proposait le même temps d'acquisition décomposé en 600 accumulations d'une seconde. La même année, H. Mehdizadeh [136] présentait un temps d'acquisition de 75 × 10 s pour le développement d'un modèle de régression générique. C'est en 2012 dans l'article de J. Whelan [134] que la détermination du temps d'acquisition est le plus détaillée. Le temps d'exposition est fixé à 10 s afin d'atteindre 50 % de la gamme du convertisseur analogique-numérique. Ensuite, certaines simulations les conduisent à ne pas travailler sur un temps d'acquisition total dépassant 6 min. Ainsi, ils fixèrent leur temps total d'acquisition à 5 min, soit 10 × 30 s. Bien que ces différents travaux portent tous sur les cultures de cellules CHO, mettant en jeu un système Raman Kaiser Rxn2™, les temps d'acquisition proposés sont, à l'évidence, sensiblement différents.

2.1 Méthodologie employée

Le temps d'acquisition est important, non seulement pour obtenir des spectres de qualité, mais aussi afin de conserver toute la signification chimique du contenu du

bioréacteur aux temps des prélèvements et garder le lien qui existe entre un spectre Raman et les données de référence acquises hors-ligne. Si le temps d'acquisition est trop long, les spectres Raman ne retranscriront donc pas la juste information aux temps de prélèvement et cela même si le rapport SNR est amélioré. Si au contraire le temps d'acquisition est trop court, la qualité spectrale risque de ne pas être suffisante pour construire des modèles PLS performants. Nous proposons alors une méthode originale pour la détermination du temps optimal d'acquisition.

2.1.1 Détermination du temps d'exposition

Dans cette méthodologie, il est nécessaire de déterminer dans un premier temps le temps d'exposition de chaque accumulation des spectres. En effet, ce temps est une des deux composantes du temps d'acquisition global et permet notamment d'amplifier l'intensité des signaux Raman mesurés. Ceci a pour effet d'augmenter le rapport SNR, marqueur de la qualité des spectres.

Le logiciel iC Raman™ propose son propre outil de détermination du temps d'exposition avant de lancer toute acquisition spectrale, représenté Figure 3.3. L'utilisateur entre un premier temps d'exposition. L'instrument va alors acquérir et représenter un spectre acquis au temps indiqué. De plus, le logiciel permet de représenter le niveau de saturation des pixels du détecteur CCD équipé sur le spectromètre.

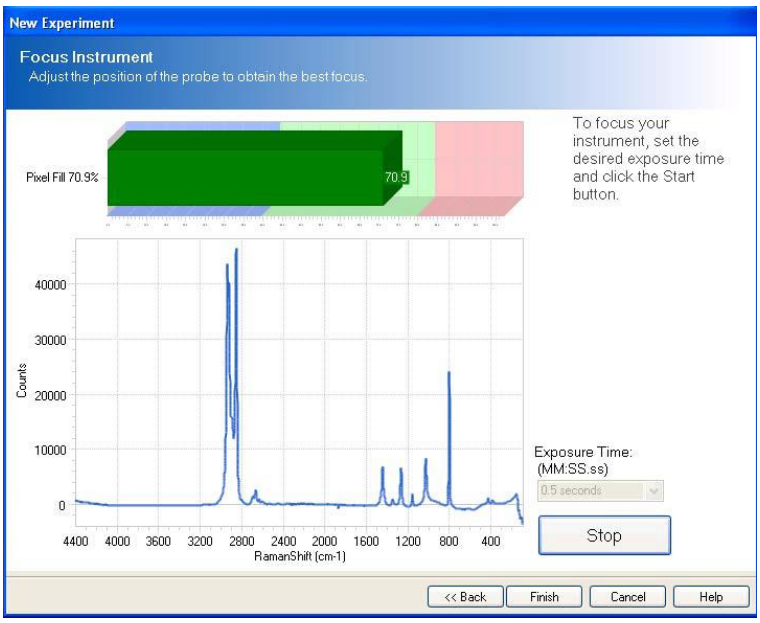

Figure 3.3 Représentation de l'outil de détermination du temps d'exposition proposé par le logiciel iC Raman™. Image issue du guide d'utilisation du logiciel.

Nous pouvons également noter que dans le cas de suivis spectroscopiques de cultures cellulaires, un phénomène de fluorescence apparaît, notamment à cause du matériel

biologique mis en jeu. Bien que minimisé par l'utilisation d'une source excitatrice à 785 nm, il est nécessaire de prendre en compte ce phénomène croissant qui influe inévitablement sur l'intensité spectrale. La Figure 3.2 permet de visualiser l'évolution de ce fond au cours du temps. De ce fait, le temps d'exposition est déterminé de façon à atteindre entre 40 % et 80 % de la gamme du convertisseur analogique-numérique.

À l'aide d'une approche essai-erreur, nous avons donc modifié progressivement le temps d'exposition afin de faire varier le taux de saturation du détecteur. Nous avons finalement conclu qu'un temps d'exposition de 30 s était le plus judicieux pour la suite des travaux. Ce travail a été réalisé pour les trois types de cellule (CHO, HeLa, Sf9) et le temps d'exposition précédemment déterminé a été appliqué à tous les bioprocédés mis en jeu.

2.1.2 Détermination du nombre d'accumulations

Une fois le temps d'exposition établi, il faut déterminer le nombre d'accumulations à effectuer pour obtenir le temps d'acquisition total. En sommant ou moyennant les spectres accumulés, il est alors possible de réduire le bruit, phénomène aléatoire observé sur les spectres finaux et donc améliorer le rapport SNR (d'un ordre $\sqrt{n}$, n étant le nombre d'accumulations). Cependant, nous ne pouvons pas nous permettre de travailler sur un très grand nombre d'accumulations sous peine de générer un spectre moyen de moins en moins représentatif de la mesure de référence du prélèvement réalisé à un instant t. Ainsi, nous avons mis au point un protocole d'optimisation du nombre d'accumulations. Celui-ci est représenté de manière schématique Figure 3.4.

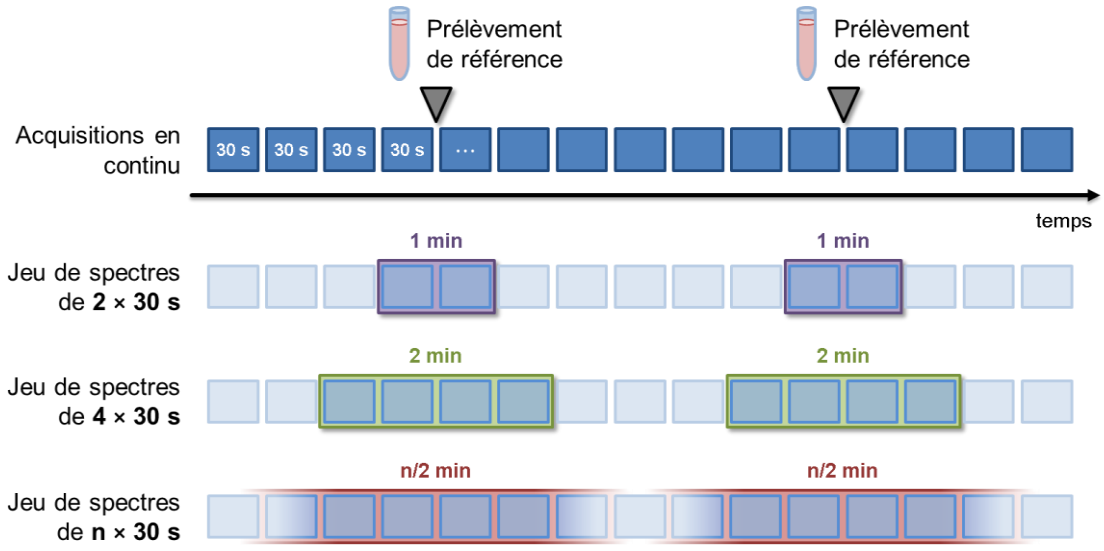

Figure 3.4 Schéma représentatif de la stratégie de construction des jeux de données spectrales pour la détermination du nombre d'accumulations et l'optimisation du temps d'acquisition total.

Sur une seule et même culture cellulaire, nous avons acquis en continu des spectres de 30 s (une seule accumulation, soit 1 × 30 s). Pendant toute la durée de la culture, des prélèvements de référence ont été effectués et notés à des temps précis. Ensuite, en moyennant autour des temps de prélèvement un certain nombre de spectres de 30 s, il est possible de générer artificiellement des spectres Raman à temps d'acquisition désirés. Il est donc possible *in fine* de produire, pour une seule et même culture, plusieurs jeux de spectres à temps d'acquisition différents pour étudier l'influence du nombre d'accumulations.

Étant donné que le rapport SNR n'évolue que d'un rapport $\sqrt{n}$ en fonction du nombre n d'accumulations, nous avons étudié six temps d'acquisition globaux différents: 1 min (n = 2), 2 min (n = 4), 5 min (n = 10), 10 min (n = 20), 16 min (n = 32) et 25 min (n = 50). Ce travail a été réalisé pour trois types de cellule : les cellules CHO, les cellules HeLa et les cellules Sf9.

2.2 Temps d'acquisition pour les cultures de cellules CHO

Pour l'étude du temps d'acquisition total des cellules CHO, les essais portent sur une culture de cellules de plus de 330 h durant laquelle près de 70 prélèvements ont été réalisés. Cependant, suite à certaines contraintes temporelles, seules 63 mesures de référence ont été prises en compte. Les gammes de concentration de chaque biomarqueur sont présentées Tableau 3.1.

Tableau 3.1 Paramètres obtenus pour les modèles PLS les plus performants par biomarqueur dans l'étude du temps d'acquisition d'une culture de cellules CHO.

Biomarqueur (unité)	Temps	R^2	LV	RMSEP	Gamme
Glucose (mM)	5 min	0,983	4	2,03	11,54 – 58,70
Glutamine (mM)	5 min	0,924	3	0,19	0,00 – 1,66
Glutamate (mM)	5 min	0,967	3	0,27	1,85 – 6,82
Lactate (mM)	5 min	0,969	5	0,70	0,00 – 10,06
Ammonium (mM)	2 min	0,970	3	0,74	0,61 – 16,97
VCD (10^6 cell/mL)	10 min	0,964	3	0,52	0,27 – 13,21
TCD (10^6 cell/mL)	10 min	0,969	3	0,51	0,27 – 16,50

Les spectres de 1 × 30 s ont été acquis en continu tout au long de la culture afin d'appliquer le précédent protocole. Six jeux de données à différents temps ont alors été générés. Pour chaque jeu, les spectres ont été prétraités de la même manière : afin d'éliminer l'effet additif de la fluorescence sur la ligne de base des spectres, ces-derniers ont été dérivés. L'algorithme de Savitzky-Golay a été employé en prenant une taille de fenêtre mobile de 15 points (soit ici 15 cm^{-1}), et ajustant un polynôme d'ordre 2 sur chaque portion. L'effet correctif apporté par ce prétraitement est représenté sur la Figure 3.5. Il faut

également noter que les spectres ont été acquis sur 3425–100 cm^{-1}. Toutefois, la totalité du spectre n'est pas utile pour réaliser des modèles de régression performants. En effet, l'empreinte spectrale des composés organiques est manifeste de 1800 cm^{-1} à 365 cm^{-1} [137]. Ainsi, en travaillant uniquement sur ce domaine spectral, il est possible de réduire préférentiellement le domaine d'étude et donc améliorer les performances des modèles de régression. Dans notre cas, nous sélectionnons uniquement la région comprise entre 1775 cm^{-1} et 350 cm^{-1}, telle qu'il est indiqué sur la Figure 3.5.

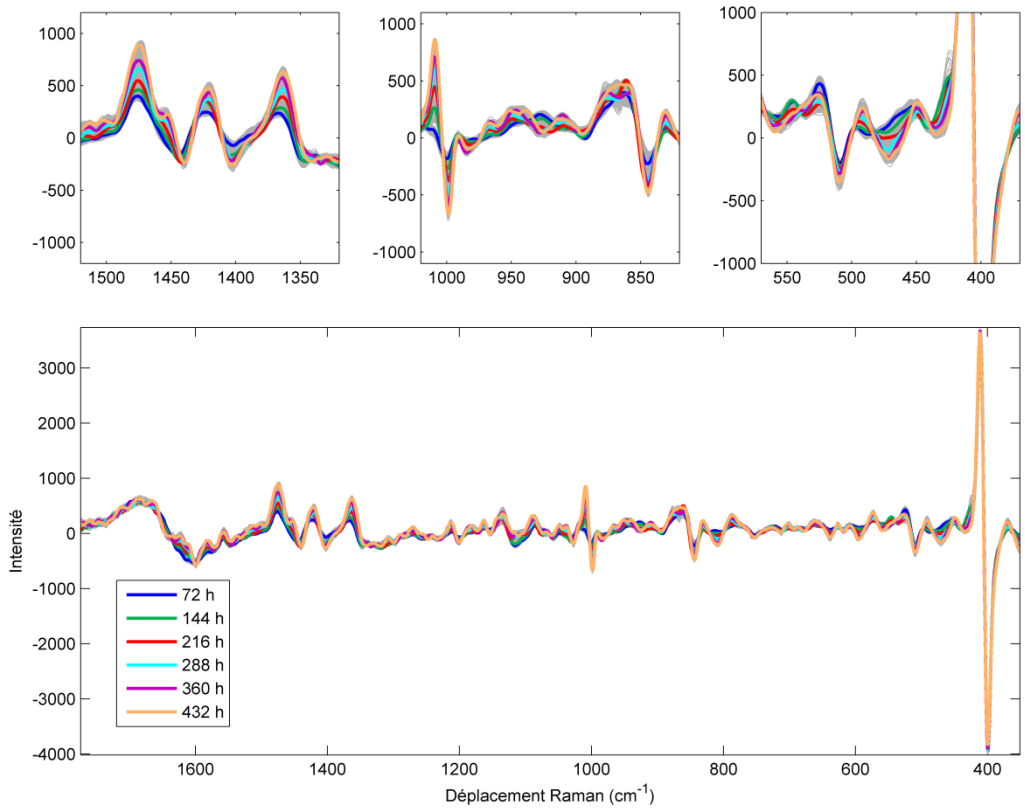

Figure 3.5 Représentation de l'effet de la correction de ligne de base par la dérivée première sur la région 1775–350 cm^{-1}. Les spectres utilisés sont les mêmes que Figure 3.2.

Nous disposons alors de six jeux de données (pour six temps d'acquisition différents) possédant tous 63 mesures de référence et 63 spectres Raman associés. Chaque jeu a été traité indépendamment. Pour cela, les données ont été séparées en deux ensembles distincts à savoir un ensemble de calibration et un ensemble test. Le modèle de régression PLS de chaque biomarqueur est réalisé sur l'ensemble de calibration (ou ensemble d'apprentissage), tandis que l'ensemble test sert uniquement à évaluer les capacités prédictives du modèle et calculer l'erreur RMSEP. Néanmoins, pour chaque modèle calculé, une première validation (validation croisée) est effectuée pour notamment déterminer le nombre de variables latentes à prendre en compte (dans notre cas, une « *10-fold cross-validation* »). Pour chacun des sept biomarqueurs mis en jeu au cours de ces travaux de

recherche (glucose, glutamine, glutamate, lactate, ammonium, TCD, VCD), nous évaluons les différents modèles PLS obtenus et comparons les erreurs RMSEP des 6 jeux de données. Les résultats obtenus pour les cultures de cellules CHO sont représentés Figure 3.6.

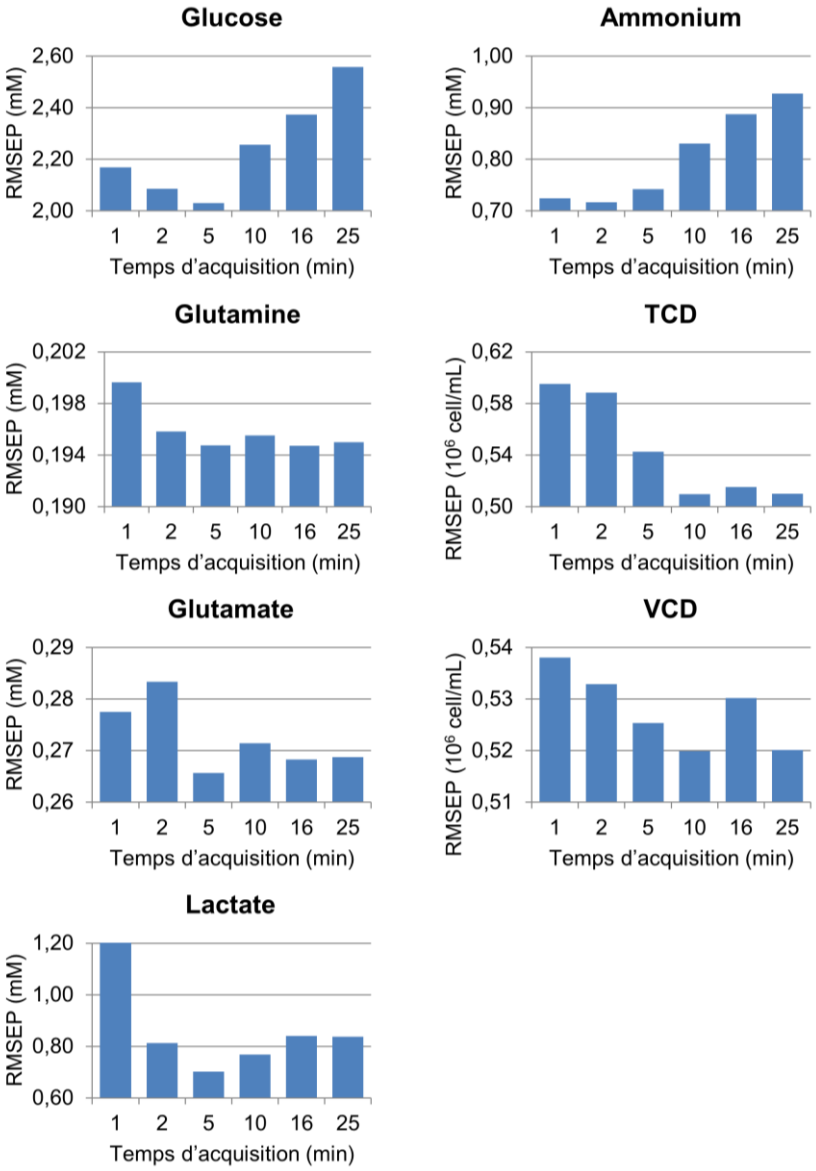

Figure 3.6 Erreurs RMSEP obtenues pour chacun des jeux de données construits (en fonction du temps d'acquisition) sur la base de spectres d'une culture CHO et pour chaque paramètre biochimique étudié.

Nous pouvons observer sur ces résultats que les erreurs RMSEP présentent une tendance pour chaque biomarqueur qui ressort un temps d'acquisition optimal, c'est-à-dire celui dont l'erreur RMSEP est la plus faible. Les caractéristiques de chacun de ces modèles optimaux sont reprises dans le Tableau 3.1. Cependant, il convient ensuite d'analyser les résultats obtenus et interpréter les modèles déterminés.

2.2.1 Modèles pour les concentrations en glucose et lactate

Deux des métabolites les plus importants lors des cultures CHO sont les métabolites glucose et lactate. En effet, le glucose est le produit consommé par les cellules pour entamer la reproduction cellulaire, tandis que le lactate peut être considéré comme le déchet majeur produit par la multiplication des cellules CHO. Pour ces deux paramètres biochimiques, un temps de 5 min d'acquisition a été déterminé. Les graphiques présentés Figure 3.7 permettent d'interpréter les modèles obtenus.

Sur les modèles du glucose et du lactate, les niveaux d'erreurs obtenus sont acceptables par rapport aux gammes de concentration. L'erreur RMSEP du modèle glucose est calculée à 2,03 mM pour une gamme de concentration comprise entre 11,54 mM et 58,70 mM. Pour le modèle lactate, nous déterminons l'erreur RMSEP à 0,70 mM pour une gamme 0,00–10,06 mM. Nous pouvons clairement observer Figure 3.7, sur les graphiques des concentrations prédites par rapport aux concentrations calculées, que les modèles calculés permettent bien de prédire les concentrations en glucose et lactate respectivement. En observant les coefficients de régression, nous pouvons observer un fort niveau de bruit pour les différents modèles, dû aux faibles intensités des signaux et au faible nombre d'échantillons pour construire les modèles (42 échantillons).

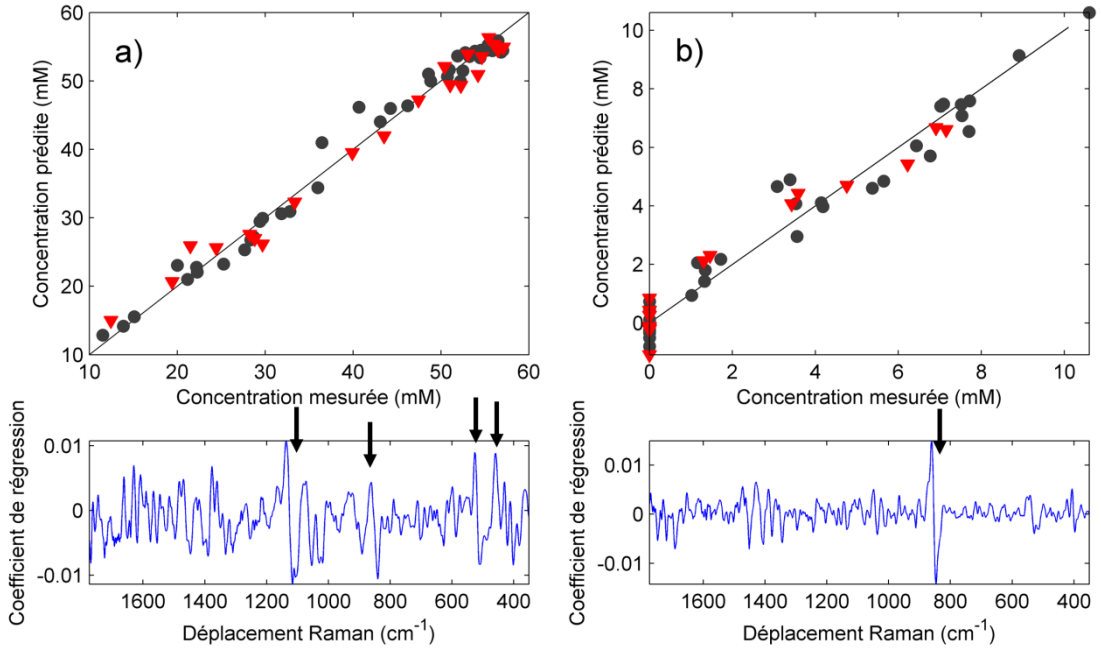

Figure 3.7 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour a) le glucose et b) le lactate d'une culture de cellules CHO. Les spectres des deux modèles possèdent des temps d'acquisition de 10 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Nous pouvons toutefois interpréter les modèles à travers les coefficients obtenus. Pour le glucose d'une part, nous observons trois régions importantes discriminant la molécule de

glucose. D'une part, la zone 1200–1010 cm^{-1} présente des coefficients de régression importants et peut être attribuée aux élongations ν(C–C) et ν(C–O) de la molécule [137]. La région 910–812 cm^{-1} est assignée aux déformations δ(COH), δ(CCH) et δ(OCH) [137]. Enfin, la région spectrale entre 540 cm^{-1} et 400 cm^{-1} désigne les déformations exocycliques (autour de 520 cm^{-1}) et endocycliques (autour de 450 cm^{-1}) de la molécule [137]. Enfin, les bandes autour de 1400 cm^{-1} peuvent être assignées aux déformations δ(CH_2) et δ(CH_2OH) [137], que nous retrouvons également sur les coefficients de régression du modèle lactate. Sur ces derniers pour finir, une bande principale ressort autour de 860 cm^{-1} qui traduit l'élongation ν(C–CO_2^-), caractéristique de la molécule de lactate [138].

2.2.2 Modèles pour les niveaux en glutamine, glutamate et ammonium

Ces trois métabolites présentent également un intérêt pour le suivi de cultures cellulaires. Cependant, dans le cas des cellules CHO, leur action est moins déterminante que le glucose ou le lactate, dont les niveaux de concentration restent plus élevés. Les modèles calculés pour ces trois paramètres biochimiques sont présentés Figure 3.8.

D'après la Figure 3.8 et les erreurs RMSEP présentées Tableau 3.1, nous pouvons dire à première vue que les modèles calculés sont acceptables. En effet, pour la glutamine d'abord, bien que le graphique des concentrations prédites contre les concentrations mesurées (Figure 3.8a) présente des résidus plus importants que les modèles lactate et glucose, l'erreur RMSEP de 0,19 mM reste faible compte tenu de la gamme d'étalonnage (0,00–1,66 mM). En observant les coefficients de régression, nous pouvons voir que deux contributions principales ressortent à 850 cm^{-1} et 1020 cm^{-1}. Elles peuvent être attribuées aux vibrations ν(C–C) de la chaine carbonée [139,140]. De plus, nous pouvons observer des bandes importantes entre 1350 cm^{-1} et 1450 cm^{-1} pouvant être assignées aux groupements amide (élongations ν(C–N) et déformations δ(N–H)), aux groupements carboxyle ainsi qu'aux déformations de la chaine carbonée δ(CH_2) [139,140]. En ce qui concerne le modèle glutamate (Figure 3.8b), le modèle obtenu permet de prédire les échantillons test convenablement avec une faible erreur RMSEP (0,27 mM) par rapport à la gamme d'étalonnage (1,85–6,82 mM). Par rapport à la glutamine, le graphique des concentrations prédites par rapport aux concentrations mesurées présente moins de résidus dans les prédictions, notamment grâce à une gamme de concentration plus élevée et qui varie de manière plus prononcée. Le glutamate est donc plus facile à détecter en spectroscopie Raman. Les molécules de glutamine et glutamate présentant des structures proches, les coefficients de régression sont similaires. Cependant, les signaux observés entre 1350 cm^{-1} et 1450 cm^{-1}, plus intenses pour la molécule de glutamate que la molécule de glutamine,

sont attribués non pas aux groupements amide, mais plutôt aux vibrations $\nu(CO_2^-)$ des groupements carboxylate [137].

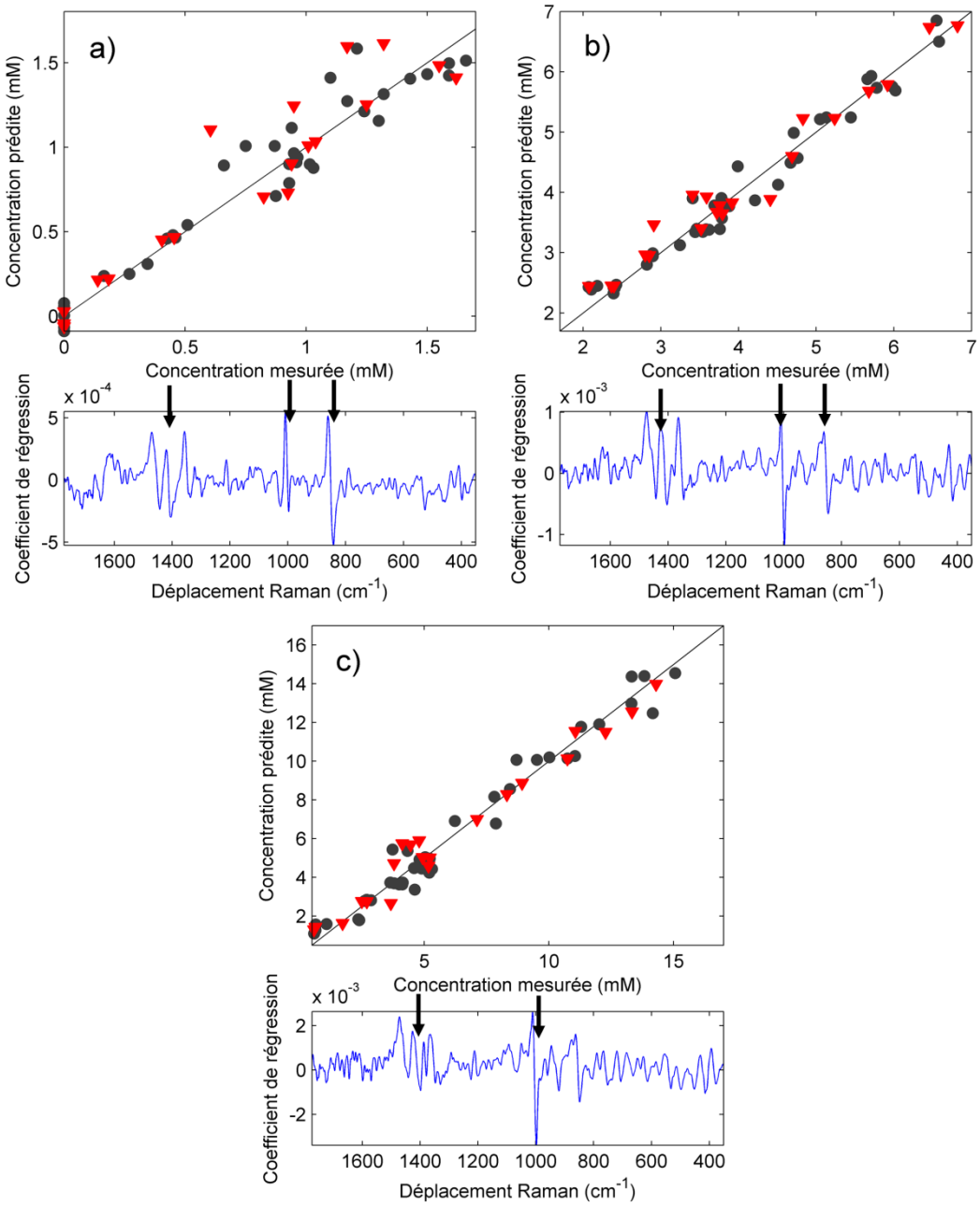

Figure 3.8 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour a) la glutamine, b) le glutamate et c) l'ammonium d'une culture de cellules CHO. Les spectres des modèles glutamine et glutamate possèdent des temps d'acquisition de 10 × 30 s tandis que les spectres du modèle ammonium sont acquis sur 2 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

En ce qui concerne le modèle de régression pour l'ammonium, en prenant en compte le graphique des concentrations prédites par rapport aux concentrations calculées (Figure 3.8c) et l'erreur RMSEP calculée (Tableau 3.1), nous pouvons voir que le modèle est acceptable. Toutefois, les travaux de T. Ujike et Y. Tominaga [141] ont permis de montrer que la molécule d'ammonium (NH_4^+) en solution aqueuse présentait deux bandes spectrales autour

de 1040 cm^{-1} et 1650 cm^{-1}. Hors, les coefficients de régression obtenus ici présentent uniquement une bande caractéristique de la molécule, autour de 1020 cm^{-1}. Nous observons cependant un groupe de bandes autour de 1400 cm^{-1}, similaire à ce que nous avions obtenu pour les molécules de glutamine et glutamate. Ceci traduit la difficulté à modéliser chimiométriquement les teneurs en ammonium pour les suivis de cultures de cellules CHO. Les modèles s'appuient non seulement sur une bande caractéristique de l'ammonium, mais aussi sur celles d'autres produits présents en solution pour les prédictions, ce qui peut devenir problématique par la suite dans la recherche de modèles de régression robustes.

2.2.3 Modèles pour les densités TCD et VCD

Les déterminations des densités cellulaires sont des modèles distincts. En effet, ces-derniers ne se basent pas uniquement sur la seule structure chimique de la molécule tels que les biomarqueurs présentés jusque là, mais sur un ensemble biologique singulier traduisant la présence de cellules ou non. Les caractéristiques des modèles optimaux obtenus pour TCD et VCD sont reprises dans le Tableau 3.1. Les temps d'acquisition permettant d'avoir les erreurs RMSEP les plus faibles proviennent des jeux dont les spectres sont acquis sur 10 min (20 × 30 s). Les représentations de ces modèles chimiométriques sont disponibles Figure 3.9.

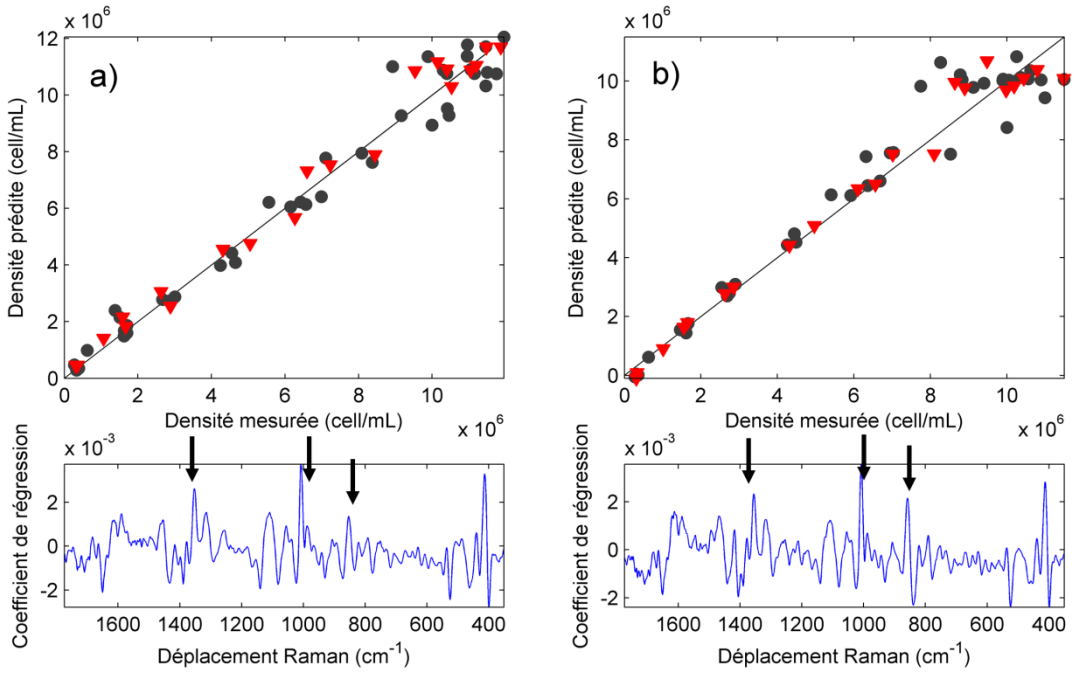

Figure 3.9 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour les paramètres a) TCD et b) VCD d'une culture de cellules CHO. Les spectres des deux modèles possèdent des temps d'acquisition de 20 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Nous pouvons voir que les modèles parviennent tous deux à modéliser les paramètres. De plus, les erreurs RMSEP calculées pour TCD et VCD ($0{,}51 \cdot 10^6$ cell/mL et $0{,}52 \cdot 10^6$ cell/mL respectivement) sont faibles comparées aux gammes de densité ($0{,}27 \cdot 10^6$–$16{,}50 \cdot 10^6$ cell/mL et $0{,}27 \cdot 10^6$–$13{,}21 \cdot 10^6$ cell/mL respectivement). Toutefois, nous pouvons noter des résidus plus importants pour les fortes densités sur les graphiques des densités prédites en fonction des densités mesurées. Ceci provient du grand nombre d'échantillons dans ces fortes gammes. En effet, nous disposons d'un grand nombre de mesures de prélèvement pour les fortes densités cellulaires puisqu'il s'agit de la phase de plateau durant laquelle la viabilité au sein du bioréacteur stagne à son plus haut niveau (Figure 1.5, page 31, pour les tendances des densités cellulaires pour les cultures de cellules CHO). De ce fait, le nombre de répétitions présentant des spectres différents exalte la variabilité de la méthode de référence sur le modèle de régression, ce qui se traduit par des résidus plus importants Figure 3.9.

En ce qui concerne les coefficients de régression sur lesquels se reposent les deux modèles calculés pour les densités cellulaires, nous pouvons voir que les résultats sont similaires. Tout d'abord, une bande majoritaire ressort autour de 1000 cm^{-1}. Celle-ci est assignée à la respiration du cycle benzénique (élongations v(C–C)) de la molécule de phénylalanine, molécule omniprésente dans les cellules cultivées qui en fait donc un très bon traceur [137,142]. De plus, les bandes observées entre 1450 cm^{-1} et 1300 cm^{-1} peuvent être attribuées aux respirations des cycles des molécules adénine et guanine, constitutives des chaines d'ADN ou d'ARN [137,142]. La différence notable qui peut être observée entre les coefficients de régression des deux modèles provient de la bande à 850 cm^{-1}. Cette dernière est plus importante pour le modèle calculé pour le paramètre VCD (Figure 3.9b) tandis qu'elle ne présente pas d'impact majeur sur les coefficients du modèle TCD (Figure 3.9a). Certains travaux permettent d'attribuer cette bande à la molécule de tyrosine [142], dont l'effet bénéfique sur la survie cellulaire a été démontré [143]. L'attribution de cette bande est donc cohérente quant à la différenciation entre les deux modèles par rapport à la viabilité des cellules.

Finalement, la détermination du temps d'acquisition des spectres Raman pour les cultures de cellules CHO fait ressortir deux temps : 5 min et 10 min (le modèle ammonium n'étant pas pris en compte puisque les coefficients de régression ne traduisent pas tout à fait la nature de la molécule). Cependant, d'un point de vue pratique, il est plus simple d'utiliser un seul temps d'acquisition. Ainsi, nous avons établi le temps d'acquisition des spectres Raman à 5 min (10 × 30 s) puisque d'une part, quatre des sept paramètres métaboliques présentent les modèles les plus performants pour ce temps d'acquisition. D'autre part, les modèles pour les densités cellulaires (seuls modèles à 10 min d'acquisition) présentent tout

de même des modèles suffisamment prédictifs pour réaliser un suivi de culture de cellules CHO à 5 min d'acquisition. Néanmoins, si nous avions besoin de temps d'acquisition différents, nous pourrions générer numériquement n'importe quel temps d'acquisition sur la base de spectres acquis toutes les 30 s.

2.3 Temps d'acquisition pour les cultures de cellules HeLa

Après avoir déterminé le temps d'acquisition optimal pour les cultures de cellules CHO, nous réalisons le même travail pour les cultures virales mettant en jeu les cellules HeLa. Ces procédés se décomposent en deux phases distinctes : l'amplification cellulaire pour la croissance des cellules et la culture virale durant laquelle les virus injectés prolifèrent en s'appuyant sur les cellules cultivées. Dans notre cas, le procédé n'est pas interrompu entre les deux phases, la détermination du temps d'acquisition est donc réalisée sur un lot de données provenant d'une seule culture mettant en jeu les deux phases du bioprocédé de culture virale.

La culture mise en jeu pour la détermination du temps d'acquisition a duré plus de 215 h (soit 9 jours). L'inoculation du virus est réalisée au bout de 144 h (6 jours d'amplification cellulaire) dans le mode opératoire dont il est question ici. De plus, un *feed* est réalisé à 72 h afin d'apporter glucose et glutamine au sein du bioréacteur. Au cours du procédé, un total de 70 prélèvements a été réalisé pour le développement des modèles de régression dont 40 durant la phase de croissance cellulaire et 30 tout au long de l'infection virale. Un ensemble de 47 échantillons a été conservé pour l'étalonnage des modèles tandis que les 23 autres forment le lot test.

De plus, tout au long de la culture, des spectres Raman (1 × 30 s) ont été acquis en continu. La même méthode de construction des jeux de données ainsi que les mêmes temps que pour les cultures de cellules CHO ont été employés pour les bioprocédés HeLa. Pour les six jeux de données produits, les prétraitements et le domaine spectral restent inchangés.

Les calculs chimiométriques sont réalisés pour chacun des six différents lots de données et pour les sept principaux métabolites. Les modèles de régression PLS sont obtenus à partir des 47 mesures de référence pour l'étalonnage et des spectres Raman associés dans le temps. Analogiquement aux calculs chimiométriques réalisés pour les données des cellules CHO, une *10-fold cross-validation* est réalisée pour estimer le nombre de variables latentes à considérer et les performances des modèles sont estimées à travers l'erreur RMSEP calculée en appliquant les modèles sur les 23 échantillons du lot test. Les résultats obtenus sont récapitulés dans le Tableau 3.2 et représentés Figure 3.10.

Tableau 3.2 Paramètres obtenus pour les modèles PLS les plus performants par biomarqueur dans l'étude du temps d'acquisition d'une culture de cellules HeLa.

Biomarqueur (unité)	Temps	R^2	LV	RMSEP	Gamme
Glucose (g/L)	5 min	0,986	4	0,18	0,09 – 5,13
Glutamine (mM)	5 min	0,986	4	0,14	0,12 – 4,24
Glutamate (mM)	5 min	0,919	4	0,11	0,95 – 2,55
Lactate (g/L)	16 min	0,965	5	0,05	0,57 – 1,69
Ammonium (mM)	5 min	0,952	4	0,13	1,12 – 3,77
VCD (10^6 cell/mL)	10 min	0,851	4	0,56	0,67 – 10,41
TCD (10^6 cell/mL)	10 min	0,919	5	0,57	0,67 – 10,66

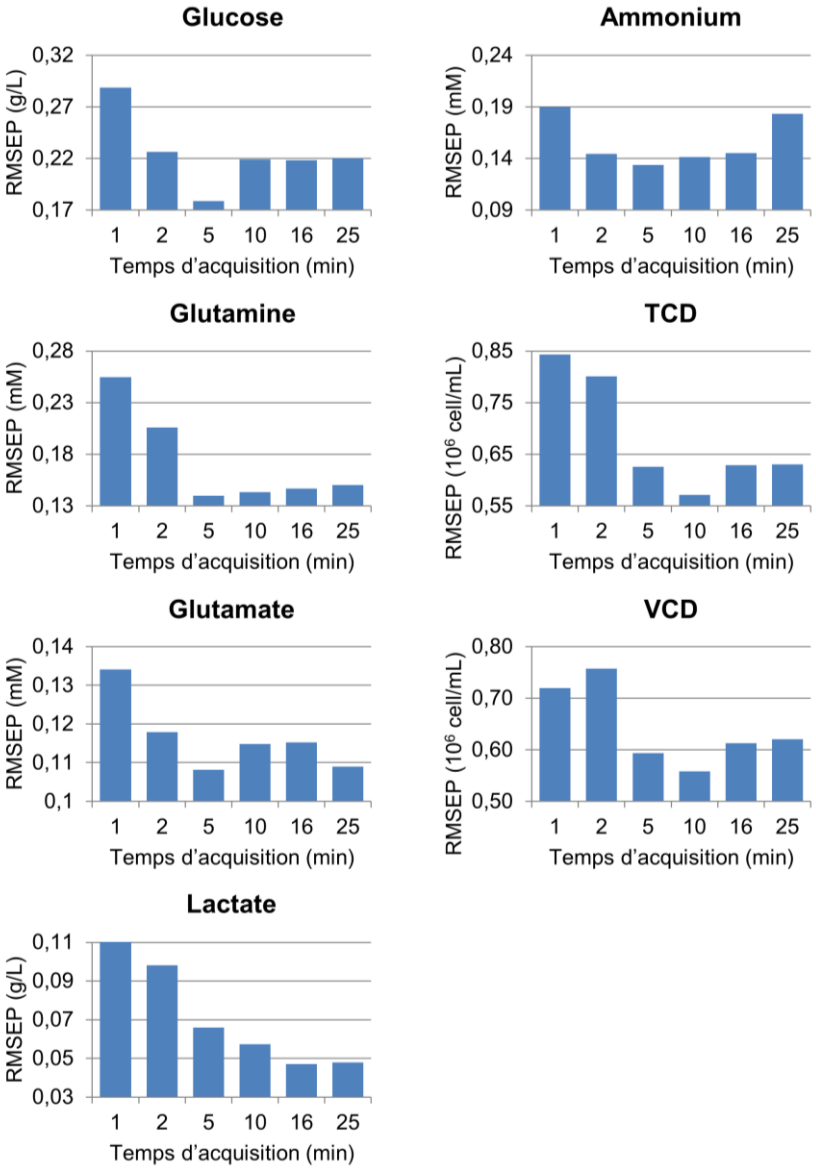

Figure 3.10 Erreurs RMSEP obtenues pour chacun des jeux de données construits (en fonction du temps d'acquisition) sur la base de spectres d'une culture HeLa (amplification cellulaire puis culture virale) et pour chaque paramètre biochimique étudié.

2.3.1 Modèles pour les concentrations en glucose et glutamine

Pour les procédés mettant en jeu des cellules HeLa, le glucose et la glutamine sont deux paramètres métaboliques importants dans le sens où il s'agit des deux produits consommés par les cellules pour la multiplication cellulaire. De ce fait, il est important de suivre ces deux produits afin non seulement d'évaluer la consommation de la cellule, mais aussi d'assurer l'abondance de ces métabolites dans le milieu de culture.

Pour ces deux biomarqueurs, nous avons obtenu le même temps d'acquisition optimal de 5 min (10 × 30 s). En effet, les résultats représentés Figure 3.10 permettent de montrer qu'il existe un RMSEP minimum pour les molécules de glucose et glutamine en fonction du temps d'acquisition global. Les deux modèles calculés sur la base des spectres de 5 min sont représentés Figure 3.11.

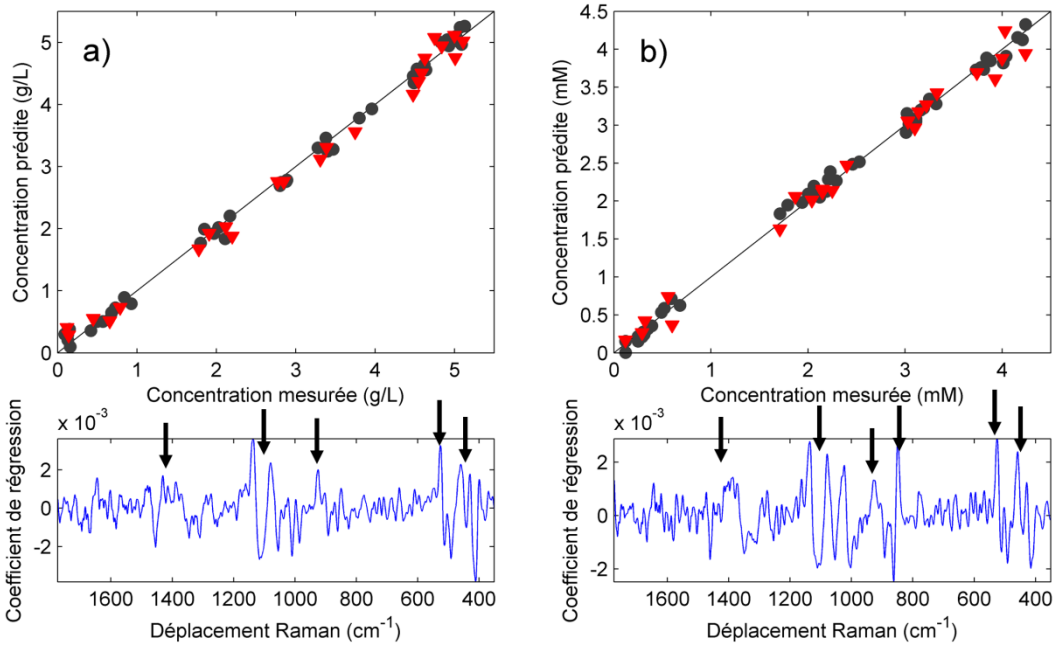

Figure 3.11 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour a) le glucose et b) la glutamine d'une culture de cellules HeLa. Les spectres des deux modèles possèdent des temps d'acquisition de 10 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Pour le glucose d'une part, nous pouvons voir que le modèle de régression PLS présente une faible erreur RMSEP (0,18 g/L) compte-tenu des variations de la gamme de concentration (0,09–5,13 g/L), données reprises Tableau 3.2. Nous pouvons également constater sur le graphique des concentrations prédites en fonction des concentrations mesurées (Figure 3.11a) que le modèle est tout à fait capable de prédire les concentrations en glucose du jeu test. De plus, d'après les coefficients de régression du modèle glucose, nous pouvons assigner les bandes principales à la molécule. Ainsi, les bandes majeures apparaissant entre 1200 cm^{-1} et 1000 cm^{-1} sont attribuées aux élongations ν(C–C) et ν(C–O) de la molécule de glucose [137]. Nous pouvons également observer les bandes

caractéristiques des déformations exocycliques (520 cm^{-1}) et endocycliques (450 cm^{-1}) de la molécule de glucose [137]. Moins intenses, les bandes autour de 900 cm^{-1} peuvent être assignées aux déformations δ(COH) et δ(CCH), quant aux bandes autour de 1400 cm^{-1}, elles sont attribuées aux déformations δ(CH_2) et δ(CH_2OH) du groupement hydroxyméthyle de la molécule de glucose [137].

Le modèle glutamine présente également des performances suffisamment élevées pour réaliser les prédictions du jeu test, en témoigne l'erreur RMSEP calculée (0,14 mM) pour la gamme de concentration mise en jeu (0,12–4,24 mM), Tableau 3.2. De plus, nous pouvons observer que les coefficients de régression (Figure 3.11b) sont très similaires à ceux de la molécule de glucose. L'unique différence qui existe entre les graphiques proposés est l'apparition d'une bande importante à 850 cm^{-1}, ici assignée aux élongations ν(C–C) de la chaîne carbonée, bande très intense sur les spectres de glutamine [140]. De plus, les signaux autour de 1400 cm^{-1}, assignés aux élongations ν(C–N) et aux déformations δ(CH_2) permettent de caractériser la molécule de glutamine. Toutefois, nous pouvons observer que le modèle de régression s'appuie en grande partie sur les variations spectrales observées entre 500–400 cm^{-1}, sans que la molécule de glutamine ne présente de bande de vibration caractéristique en spectroscopie Raman dans cette région. Ceci est crédité à l'évolution de la concentration en glutamine durant la culture, évolution quasi-similaire à celle du glucose (Figure 1.6, page 32 pour l'amplification cellulaire et Figure 1.7, page 33 pour la culture virale), mais présentant des niveaux de concentration inférieurs. De ce fait, le modèle de régression s'appuie sur les variables présentant la variabilité la plus à même de retranscrire l'évolution et la tendance des réponses apportées à l'ensemble de calibration. Autrement dit, bien qu'il s'agisse de mesures de la quantité de glutamine dans le milieu, les bandes spectrales entre 500–400 cm^{-1} permettent en partie d'expliquer l'évolution du biomarqueur, sans que celui-ci ne soit responsable des variations observées sur les spectres Raman. La robustesse du modèle est donc remise en cause quant à d'éventuelles variations dans l'évolution du paramètre glutamine sur d'autres cultures. Cependant, dans la présente étude visant à déterminer le temps optimal d'acquisition, nous travaillons sur une unique culture et ne pouvons donc pas évaluer les six modèles (correspondant aux six temps d'acquisition) sur d'autres jeux test. Un total de dix accumulations de 30 s pour un temps total de 5 min est donc ici le choix le plus convenable pour la glutamine.

2.3.2 Modèles pour les concentrations en glutamate, lactate et ammonium

Lors des cultures virales mettant en jeu les cellules HeLa, les trois molécules glutamate, lactate et ammonium sont produites par les cellules. Les erreurs RMSEP des modèles obtenus sur la base des spectres Raman acquis à différents temps et des mesures de

référence sont présentées Figure 3.10. Nous pouvons voir que pour le glutamate et l'ammonium, un RMSEP minimum est atteint pour des spectres de 5 min tandis que le modèle lactate ayant l'erreur la plus faible concerne les spectres acquis sur 16 min.

Le modèle glutamate, obtenu à partir de spectres de 10 × 30 s, est représenté Figure 3.12a. L'erreur RMSEP obtenue pour ce modèle (0,11 mM) est tout à fait satisfaisante par rapport à la gamme de concentration étudiée (0,95–2,55 mM). Toutefois, sur le graphique des concentrations prédites en fonction des concentrations mesurées, nous pouvons voir que le modèle présente certains écarts sur les références entre 1,5 mM et 1,7 mM. Néanmoins, l'étude des coefficients de régression permet de montrer que le modèle se repose sur des bandes spectrales caractéristiques de la molécule de glutamate. En effet, trois zones ressortent principalement sur ces coefficients, Figure 3.12a. D'une part, les bandes principales sont observées à 850 cm^{-1} et 1020 cm^{-1}, tout comme elles l'avaient été pour le modèle glutamate de la lignée cellulaire CHO proposé Figure 3.8b. Ces bandes sont directement assignées aux élongations ν(C–C) de la chaine carbonée [137]. D'autre part, les bandes entre 1450–1350 cm^{-1} sont assignées aux élongations symétriques du groupement carboxylate $\nu(CO_2^-)$ et aux déformations $\delta(CH_2)$ de la molécule de glutamate [137,140]. Enfin, autour de 1100 cm^{-1}, les bandes de vibration sont attribuées aux élongations ν(C–N) du groupement amine de la molécule [140].

Pour ce qui est du modèle basé sur les références lactate, bien que les différences soient non-significatives, ce sont des spectres de 16 min qui proposent le modèle PLS présentant le RMSEP le plus faible, Figure 3.10. Cette valeur fait exception par rapport aux autres métabolites (pour ce type de cellule) basés sur des temps de 5 min. L'observation des coefficients de régression permet de montrer que ceux-ci caractérisent bien la molécule de lactate à première vue. En effet, la bande caractéristique du lactate à 850 cm^{-1}, traduisant la présence des élongations symétriques $\nu(CO_2^-)$ est très importante, chose que nous avions déjà observée pour le modèle lactate des cellules CHO (Figure 3.7b). Toutefois, nous pouvons voir que le modèle de régression se base également sur la région autour de 1000 cm^{-1}. Ceci n'est pas anodin puisqu'il s'agit des élongations ν(C–C), certes présentes au sein de la molécule de lactate, mais caractérisant plutôt les molécules telles que glutamine ou glutamate pour leurs chaines carbonées plus longues.

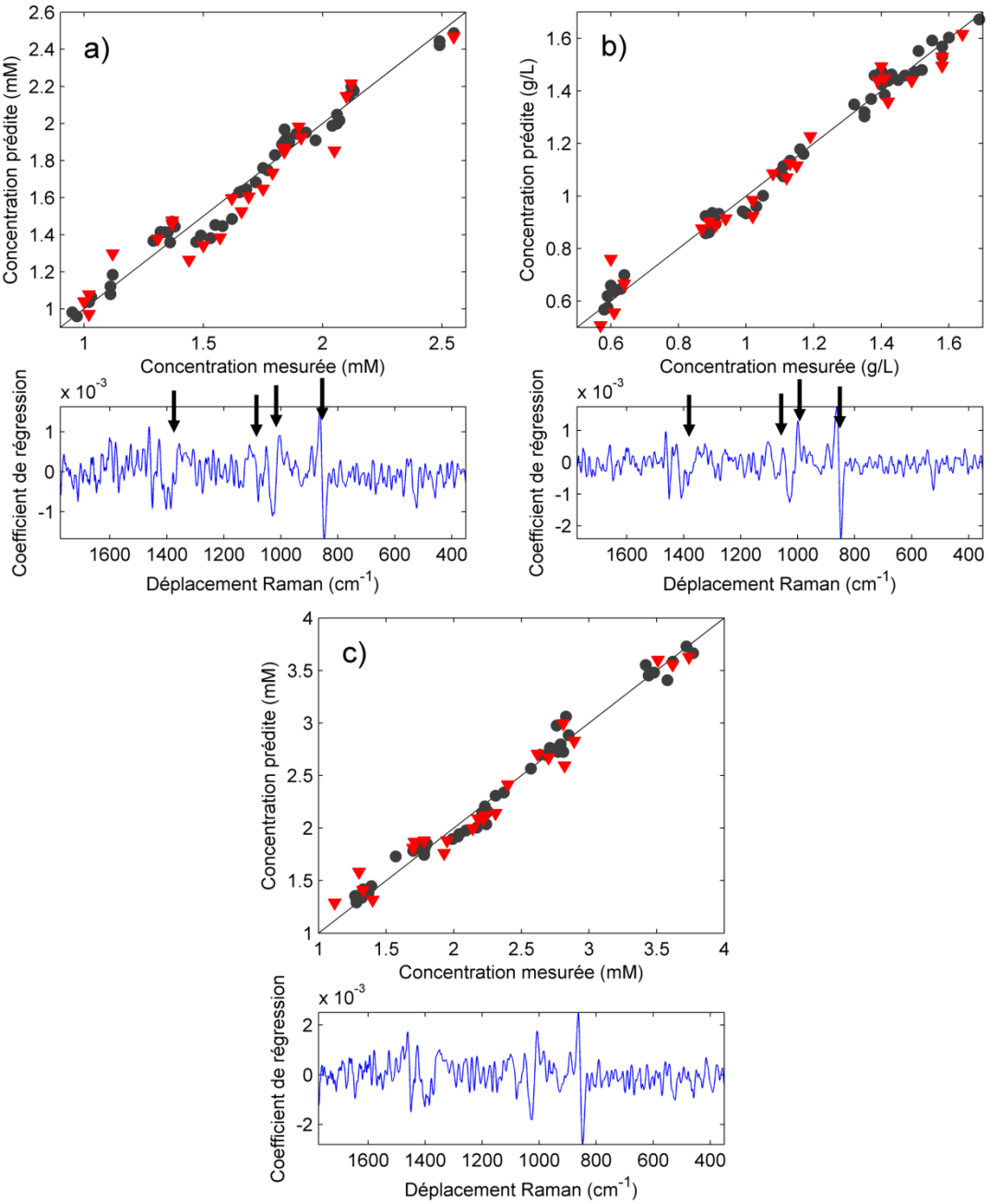

Figure 3.12 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour les molécules a) de glutamate, b) de lactate et c) d'ammonium d'une culture de cellules HeLa. Les spectres des modèles glutamate et ammonium possèdent des temps d'acquisition de 10 × 30 s tandis que les spectres du modèle lactate présenté sont acquis sur 32 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

De manière plus générale, nous pouvons voir que les coefficients de régression des modèles basés sur les molécules glutamate, lactate et ammonium sont très similaires. L'explication repose une nouvelle fois sur les tendances des métabolites durant toute la durée du bioprocédé. En effet, nous pouvons voir Figure 1.6 (page 32) et Figure 1.7 (page 33) que les évolutions de ces trois produits sont les mêmes. Nous nous retrouvons alors dans le même cas que le glucose et la glutamine, section précédente. Bien que les modèles présentent des performances satisfaisantes, ils ne reposent pas sur les variables

caractéristiques de chacune des molécules. C'est notamment le cas du lactate et de l'ammonium ici.

Il n'est donc pas pertinent de prendre en compte les résultats obtenus pour les modèles de régression des molécules de lactate et d'ammonium pour l'optimisation du temps d'acquisition. D'après les évaluations des coefficients de régression des modèles calculés, ce sont les biomarqueurs glucose et glutamate qui présentent le plus de fiabilité.

2.3.3 Modèles pour les densités TCD et VCD

Les biomarqueurs TCD et VCD présentent tous deux une erreur RMSEP minimum pour des temps d'acquisition de 10 min (Figure 3.10) et satisfaisante compte tenu des gammes de densité mises en jeu (Tableau 3.2). Les deux modèles de régression basés sur les spectres de 10 min pour les références TCD et VCD sont représentés Figure 3.13.

Nous pouvons observer l'existence de 4 échantillons (3 étalons, 1 test) particuliers sur les graphiques des densités prédites en fonction des densités mesurées. Il s'agit en fait de mesures réalisées en fin d'amplification cellulaire, donc au moment où le seuil maximal de cellules est atteint dans le bioréacteur. L'absence de mesure entre ces échantillons et le reste de la gamme est due au calendrier. Ces échantillons présentent donc un fort levier comparé aux autres, mais sont néanmoins nécessaires dans la construction des modèles afin de capter la croissance exponentielle des densités cellulaires.

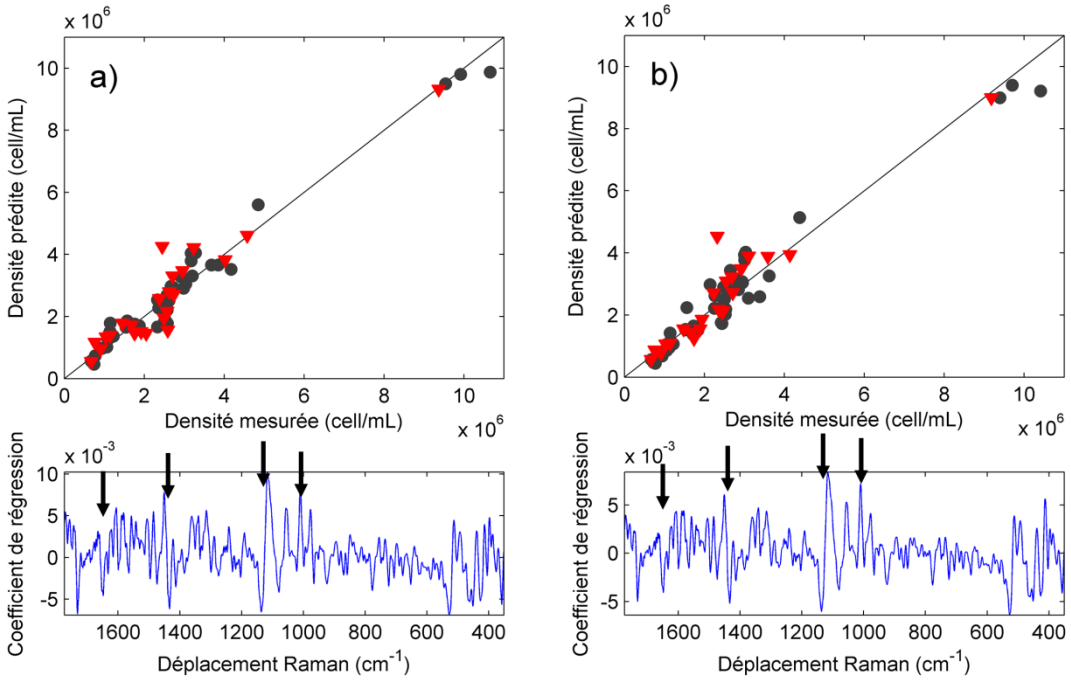

Figure 3.13 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour les paramètres a) TCD et b) VCD d'une culture de cellules HeLa. Les spectres des deux modèles possèdent des temps d'acquisition de 20 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

En observant les coefficients de régression des modèles, nous pouvons remarquer encore une fois que ceux-ci sont similaires. En effet, les tendances des deux biomarqueurs sont sensiblement les mêmes durant le procédé. Dans un premier temps, durant la multiplication cellulaire, les deux tendances sont immanquablement liées. La différenciation entre les deux paramètres métaboliques intervient suite à la phase plateau durant la culture de cellules. Cependant, les cellules HeLa sont ici utilisées comme support pour la culture virale. De ce fait, la phase de croissance cellulaire est stoppée le plus rapidement possible une fois la phase plateau atteinte, avant la phase de sénescence. Nous n'avons donc presqu'aucune différence entre VCD et TCD. Lors de la phase de culture virale, bien que le virus intervienne, les cellules continuent de se multiplier de la même manière. Nous pouvions alors nous attendre à une différenciation entre les paramètres VCD et TCD, mais la méthode de référence ne doit pas parvenir à détecter les cellules lysées. C'est pourquoi, les deux paramètres métaboliques sont si semblables tout au long du bioprocédé, même pendant la phase de culture virale.

Les coefficients de régression de ces deux modèles permettent toutefois de montrer que nous suivons bien les paramètres mettant en jeu la physiologie des cellules HeLa. Tout d'abord, les coefficients les plus importants sont situés autour de 1100 cm^{-1}, assignés aux élongations symétriques $\nu(PO_2^-)$ des colonnes de macromolécules d'ADN [144]. Nous pouvons également souligner les coefficients de régression compris à 1000 cm^{-1}, traduisant les élongations ν(C–C) du cycle benzénique de la molécule de phénylalanine, abondante dans les cellules [145]. Les bandes autour de 1400 cm^{-1} sont associées aux déformations $\delta(CH_2)$ de diverses molécules des cellules [145], tandis que les bandes autour de 1600 cm^{-1} reflètent les modes de vibrations des amides primaires présentes sur les liaisons peptidiques des acides aminés [145].

Finalement, en ce qui concerne la détermination du temps d'acquisition pour les cellules HeLa, nous nous retrouvons dans le même cas de figure que pour les cellules CHO, à savoir des temps de 5 min pour les métabolites et 10 min pour les densités cellulaires. Nous choisissons donc de travailler uniquement sur des temps d'acquisition de 5 min (10 × 30 s), ce qui permet d'une part de simplifier l'acquisition pour le suivi de culture HeLa (un seul et unique temps d'acquisition), et d'autre part de permettre à terme de fusionner les données provenant de différents types de cellules (HeLa et CHO) pour la construction d'un modèle global de prédiction.

2.4 Temps d'acquisition pour les cultures de cellules Sf9

Les cellules Sf9 employées au cours de ces travaux jouent le rôle d'hôte pour la culture de virus, telles que les cellules HeLa. Le procédé de culture se décompose aussi en deux

phases : une première permettant de cultiver les cellules, une seconde pour cultiver le virus. Toutefois, contrairement au bioprocédé des cellules HeLa, les deux phases ne sont pas réalisées à la suite, ce qui empêche tout suivi en continu. C'est pourquoi la culture mise en jeu pour la détermination du temps d'acquisition des spectres Raman pour les cellules Sf9 présente uniquement la phase de croissance cellulaire. Il s'agit d'une culture ayant duré 192 h, soit un total de 8 jours. 63 prélèvements pour mesure de référence ont été réalisés au cours de la culture, séparés en 47 échantillons de calibration et 16 échantillons test, répartis de façon homogène sur l'ensemble des gammes de concentration et de densité des biomarqueurs.

Les spectres ont été acquis de la même façon que pour les cellules HeLa et CHO, à savoir en continu sur 1 × 30 s. Afin d'atteindre un temps d'acquisition optimal, nous procédons de la même façon que pour les deux autres types de cellule en construisant six différents jeux de données à des temps différents. Les spectres des jeux de données sont traités de manière similaire aux spectres acquis lors des suivis des cultures de cellules CHO et HeLa (dérivée première de Savitzky-Golay prenant une fenêtre mobile de 15 points et un polynôme d'ordre 2 puis sélectionnant uniquement la région spectrale 1775–350 cm^{-1}). Pour chacun des six jeux de données formés, en prenant les 47 échantillons de calibration, le nombre de variables latentes à prendre en compte pour le modèle est déterminé à l'aide des calculs d'erreurs RMSECV obtenus grâce à une *10-fold cross-validation*. Les modèles au nombre de variables latentes optimal sont finalement appliqués sur les 16 échantillons test afin d'évaluer l'erreur RMSEP pour chaque temps d'acquisition.

Les résultats obtenus pour l'étude des temps sont représentés Figure 3.14. Les caractéristiques des modèles présentant les erreurs RMSEP les plus faibles, considérés comme étant les modèles les plus performants pour les paramètres métaboliques glucose, glutamine, glutamate et TCD sont récapitulées dans le Tableau 3.3.

Tableau 3.3 Paramètres obtenus pour les modèles PLS les plus performants par biomarqueur dans l'étude du temps d'acquisition d'une culture de cellules Sf9.

Biomarqueur (unité)	Temps	R^2	LV	RMSEP	Gamme
Glucose (g/L)	5 min	0,940	3	0,21	8,66 – 10,80
Glutamine (mM)	5 min	0,979	4	0,58	1,86 – 6,98
Glutamate (mM)	16 min	0,939	4	0,41	9,28 – 12,78
TCD (10^6 cell/mL)	5 min	0,989	3	0,15	2,04 – 4,32

Il est important de noter que certains métabolites ne seront pas représentés pour cette étude. En effet, la lignée Sf9 mise en jeu étant très peu consommatrice de produits du milieu de culture, elle est également très peu productrice de lactate. Ainsi, les teneurs en lactate

dans le milieu de culture sont systématiquement sous la limite de quantification (Figure 1.9, page 34). De plus, étant donné que les cultures de cellules Sf9 ont pour but final de permettre de cultiver du virus, il convient d'arrêter la culture le plus rapidement possible une fois la phase de plateau atteinte (en termes de densité cellulaire). Ainsi, nous pouvons réaliser les mêmes observations que pour la cellule HeLa en ce qui concerne les paramètres VCD et TCD : ceux-ci présentent des évolutions très similaires durant la culture. Nous choisissons donc de ne travailler que sur le paramètre TCD, proposant une gamme plus importante que VCD. Enfin, les observations que nous avons pu faire quant aux modèles de régression s'appuyant sur la concentration en ammonium pour les deux premiers types de cellule (CHO et HeLa) sont les mêmes pour la cellule Sf9. De ces faits, nous ne présenterons pas de résultats ni pour les concentrations en lactate et en ammonium, ni pour la densité en cellules vivantes VCD.

En dernier point concernant la culture, nous pouvons souligner à nouveau que les cellules nécessitent peu de produits pour se multiplier. Ces derniers sont alors apportés en abondance dès le début de la culture de cellules. Ainsi, aucun *feed* n'est nécessaire durant toute la durée du bioprocédé.

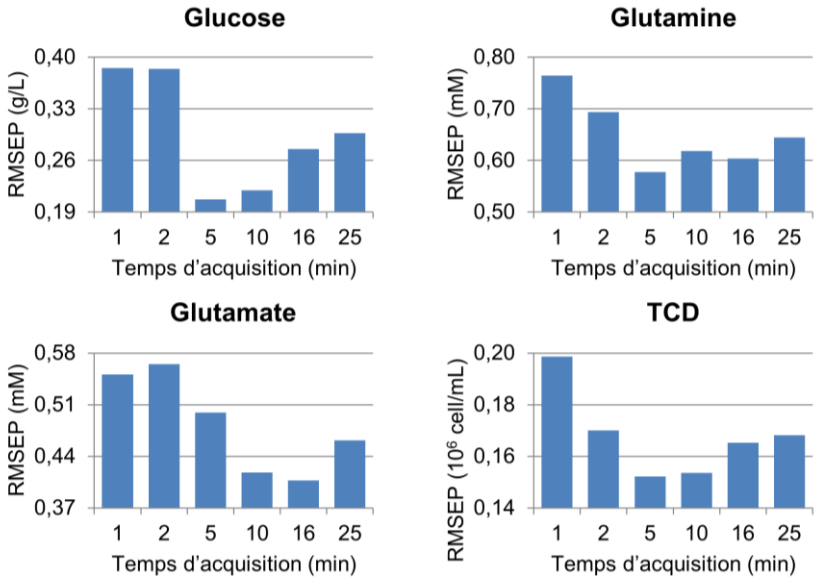

Figure 3.14 Erreurs RMSEP obtenues pour chacun des jeux de données construits (en fonction du temps d'acquisition) sur la base de spectres d'une culture Sf9 (amplification cellulaire uniquement) et pour les paramètres biochimiques glucose, glutamine, glutamate et TCD.

2.4.1 Modèles pour les concentrations en glucose, glutamine et glutamate

Dans un premier temps, nous évaluons les résultats obtenus pour les métabolites conservés pour l'étude du temps d'acquisition total. Il s'agit des paramètres glucose, glutamine et glutamate, dont les résultats sont représentés Figure 3.15. Dès lors, nous pouvons observer que les coefficients de régression de chaque modèle présentent des

niveaux de bruit élevés. Ceci provient notamment des faibles variations des gammes de concentrations. En effet, nous rappelons ici que les cellules sont peu consommatrices des produits présents dans le milieu.

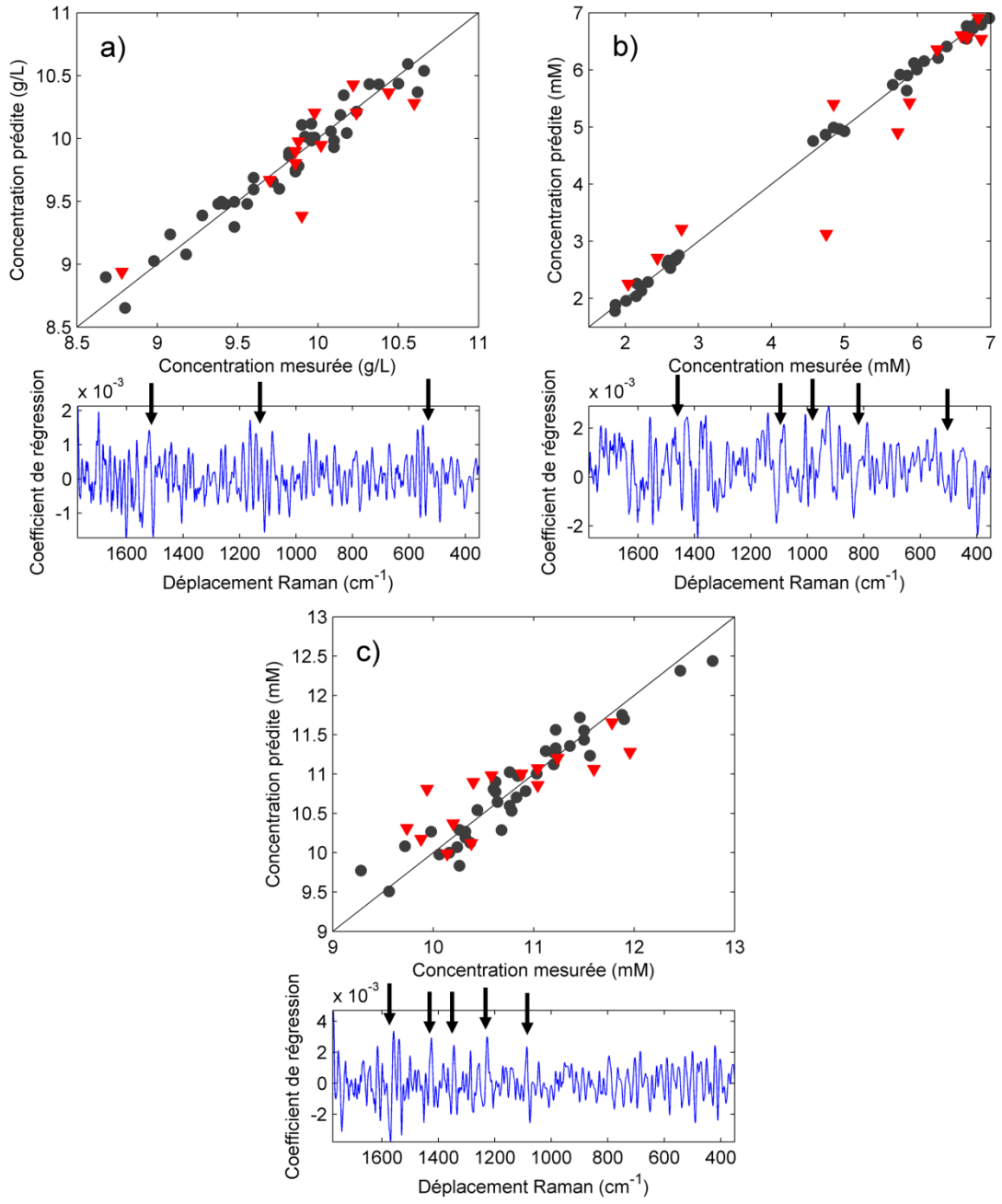

Figure 3.15 Représentations graphiques des modèles calculés lors de la détermination du temps d'acquisition pour les paramètres a) glucose, b) glutamine et c) glutamate d'une culture de cellules Sf9. Les spectres des modèles glucose et glutamine possèdent des temps d'acquisition de 10 × 30 s tandis que les modèles glutamate sont acquis sur 32 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

D'après les résultats obtenus pour chacun des modèles optimaux, disponibles Tableau 3.3, nous pouvons voir que la majeure partie des modèles tendent à proposer de travailler sur des spectres dont le temps d'acquisition est de 5 min (10 × 30 s), Figure 3.14. Tout

d'abord en ce qui concerne le modèle de régression calculé pour les concentrations en glucose, nous pouvons observer qu'il présente une erreur RMSEP faible comparée à la gamme de concentration à l'étude (0,21 g/L par rapport 8,66–10,80 g/L). Nous en étudions alors les coefficients de régression, présentés Figure 3.15a. Nous pouvons dégager trois régions dominantes sur ces coefficients de régression : 1600–1400 cm^{-1}, 1200–1050 cm^{-1} et 600–450 cm^{-1}. La première région est assignée aux déformations $\delta(CH_2)$ et $\delta(CH_2OH)$ de la molécule de glucose [137]. La deuxième zone est attribuée aux élongations ν(C–C) du cycle benzénique de la molécule ainsi qu'aux élongations ν(C–O) des fonctions alcools [137]. Enfin, la région située entre 600–450 cm^{-1} caractérise les déformations endo- et exocycliques du noyau de la molécule de glucose [137].

Pour ce qui est du modèle de régression PLS caractérisant la molécule de glutamine, nous pouvons observer les résultats Figure 3.15b. Il s'agit de la molécule présentant le plus fort niveau de variation dans sa gamme (1,86–6,98 mM). L'erreur RMSEP de 0,58 mM est faible compte tenu de la gamme proposée. Toutefois, nous pouvons souligner quelques échantillons mal prédits sur le graphique des concentrations prédites par rapport aux concentrations mesurées (Figure 3.15b). En étudiant les coefficients de régression, nous pouvons voir que le modèle ne s'appuie pas seulement sur les bandes caractéristiques de la molécule de glutamine. Nous pouvons assigner les bandes observées entre 1600–1400 cm^{-1} aux déformations $\delta(CH_2)$ de la chaine carbonée [140], ainsi qu'aux déformations $\delta(NH_2)$ et élongations ν(C–N) des groupements amine [140]. Il en est de même pour les bandes autour de 1100 cm^{-1}, également attribuées aux élongations ν(C–N) [140]. Les bandes situées entre 1000–900 cm^{-1} traduisent quant à elles les élongations ν(C–C) et les déformations $\delta(CH_2)$ de la chaine carbonée [137,140]. Bien que ces différentes bandes servent à caractériser la molécule de glutamine, nous pouvons observer sur les coefficients de régression des signaux importants autour de 800 cm^{-1} d'une part, entre 500 cm^{-1} et 400 cm^{-1} d'autre part. Cependant, les spectres effectués sur des solutions de glutamine permettent de montrer qu'il n'existe pas de signaux importants pour la molécule dans ces zones spectrales [140]. Nous pouvons donc conclure que ces dernières bandes ne sont pas directement dues aux variations de concentration de ce biomarqueur pendant la culture cellulaire. Le modèle de régression proposé pour la molécule de glutamine n'est donc pas parfaitement robuste pour la prédiction d'échantillons inconnus (absents de la construction du modèle, soit du jeu de calibration).

Enfin, en ce qui concerne le modèle de régression pour le paramètre glutamate, les temps d'acquisition utilisés sont de 16 min (soit 32 × 30 s). Bien que l'erreur RMSEP soit très satisfaisante (0,41 mM) compte tenu de la gamme (9,28–12,78 mM), Tableau 3.3, nous pouvons observer sur la Figure 3.15c que les coefficients de régression présentent peu de

signaux sortant du bruit. Nous pouvons sortir certaines régions intenses, dont celle autour de 1600 cm^{-1} (déformations $\delta(NH_2)$) [140], vers 1420 cm^{-1} (déformations $\delta(CH_2)$) [137], vers 1350 cm^{-1} (élongations ν(C–N) et $\nu(CO_2^-)$) [137], ou encore à 1100 cm^{-1} (ν(C–N)) [140]. Toutefois, nous pouvons noter l'absence de signal caractéristique majeur à 850 cm^{-1} traduisant principalement la présence du groupement carboxylate, signal que nous avions pu observer pour les cellules CHO (Figure 3.8b) et HeLa (Figure 3.12a). Le temps d'acquisition obtenu n'est donc pas à être considérée.

2.4.2 Modèle pour la densité TCD

En ce qui concerne le modèle TCD obtenu, nous pouvons une nouvelle fois observer que le modèle obtenu est satisfaisant, avec une erreur RMSEP de $0{,}15 \cdot 10^6$ cell/mL pour une gamme de variation allant de $2{,}04 \cdot 10^6$ cell/mL à $4{,}32 \cdot 10^6$ cell/mL (considérant des spectres Raman acquis sur 5 min tout au long de la culture). Toutefois, manquant de matière pour évaluer les coefficients de régression du modèle de régression PLS (Figure 3.16), nous pouvons étudier le présent modèle par analogie avec les précédents cas des cellules CHO et HeLa.

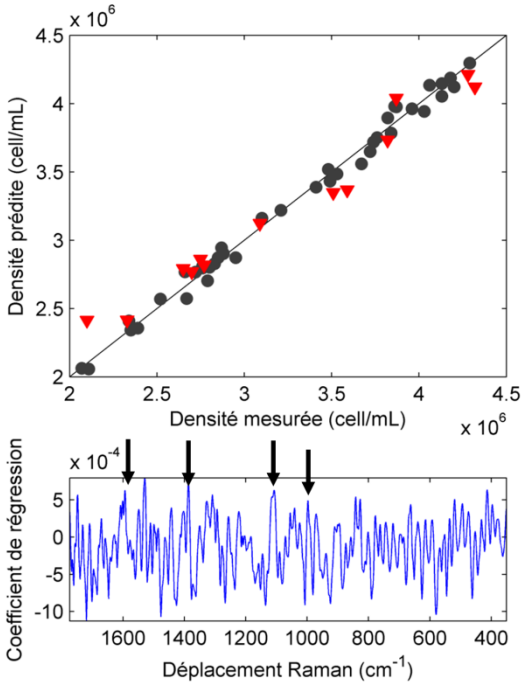

Figure 3.16 Représentation graphique du modèle calculé lors de la détermination du temps d'acquisition pour le paramètre TCD d'une culture de cellules Sf9. Les spectres sont acquis sur des temps d'acquisition de 10 × 30 s. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Ainsi, les bandes situées autour de 1600 cm^{-1} sont attribuées aux déformations $\delta(NH_2)$ des groupements amine des amides secondaires [140] situés dans les cellules. Les signaux autour de 1400 cm^{-1} sont directement assignés aux déformations $\delta(CH_2)$ des chaines

carbonées des acides aminés présents dans les cellules. À 1100 cm^{-1}, nous pouvons attribuer les déformations $\nu(PO_2^-)$ des colonnes des macromolécules d'ADN [145]. Enfin, nous pouvons assigner aux signaux observés à 1000 cm^{-1} les élongations ν(C–C) du cycle benzénique de la molécule de phénylalanine [137,142,145]. Cependant, cette dernière bande a toujours été prédominante pour les autres types de cellules. Le fait qu'ici, elle ne soit pas présente de manière majoritaire sur les coefficients de régression traduit surement la différence qu'il existe entre ces cellules Sf9 et les autres étudiées jusque là. La nature des cellules est belle et bien différente d'une famille à l'autre, c'est pourquoi il faut prendre cette dernière analyse avec précaution, puisque d'une part, il s'agit d'une autre famille de molécule et d'autre part, le modèle repose uniquement sur un seul lot de culture.

En effet, les coefficients de régression calculés pour les différents modèles présentés jusqu'ici, qu'il s'agisse des derniers pour les cellules Sf9, présentant des niveaux de bruit importants en partie seulement dus à des gammes de variation faibles, ou des autres types de cellule étudiés précédemment, nous avons pu voir qu'ils présentaient tous un certain niveau de bruit synonyme de manque de robustesse. Ceci vient du fait que nous travaillions uniquement sur un seul lot de données pour chaque type de cellule.

L'étude des temps d'acquisition est une étape initiale déterminante pour la suite des travaux présentés puisqu'elle conditionne le paramétrage de l'acquisition de tous les spectres à venir. Pour chaque type de cellule (CHO, HeLa, Sf9), nous avons acquis les spectres d'une culture de cellules en continu pour un temps d'acquisition de 1 × 30 s. En théorie, cela pourrait être reproduit pour tous les lots suivants afin de se placer dans les mêmes conditions et faire varier à souhait le temps d'acquisition suivant le métabolite étudié. Toutefois, le volume de données généré par de tels protocoles est trop lourd pour le logiciel iC Raman™. De ce fait, l'acquisition en continu doit être renouvelée plusieurs fois afin de recouvrir toute la durée d'une même culture cellulaire. De plus, le fait d'acquérir uniquement des spectres de 30 s oblige l'utilisateur à cumuler les spectres artificiellement pour chaque métabolite. C'est pourquoi dans nos travaux, nous fixons le temps d'acquisition pour les acquisitions des spectres Raman des cultures suivantes. Dans notre cas, les études ont montré qu'un temps de 10 × 30 s était le plus à même de produire des jeux de données permettant d'accéder aux modèles de régression les plus performants.

Il est toutefois nécessaire de rappeler une nouvelle fois que les modèles de régression PLS construits pour l'étude du temps d'acquisition reposent, pour chaque type de cellule, sur un lot de spectres Raman acquis tout au long d'une seule et unique culture. De ce fait, il est nécessaire par la suite de tester les performances du modèle de régression non pas sur un

lot de spectres test provenant de la même culture, mais sur un ensemble de spectres provenant d'autres cultures et acquis dans les mêmes conditions.

3 Effets du montage de la sonde

Les investigations concernant le montage de la sonde proviennent directement de l'étude des capacités prédictives des modèles de régression construits jusqu'ici, ces derniers reposant sur une seule et unique culture. L'acquisition de données provenant d'une nouvelle culture permet alors d'étudier les réelles capacités des modèles par rapport à des échantillons complètement extérieurs.

3.1 Biais existant entre deux lots

Dans cette partie, nous reprendrons les travaux concernant les cellules CHO afin d'exposer les observations réalisées pour tous les types de cellule employés. Ainsi, une première culture (nous parlerons aussi de *batch*) a été réalisée pour l'étude du temps d'acquisition, étude conclue par le choix d'un temps de 5 min (10 × 30 s) pour l'acquisition des spectres Raman.

Nous avons repris l'intégralité des données fournies par la première culture, soit les 63 mesures de références et leurs spectres associés. Les spectres sont prétraités de la même façon que pour les travaux précédents, soit dérivés (algorithme de Savitzky-Golay, fenêtre de 15 points et polynôme d'ajustement d'ordre 2) puis restreints au domaine spectral compris entre 1775–350 cm^{-1}. Les données acquises pour cette première culture constituent alors l'ensemble de calibration pour les nouveaux modèles de régression.

Les paramètres de la seconde culture sont similaires à ceux de la première. La culture de cellules CHO a duré 336 h (soit 14 jours) et un total de 56 prélèvements a été réalisé pour les analyses chimiométriques. Tout au long de la culture, les spectres Raman *in situ* ont été acquis de la même façon que la première culture : en continu des spectres de 1 × 30 s. Bien qu'un temps d'acquisition de 5 min ait été déterminé pour la suite des travaux, le fait de conserver ce format d'acquisition pour ce deuxième *batch* permet de se positionner exactement dans les mêmes dispositions que pour la première culture et ainsi éviter de générer d'éventuelles variations d'un *batch* à l'autre. Nous ne travaillerons néanmoins que sur des temps d'acquisition de 10 × 30 s. Toutes ces données acquises pour la deuxième culture constituent l'ensemble test pour vérifier que le modèle de régression établi sur le premier batch peut être exporté à d'autres cultures de cellules CHO.

Afin de simplifier l'analyse, nous représenterons uniquement le biomarqueur VCD car ce dernier traduit parfaitement les observations réalisées pour tous les différents paramètres

biochimiques. Ainsi, les résultats obtenus quant aux prédictions des données de la seconde culture par un modèle de régression PLS construit sur le premier batch sont représentés Figure 3.17. Celle-ci comprend d'une part le graphique des densités cellulaires prédites en fonction des densités mesurées, mais aussi un aperçu des résidus (différences entre les densités mesurées et les densités prédites) pour le graphique précédemment évoqué.

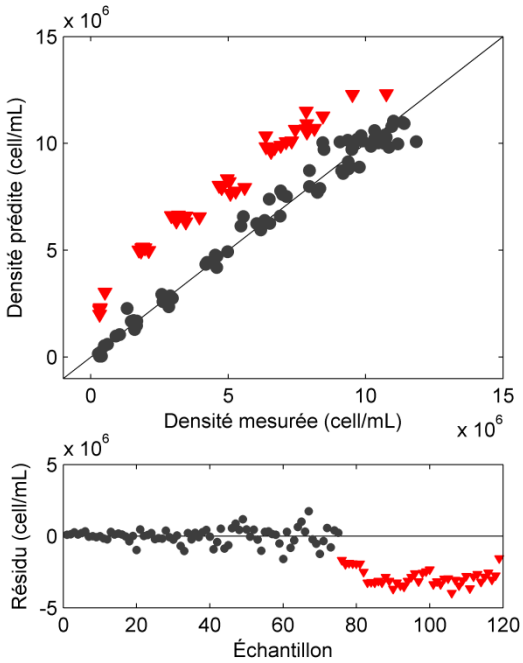

Figure 3.17 Représentation des résultats obtenus lors de la prédiction des densités en cellules vivantes des échantillons du nouveau batch (triangles rouges) par le modèle de régression VCD basé sur la première culture (points noirs). Le graphique supérieur reprend les densités prédites en fonction des densités mesurées tandis que le graphique inférieur représente les résidus (densité mesurée – densité prédite) pour chaque individu.

Dans un premier temps, le modèle de régression construit sur les 63 échantillons du premier *batch* a été validé par une validation croisée qui a permis de sélectionner 3 variables latentes pour le modèle. L'erreur RMSECV a été évaluée à $0{,}64 \cdot 10^6$ cell/mL, ce qui reste tout à fait cohérent vis-à-vis des résultats obtenus lors de la détermination du temps d'acquisition (nous rappelons que nous avions obtenu une erreur RMSEP de $0{,}52 \cdot 10^6$ cell/mL, Tableau 3.1, page 89).

Les prédictions des échantillons test (seconde culture) sont représentées par des triangles rouges sur la Figure 3.17. Un biais de prédiction apparaît alors de manière évidente. En effet, l'erreur RMSEP est évaluée à $2{,}84 \cdot 10^6$ cell/mL, ce qui est largement supérieur aux erreurs calculées jusque là. Ce biais, évalué à $-2{,}78 \cdot 10^6$ cell/mL, apparaît de façon très nette sur le graphique des densités prédites en fonction des densités mesurées. Nous pouvons voir que toutes les densités cellulaires des échantillons test sont surestimées. Ceci est d'autant plus visible sur le graphique inférieur où tous les résidus calculés pour les échantillons test présentent des valeurs négatives. Il est alors évident que non seulement le

modèle de régression développé sur la base des données de la première culture de cellules n'est pas assez robuste pour prédire de façon convenable les données d'une nouvelle culture, mais en plus il apparaît de manière évidente qu'il existe un certain biais systématique entre deux cultures cellulaires.

En faisant une étude ACP des deux lots de données, nous pouvons tenter de déterminer la provenance du biais apparent. Les résultats de cette analyse ACP sont représentés Figure 3.18. Le graphique des *scores* permet de montrer que la première composante principale traduit simplement l'évolution du système au cours du temps tandis que la deuxième démontre clairement une différence entre les lots de données provenant des deux cultures.

En regardant les *loadings* de la première composante principale obtenue (64,12 % de variance du jeu de données), nous pouvons voir que cette dernière est soutenue principalement par les bandes spectrales autour de 1000 cm^{-1}. Hors, nous avions pu montrer que cette bande caractérise principalement la respiration du noyau benzénique des molécules de phénylalanine présentes dans les cellules. Ainsi, l'évolution observée sur le graphique des *scores* pour la composante PC1 traduit l'évolution du bioprocédé à travers notamment la multiplication cellulaire.

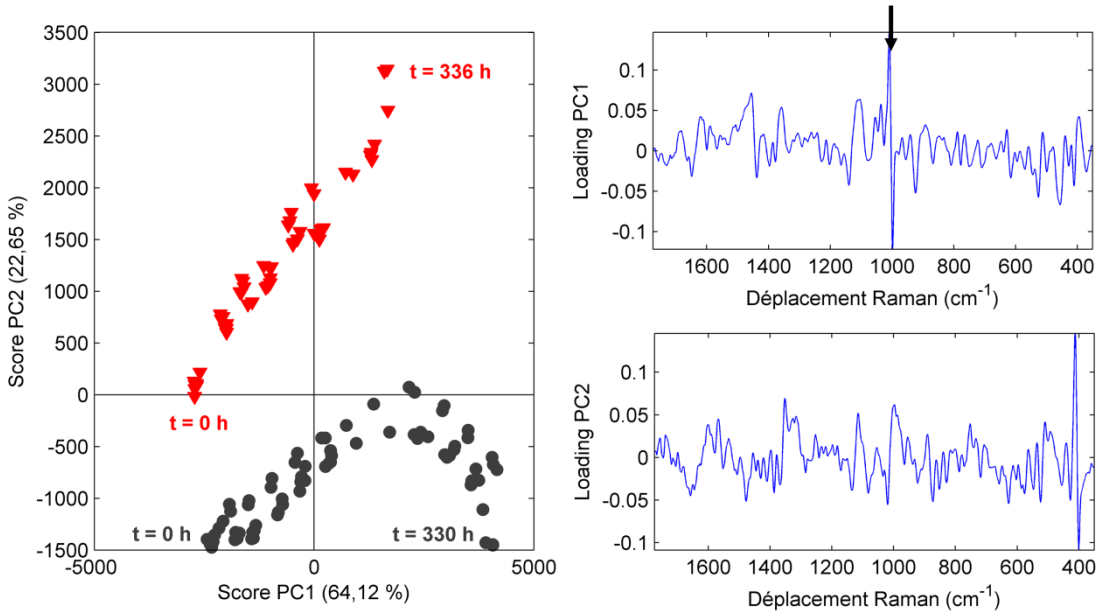

Figure 3.18 Résultats principaux obtenus via l'analyse en composantes principales réalisée sur le jeu de données combinant les échantillons des deux cultures de cellules CHO. À gauche, le graphique des *scores* des deux premières composantes principales (86,77 % de la variabilité totale). Les points noirs représentent le premier batch, les triangles rouges le second. À droite, les *loadings* des mêmes composantes principales.

La deuxième composante principale (22,65 % de la variance totale) est quant à elle majoritairement responsable de la séparation entre les spectres provenant des deux

cultures. En interprétant l'allure des *loadings* de cette composante, nous pouvons constater une similarité avec les spectres utilisés. La Figure 3.19 confronte les *loadings* de la composante PC2 et les spectres mis en jeu. Ainsi, nous pouvons voir que les bandes principales des *loadings* de la composante PC2 correspondent aux signaux les plus intenses des spectres du jeu de données mêlant les deux cultures. Ceci souligne donc des différences d'intensité spectrale qu'il peut y avoir entre les spectres de deux cultures similaires, ce qui se confirme en prenant uniquement les acquisitions spectrales, Figure 3.19. Nous pouvons remarquer que les signaux du premier batch (en noir) présentent des bandes plus intenses que ceux du second batch (en rouge).

Outre la variabilité intrinsèque qui existe entre deux cultures de cellules, nous pouvons donc voir qu'il existe des différences dues à l'acquisition des spectres. Nous proposons donc d'étudier le système d'acquisition Raman disposé sur le bioréacteur de culture et de déterminer les causes possibles de ces différences d'intensité.

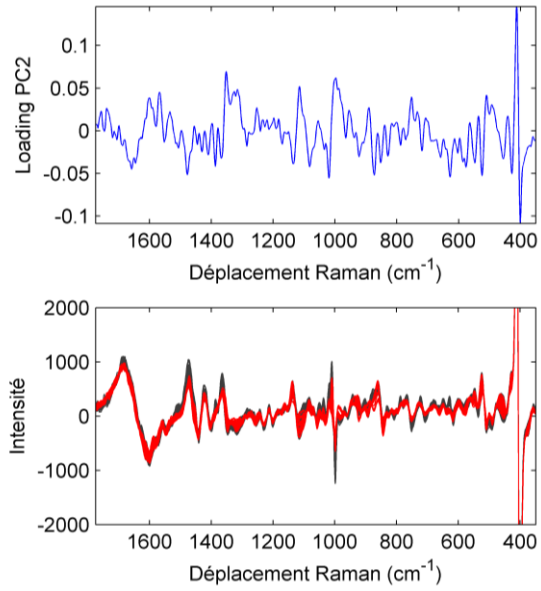

Figure 3.19 Comparaison entre les *loadings* de la composante PC2 (en haut) et les spectres dérivés des deux cultures de cellules CHO (en noir la première, en rouge la seconde).

3.2 Influence de l'installation de la sonde Raman

Le montage pour le suivi de la culture cellulaire à l'aide de la spectroscopie Raman est représenté Figure 3.1, page 82. La sonde à immersion est installée directement sur la platine du bioréacteur de culture et reliée à l'aide d'une fibre optique au spectromètre. La Figure 3.20 permet d'avoir une vue du montage de la sonde sur un bioréacteur de 7 L. Nous pouvons remarquer sur cette figure que le bioréacteur de culture est enveloppé dans du papier aluminium afin d'empêcher la lumière extérieure du laboratoire de venir pénétrer au sein du bioréacteur (dont les parois sont en verre). En effet, ces rayonnements présentent

des bandes caractéristiques dans le domaine du visible et se retrouvent inévitablement sur les spectres Raman. Ces contributions risquent ainsi de dégrader les données spectrales et donc l'analyse chimiométrique.

De plus, il est important de noter ici que les bioréacteurs subissent entre chaque culture une phase de stérilisation entre les cultures afin d'éviter que n'importe quelle contamination apparaisse ou soit transmise de la culture précédente. Il en va de même pour la partie immergée de la sonde Raman puisque celle-ci est en contact avec le milieu de culture. De ce fait, entre chaque culture, la sonde est démontée puis remontée.

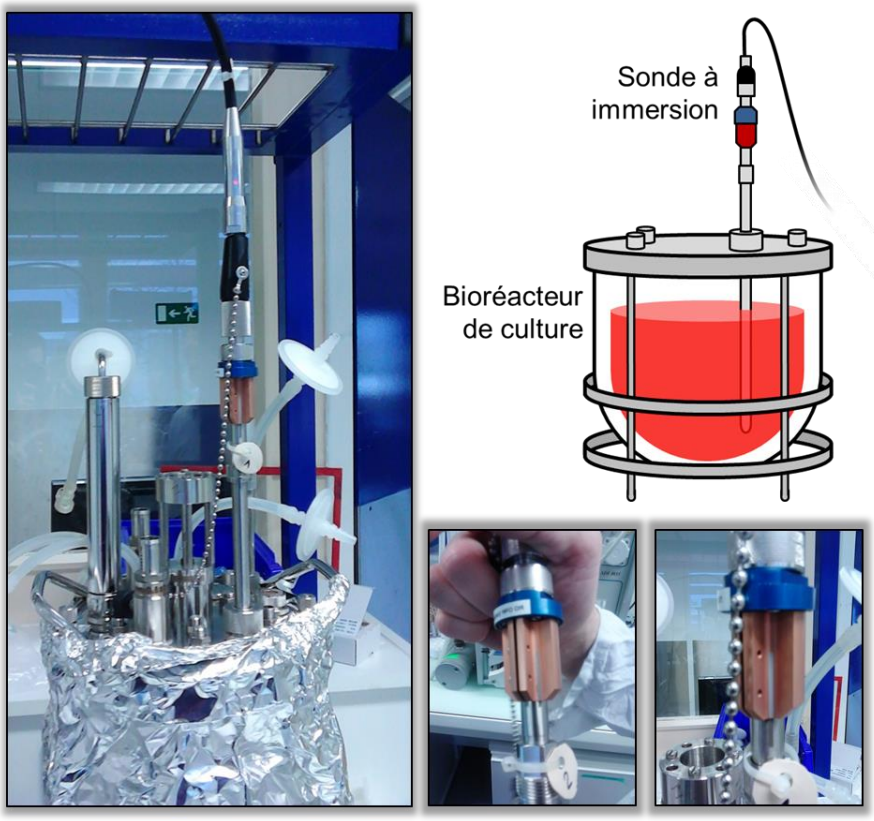

Figure 3.20 Représentation du montage de la sonde à immersion sur le bioréacteur de culture. Les vues rapprochées permettent de mettre en avant le repère gris pour l'alignement des fibres optiques de la tête et du corps de la sonde Raman.

Nous pouvons alors lister plusieurs points susceptibles de générer des variations d'une culture à l'autre en ce qui concerne l'installation de la sonde. Tout d'abord, la première chose à noter est la hauteur de sonde durant le montage de celle-ci sur le bioréacteur. En effet, aucun repère n'est présent pour fixer la distance à laquelle la sonde est immergée. De ce fait, il convient d'étudier les variations de profondeur afin d'évaluer si ce paramètre a un effet sur les mesures Raman. Ensuite, nous pouvons vérifier l'alignement des fibres optiques entre la tête de sonde (reliée au spectromètre) et le corps de sonde (partie immergée). Il existe sur la sonde un repère gris (Figure 3.20) pour parfaire l'alignement. Cette procédure très sommaire ne permet certainement pas un alignement systématique de la sonde entre

différents suivis cellulaires. Enfin, nous pouvons également évaluer l'impact de la position de la fibre optique reliant la sonde au spectromètre sur les mesures, celle-ci étant longue de près de 5 m afin d'assurer une longue distance dans l'éventualité où le bioréacteur et le spectromètre ne pourraient être disposés conjointement.

3.2.1 Influence du positionnement de la tête de sonde

Pour l'étude des effets potentiellement dus au vissage de la sonde, le déploiement de plusieurs lots de cellules n'était pas envisageable par souci financier. Ainsi, nous avons mis en place un protocole particulier.

L'idée est de simuler plusieurs fois l'installation de la sonde Raman sur le bioréacteur, tel que lors de la préparation d'une culture suite à la stérilisation du bioréacteur. Cependant, dans ce cas précis, l'objet de l'étude ne varie pas et reste le même du début à la fin des travaux : du milieu de culture sans cellule, n'évoluant potentiellement pas. Ainsi, nous avons démarré l'acquisition de spectres Raman en continu sur une période longue et indéterminée. Quatre opérateurs différents ont alors réalisé à tour de rôle la même opération : le démontage de la tête de sonde, puis la réinstallation de cette dernière pour reproduire l'installation du dispositif d'acquisition Raman sur une nouvelle culture. Chaque opérateur a répété trois fois la manipulation, prenant soin de laisser suffisamment de temps entre chaque épreuve pour qu'un nombre satisfaisant de spectres soit acquis une fois le système stabilisé.

Suite à cela, seulement quelques spectres de chaque plage d'acquisition furent conservés (10 spectres par plage) pour l'analyse chimiométrique, soit un total de 120 spectres Raman (4 opérateurs × 3 répétitions × 10 spectres). Les spectres ont été traités de la même façon que pour les travaux précédents : une dérivée première suivie de la sélection de la région spectrale 1775–350 cm^{-1}. Nous réalisons ensuite l'analyse ACP sur le jeu de données comprenant les 120 spectres. Les résultats sont représentés Figure 3.21.

Nous pouvons constater que les deux premières composantes (celles qui expriment le maximum de variabilité) ne représentent que 6,27 % et 3,29 %. Ceci vient du fait que nous observions les variations existantes entre les spectres d'un même objet (ici le milieu de culture). Toutefois, nous pouvons voir que la première composante principale permet en partie de séparer les manipulations effectuées par les différents opérateurs (nous ne distinguons pas les trois répétitions sur la représentation). Les installations des opérateurs #1 et #2 sont séparées du reste de l'ensemble, en particulier la première répétition de l'opérateur #1. Les *loadings* de la composante PC1, responsable de la séparation, sont représentés Figure 3.21 et comparés au spectre moyen de toutes les acquisitions (considéré alors comme le spectre du milieu de culture). Nous pouvons alors voir qu'il existe plusieurs

similarités dans les allures des deux représentations, notamment autour de 400 cm^{-1}, traduisant alors un effet dû à l'intensité des spectres. En effet, cet important signal est assigné à la fenêtre en saphir de la sonde à immersion. Nous montrons alors que le montage de la sonde en début de culture impacte l'acquisition des spectres Raman d'une culture à l'autre.

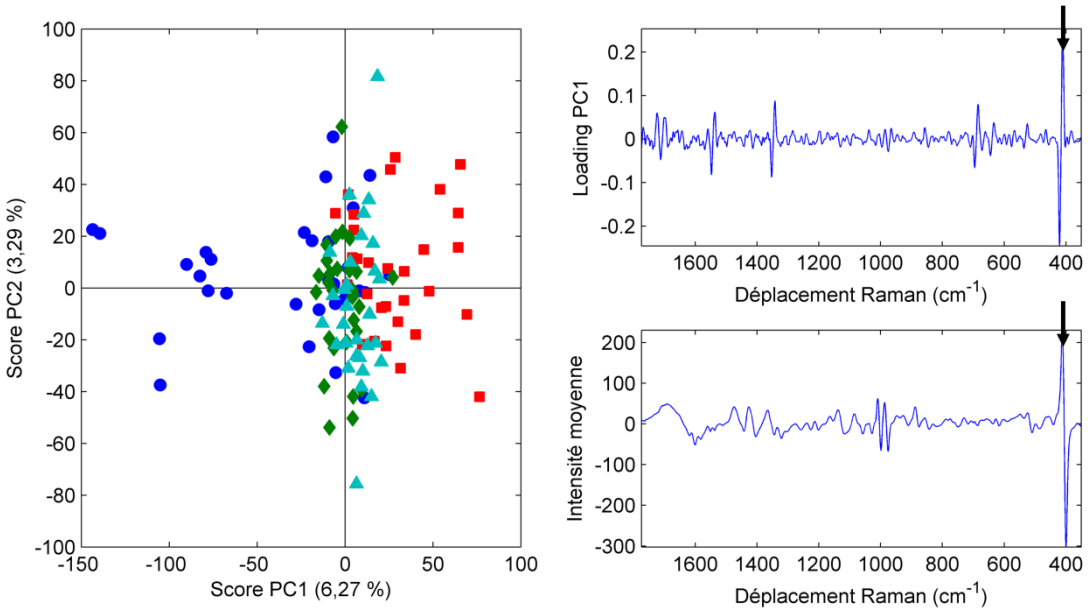

Figure 3.21 Représentation du modèle ACP réalisé sur les 120 spectres acquis pour l'étude du montage de la tête de sonde. Le prétraitement des spectres est une dérivée première, puis sélection de la région spectrale 1775–350 cm^{-1}. Sur le graphique des *scores*, les points bleus représentent l'opérateur #1, les carrés rouges l'opérateur #2, les losanges verts l'opérateur #3 et les triangles cyans l'opérateur #4.

Ne pouvant agir au niveau instrumental sur cet appareil commercial, nous avons cherché à améliorer le traitement spectral afin de minimiser, voire effacer cet effet. La première modification concerne le prétraitement des spectres. Puisque nous pouvons observer des effets d'intensité sur certaines bandes, il convient donc d'appliquer une normalisation sur les spectres Raman. Cependant, une normalisation par rapport à une bande en particulier n'est pas possible puisque l'intégralité des bandes spectrales tend à évoluer au cours du temps. Il en est de même pour la bande spectrale à 400 cm^{-1}. Bien que celle-ci représente principalement la fenêtre en saphir de la sonde, la turbidité croissante du milieu peut amener le signal à varier légèrement, rendant donc la normalisation par rapport à cette bande impossible. Dans la littérature, la plupart des travaux exploitant plusieurs cultures appliquent le prétraitement SNV en plus d'une correction de ligne de base [132-136]. Nous ajoutons alors la normalisation SNV pour tenter de corriger les différences observées à cause du montage de la sonde. Cependant, cette normalisation ne permettait pas de diminuer la séparation observée sur le graphique des *scores*, telle que nous pouvions voir Figure 3.21.

Nous avons alors cherché à modifier le domaine spectral sur lequel nous travaillions. Ainsi, nous avons ajouté la région spectrale comprise entre 3000–2800 cm^{-1} à l'ensemble. Cette région est très proche des bandes principales des molécules d'eau du milieu situées au-delà de 3000 cm^{-1} (déformations intramoléculaires et élongations asymétriques des molécules). Toutefois, la région 3000–2800 cm^{-1} traduit la présence des bandes ν(C–H) caractérisant plusieurs molécules présentes dans le milieu dont principalement le glucose [146]. De ce fait, nous avons appliqué le nouveau prétraitement (dérivée première puis normalisation SNV des spectres, puis sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}) à l'ensemble des 120 spectres simulant 120 différentes installations de la sonde. Les résultats de l'étude ACP sont représentés Figure 3.22.

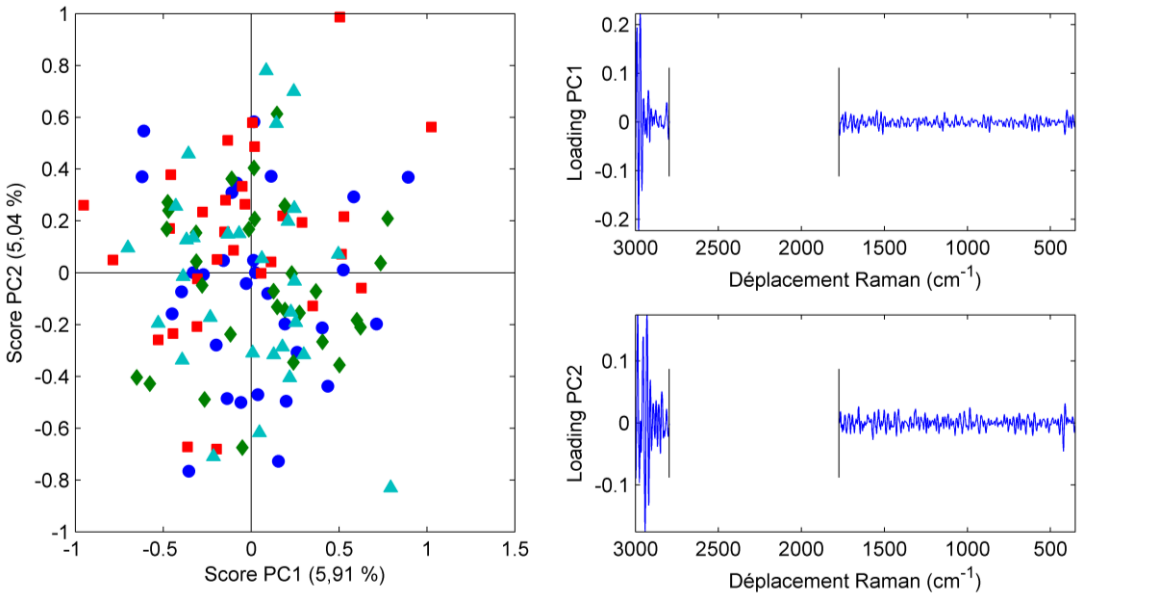

Figure 3.22 Représentation du modèle ACP réalisé sur les 120 spectres acquis pour l'étude du montage de la tête de sonde. Le prétraitement des spectres est une dérivée première suivie d'une normalisation SNV, puis sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Sur le graphique des *scores* (à gauche), les points bleus représentent l'opérateur #1, les carrés rouges l'opérateur #2, les losanges verts l'opérateur #3 et les triangles cyans l'opérateur #4.

Nous pouvons clairement observer sur le graphique des *scores* que la séparation des manipulations des différents opérateurs a été éliminée par le nouveau prétraitement des spectres. Nous pouvons notamment remarquer que même la première répétition de l'opérateur #1 n'est plus discriminée par les premières composantes de l'analyse ACP. À travers le manque de structure des *loadings* de l'analyse ACP, nous pouvons conclure que ce prétraitement permet de supprimer les variations dues au montage de la sonde sur le bioréacteur, et plus précisément l'alignement des fibres optiques de la tête de sonde et du corps immergé. Même si une solution chimiométrique a été trouvée, nous avons demandé au constructeur du spectromètre Kaiser Optical System Inc. de travailler sur une solution instrumentale permettant d'assurer un réalignement parfait de la sonde.

3.2.2 Influence de la profondeur d'immersion

La profondeur d'immersion de la sonde est un autre paramètre à contrôler pour l'analyse. Cette profondeur peut varier lors de l'assemblage de la partie immergée de la sonde (corps de la sonde, ou encore pointe de la sonde sur la Figure 2.4, page 59) sur la platine du bioréacteur, avant de connecter la tête de sonde reliée au spectromètre. Dans le même esprit que pour l'analyse précédente, nous avons démarré l'acquisition de spectres Raman du milieu de culture, sans cellule, et en continu. Cinq profondeurs de sonde ont été étudiées pour cette étude, représentées Figure 3.23.

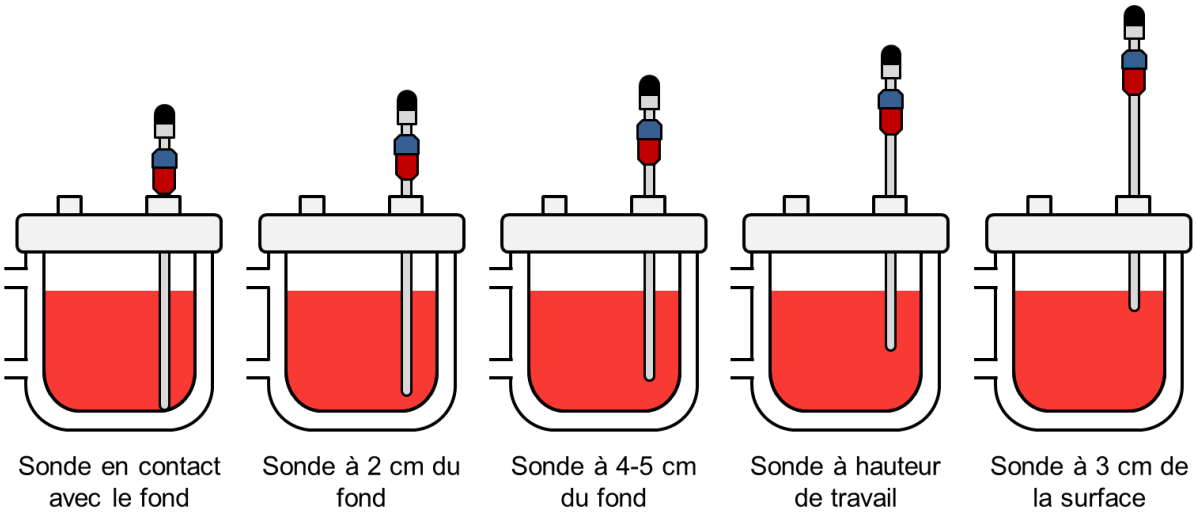

Figure 3.23 Représentation des différentes profondeurs étudiées pour les travaux sur l'installation de la sonde. La hauteur de travail correspond à une immersion d'une dizaine de centimètres de la sonde dans le milieu de culture.

Nous conservons finalement 10 spectres pour chacune des positions de la sonde, ce qui revient à un jeu de données de 50 spectres Raman (5 positions × 10 spectres). Avant de prétraiter les spectres, nous représentons les spectres bruts moyens de chacune des profondeurs à l'étude, Figure 3.24a. Nous pouvons alors remarquer l'apparition de bandes spectrales lorsque la fenêtre de la sonde est en contact avec le fond du bioréacteur entre 2000 cm^{-1} et 1000 cm^{-1}, ainsi qu'à partir de 3000 cm^{-1}. Toutefois, en appliquant le prétraitement déterminé suite à l'étude du vissage de la tête de sonde (dérivée première puis normalisation SNV, sélection spectrale 3000–2800 cm^{-1} et 1775–350 cm^{-1}), Figure 3.24b-c, nous pouvons voir que les variations dues à la profondeur de sonde interviennent non seulement dans les régions citées, mais aussi à 400 cm^{-1} et 2900 cm^{-1}.

De plus, sans prendre en compte le spectre moyen lorsque la sonde est en contact du fond de la cuve, nous pouvons remarquer que certaines bandes traduisent les variations de profondeur de sonde (Figure 3.24d), à 2880 cm^{-1}, 1550 cm^{-1}, autour de 1250 cm^{-1} et à 400 cm^{-1}. Étant donné l'écart-type obtenu en prenant en compte le spectre possédant les bandes

caractérisant le fond de la cuve (Figure 3.24c), nous pouvons remarquer que les facteurs de variation sont relativement similaires. Nous pouvons donc déceler un effet dû au fond du bioréacteur sur les spectres bien que ceux-ci ne soient pas en contact direct.

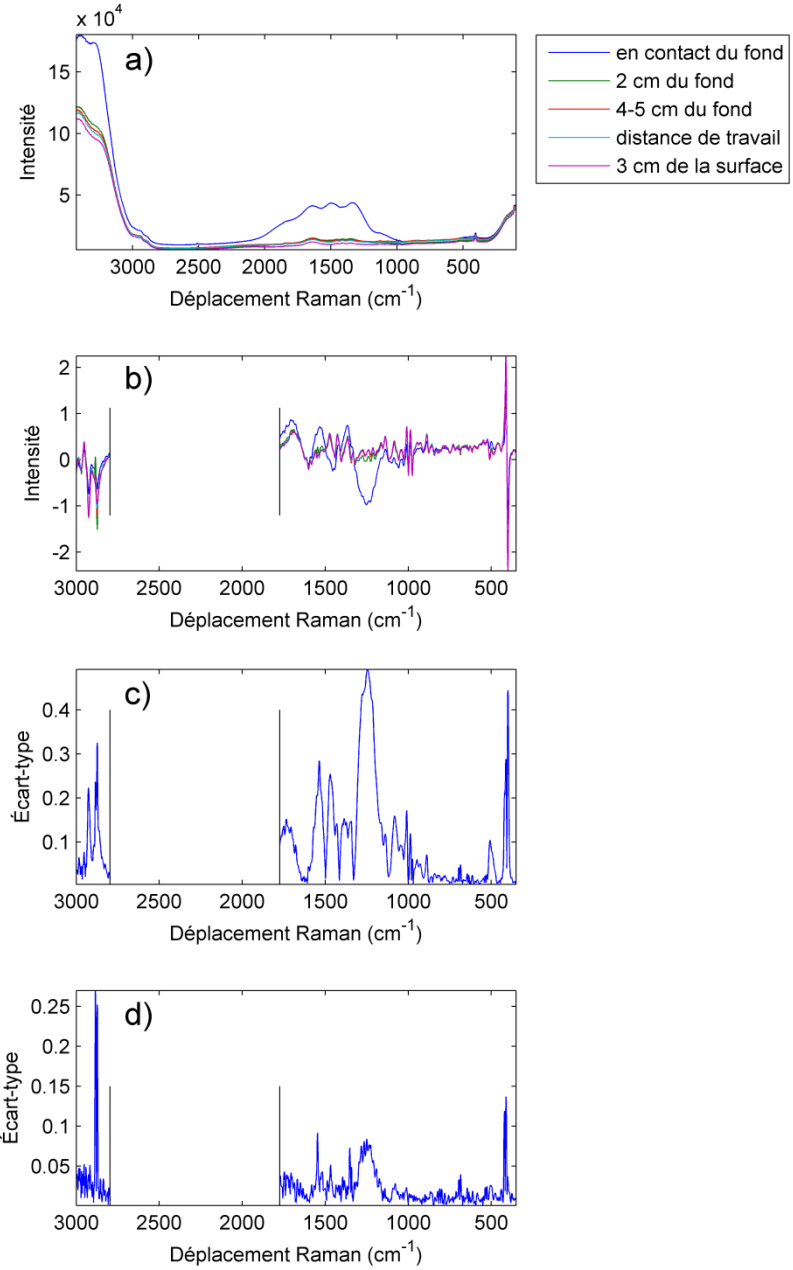

Figure 3.24 Représentations des spectres obtenus pour l'étude de la hauteur de sonde, dont a) les spectres bruts moyens pour chacune des positions, b) les spectres moyens prétraités (dérivée première puis normalisation SNV, sélection 3000–2800 cm^{-1} et 1775–350 cm^{-1}), c) l'écart-type des spectres moyens prétraités comprenant toutes les profondeurs et d) l'écart-type des spectres moyens prétraités en excluant le groupe où la sonde est en contact avec le fond du bioréacteur.

L'étude ACP est réalisée sur seulement 40 spectres Raman, excluant les 10 spectres acquis lorsque la sonde est en contact avec le fond, très différents d'un point de vue spectral. Bien que ce groupe de spectres soit écarté, nous pouvons voir Figure 3.25 que les *scores* de la première composante principale de l'analyse ACP, exprimant 29,07 % de la

variabilité totale, traduisent vraisemblablement les différences de hauteur. En effet, dans la direction des *scores* de PC1 croissants, nous pouvons voir que la hauteur de sonde diminue. Les *loadings* de la composante principale PC1 permettent de montrer que celle-ci est soutenue par les variations que nous observions sur l'écart-type des spectres moyens, Figure 3.24d. La deuxième composante principale (5,75 % de variabilité) ne traduit quant à elle aucune tendance particulière du milieu de culture, à en juger par ses *loadings*.

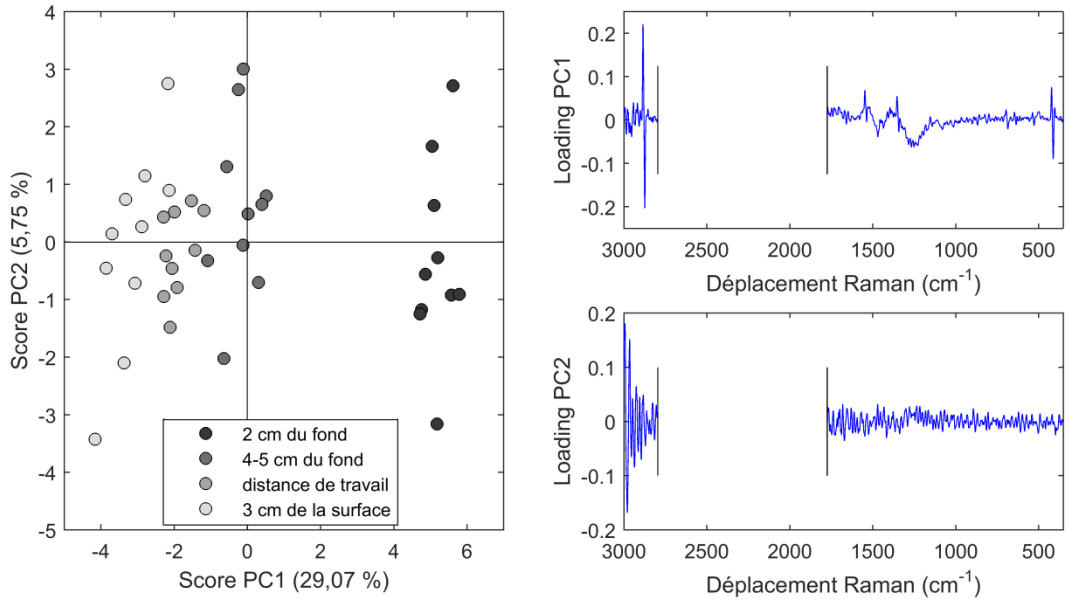

Figure 3.25 Représentation du modèle ACP calculé pour l'analyse des 40 spectres Raman acquis pour l'étude de la hauteur de sonde. Le prétraitement des spectres est une dérivée première suivie d'une normalisation SNV, puis sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Sur le graphique des *scores* (à gauche), les différentes hauteurs de sonde sont dégradées des plus claires (proches de la surface) aux plus sombres (proches du fond du bioréacteur).

Toutefois, nous pouvons noter sur la Figure 3.25 que les *scores* des spectres acquis à 2 cm du fond sont séparés du reste des acquisitions. En effet, nous observons plus de séparation entre les *scores* des spectres acquis à 2 cm du fond et 4–5 cm qu'entre les *scores* des spectres acquis à 4–5 cm et ceux acquis à 3 cm de la surface du milieu de culture. La première composante principale exprime donc bien un effet de hauteur de sonde mais pas de façon linéaire. De ce fait, nous pouvons considérer que le fait de travailler en plongeant la sonde à une distance suffisamment éloignée du fond du bioréacteur permet de prévenir tout signal susceptible de perturber la mesure Raman. En respectant ces conditions et en tachant de fixer la sonde de manière similaire pour chaque nouvelle culture, l'effet induit par la profondeur de sonde est alors négligeable. Nous ne préconisons néanmoins pas de hauteur optimale malgré l'influence de la profondeur d'immersion de la sonde.

3.2.3 Influence de la disposition du câble de fibre optique

La position de la fibre optique est un paramètre prédisposé à varier. En effet, la fibre reliant la sonde au spectromètre mesure 5 m sans être fixée. Tout mouvement du bioréacteur tend donc à faire varier la position de la fibre optique d'un suivi de bioprocédé à l'autre puisque la cuve est toujours déplacée entre deux cultures pour stérilisation.

Nous avons donc déplacé les 5 m de fibre optique de manière aléatoire pendant l'acquisition de spectres de milieu de culture en continu, tout en prenant soin de ne pas revenir sur une position similaire à ce qui avait déjà pu être réalisé. Nous avons ainsi étudié trois différentes situations de la même façon que pour l'étude du vissage de la tête de sonde ou de sa profondeur dans le bioréacteur. Ainsi, nous avons conservé 10 spectres représentatifs de chaque position, que nous avons prétraités dans le même esprit : dérivée première de Savitzky-Golay suivi d'une normalisation SNV, puis sélection des domaines spectraux 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Nous réalisons ensuite l'analyse en composantes principales sur le jeu de données comprenant les 30 spectres Raman prétraités (3 positions × 10 spectres). Les résultats sont présentés Figure 3.26.

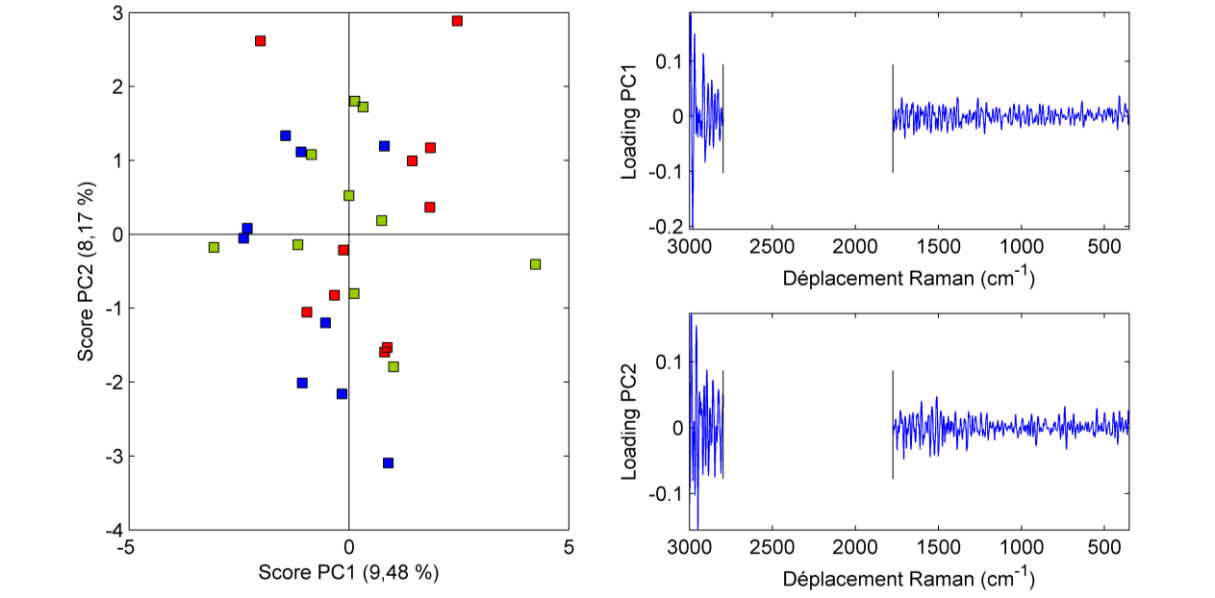

Figure 3.26 Représentation de l'analyse ACP réalisée sur les 30 spectres acquis pour l'étude de la position de la fibre optique. Le prétraitement des spectres est une dérivée première suivie d'une normalisation SNV, puis sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Sur le graphique des *scores* (à gauche), chaque position est représentée par une couleur particulière (rouge, bleu ou vert).

Nous pouvons voir sur le graphique des *scores* que nous ne sommes pas en mesure de distinguer de groupes particuliers. De plus, nous pouvons remarquer sur les *loadings* des deux premières composantes (à droite Figure 3.26) que ces derniers ne présentent pas d'allure particulière. Ceci indique qu'en appliquant le prétraitement déterminé suite aux

travaux sur le montage de la tête de sonde, nous ne discernons pas d'influence de la position de la fibre optique reliant le spectromètre Raman et la sonde immergée dans le bioréacteur de culture.

3.3 Application du nouveau traitement de données spectrales

Suite aux différentes études réalisées sur l'appareillage Raman, nous avons pu montrer que nous étions en mesure, grâce à un nouveau prétraitement, de rendre négligeable voire d'éliminer les phénomènes de fluorescence ou les variations spectrales dues au montage de la sonde. Afin de vérifier les performances de cette nouvelle correction, nous reprenons les données concernant le paramètre biochimique VCD pour les cultures de cellules CHO employées section 3.1 (Biais existant entre deux lots) afin de voir si nous pouvons réduire le biais observé sur les prédictions du nouveau lot provenant d'une seconde culture cellulaire. Nous rappelons que le modèle de calibration déterminé à partir de 63 échantillons d'une première culture était obtenu sur la base de 3 variables latentes, dont l'erreur RMSECV, calculée à l'aide d'une validation croisée initiale, atteignait $0{,}64 \cdot 10^6$ cell/mL. L'erreur RMSEP avait alors été évaluée à $2{,}84 \cdot 10^6$ cell/mL en appliquant le modèle de régression sur les 56 échantillons d'une seconde culture de cellules, présentant un biais de prédiction systématique évalué à $-2{,}78 \cdot 10^6$ cell/mL. Les résultats obtenus en appliquant le nouveau traitement de données sont représentés Figure 3.27.

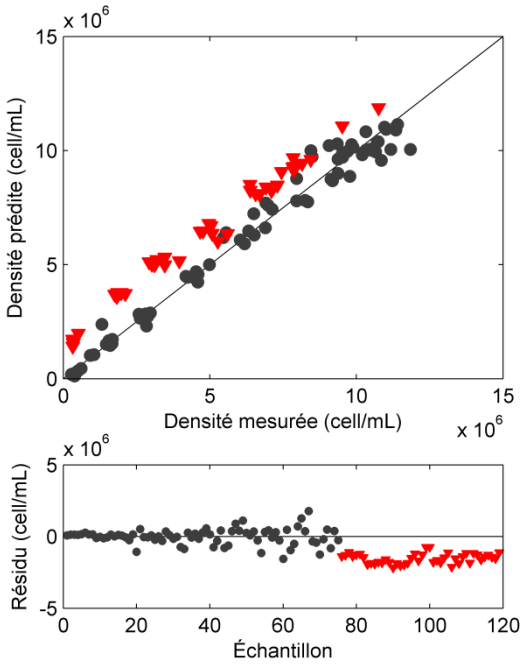

Figure 3.27 Représentation des résultats obtenus lors de la prédiction des densités en cellules vivantes des échantillons du nouveau *batch* (triangles rouges) par le modèle de régression VCD basé sur la première culture (points noirs) après avoir appliqué le nouveau prétraitement sur les spectres Raman. Le graphique supérieur reprend les densités prédites en fonction des densités mesurées tandis que le graphique inférieur représente les résidus (densité mesurée – densité prédite) pour chaque individu.

Nous obtenons un nouveau modèle de régression PLS basé sur 3 variables latentes dont l'erreur RMSECV est estimée à $0{,}63 \cdot 10^6$ cell/mL. Le coefficient de détermination R^2 (de 0,84 pour la Figure 3.17) atteint 0,94 pour cette nouvelle analyse. Toutefois, c'est le graphique des densités prédites en fonction des densités de référence qui permet de remarquer les changements les plus notables pour les prédictions. En effet, le biais observé sur la Figure 3.17 est grandement réduit par l'application du nouveau prétraitement (ici, le biais est évalué à $1{,}51 \cdot 10^6$ cell/mL). D'après les calculs, l'erreur RMSEP est estimée à $1{,}54 \cdot 10^6$ cell/mL (contre $2{,}84 \cdot 10^6$ cell/mL pour le prétraitement précédent, une dérivée première et sélection de la région spectrale 1775–350 cm^{-1}).

Toutefois, bien que ce prétraitement permette de réduire la variabilité qui existe entre deux *batches* à l'aide de corrections mathématiques des variations physiques de l'installation du dispositif Raman, il subsiste toujours un biais non négligeable sur les prédictions obtenues pour la seconde culture. Nous pouvons observer ce phénomène sur le graphique des résidus proposé Figure 3.27. Le constructeur et fournisseur du spectromètre travaille actuellement sur l'amélioration du réglage de la sonde à immersion.

4 Conclusions sur le paramétrage du suivi de cultures cellulaires

Ce chapitre présente plusieurs points clés pour la mise en place du suivi de cultures cellulaires à l'aide de la spectroscopie Raman *in situ*. L'idée principale réside dans la recherche de paramètres optimisés pour réaliser l'acquisition et le traitement des données spectrales afin d'obtenir *in fine* des modèles de régression performants pour les suivis des cultures.

Suite à la mise en place du dispositif Raman, nous avons tout d'abord étudié les caractéristiques d'acquisition des spectres pour chaque biomarqueur. Puisqu'à terme, il s'agit d'un suivi de culture (cellulaire ou virale), il est nécessaire d'être capable de faire des acquisitions en temps réel tout au long de la culture tout en conservant toute l'information chimique de l'état du bioréacteur au sein de l'acquisition spectrale. De ce fait, nous avons élaboré une stratégie d'acquisition permettant de rechercher le temps d'acquisition le plus court possible tout en gardant une signature spectrale de qualité. D'une part, nous avons déterminé le temps d'exposition du signal Raman au détecteur CCD du spectromètre à 30 s. D'autre part, nous avons étudié le nombre d'accumulations nécessaire à chaque type de cellule (CHO, HeLa, Sf9) pour obtenir les modèles de régression les plus performants au niveau prédictif dans le cas de travaux sur culture unique. D'après les erreurs RMSEP déterminées, il ressort finalement qu'un temps d'acquisition total de 5 min (soit 10 × 30 s) est

le plus performant pour les suivis de cultures. Toutefois, nous rappelons bien que ces travaux ont été réalisés sur des lots de données provenant d'une seule culture pour chaque type de cellule. Il convient donc par la suite d'étudier la faculté des modèles à prédire les niveaux de concentration ou de densité de chaque biomarqueur sur de nouvelles cultures. Nous avons ainsi mis en avant l'existence d'un biais sur les prédictions réalisées pour une nouvelle culture. L'analyse du dispositif d'acquisition a donc été engagée à travers l'étude de plusieurs points clés du matériel.

En premier lieu, nous avons évalué la robustesse de l'alignement des fibres optiques de la tête de sonde et du corps immergé. En effet, le corps de la sonde est stérilisé de la même façon que les bioréacteurs entre deux cultures. Nous avons donc simulé plusieurs montages de la sonde et réalisé, à l'aide d'analyses en composantes principales, que cette installation induisait des changements dans les signaux spectraux d'une manipulation à l'autre. Finalement, en apportant une normalisation SNV au prétraitement précédemment entrepris (une dérivée première) et en incorporant une nouvelle région spectrale à la sélection réalisée, nous pouvons très nettement diminuer les variations dues au montage de la tête de sonde. Ensuite, nous avons analysé la profondeur d'immersion de la sonde Raman. Nous avons conservé le même prétraitement pour les corrections spectrales. Toutefois, cette étude a montré qu'il existait bien un impact de ce paramètre sur les spectres Raman, notamment à cause des signaux provenant du fond en verre du bioréacteur de culture. Néanmoins, cet effet est considérablement réduit en travaillant à une distance suffisamment éloignée (au moins 4–5 cm) du fond. Il est donc nécessaire de vérifier à rester dans des conditions similaires pour toutes les cultures lors de l'installation de la sonde sur la platine du bioréacteur. Enfin, nous avons étudié l'impact des mouvements de la longue fibre optique reliant la sonde au spectromètre. Nous avons finalement montré que ces déplacements n'avaient pas d'effets particuliers, ce qui permet de travailler en déplaçant librement le système d'acquisition.

Bien que nous ayons montré que le nouveau prétraitement mis en place (dérivée première de Savitzky-Golay, fenêtre de 15 points et polynôme d'ordre 2, suivi d'une normalisation SNV puis sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}) était capable de minimiser les variations qui existent entre deux cultures, il persiste toujours un biais sur les prédictions des teneurs des biomarqueurs d'intérêt d'une seconde culture à partir de modèles de régression développés sur les données d'une première. Ceci traduit l'existence de variations inter-cultures inhérentes au développement du suivi de bioprocédés. Ainsi, afin d'améliorer les modèles de régression, il est nécessaire d'accumuler des données provenant de plusieurs lots de cultures cellulaires différents, ce qui permettrait d'incorporer la variabilité inter-*batch* aux modèles de régression pour les rendre plus robustes.

Chapitre 4
Optimisation et recherche de robustesse pour les modèles de régression des biomarqueurs

Au cours du chapitre précédent, nous avons mis en œuvre les outils spectroscopiques et statistiques afin de réaliser le suivi de cultures de cellules CHO, HeLa et Sf9. Nous avons déterminé les paramètres d'acquisition et de traitement les plus pertinents pour le développement de modèles de régression.

Toutefois, les travaux précédents ne mettent en œuvre que deux lots de données provenant uniquement de deux cultures de chaque type de cellule. Hors, nous avons pu voir dans la littérature [132,134-136] que la robustesse des modèles de régression passait avant tout par l'accumulation de données provenant de plusieurs cultures, ce qui permet d'introduire un maximum d'information quant à la variabilité inter-culture au sein des modèles de régression. Nous avons très bien pu constater l'existence de ce biais au cours de la section 3.3 du Chapitre 3 (Application du nouveau traitement de données spectrales) alors que nous réduisions les variations dues à l'installation du dispositif Raman. Rappelons que les travaux présentés au cours du chapitre précédent sont nécessairement effectués en début de projet afin de paramétrer l'acquisition, impliquant donc un nombre réduit de culture à disposition. Une fois le projet avancé, le nombre de cultures réalisées est plus important. L'objectif du travail présenté au cours de ce chapitre est alors de mettre en place des modèles de régression robustes prenant en compte les variations inter-*batches* qui existent en accumulant les données provenant de plusieurs cultures cellulaires.

1 Accumulation de cultures cellulaires

Dans un premier temps, afin de réduire la variabilité existante entre les cultures, nous accumulons plusieurs *batches* réalisés dans les mêmes conditions afin de conserver des profils similaires d'un point de vue biologique. Ainsi, pour chaque type de cellule, les mêmes paramètres physiques conditionnant la culture (aussi appelés CPP pour *Critical Process Parameters*) sont mis en place, soit les conditions normales pour le bioprocédé. De même,

les conditions d'acquisition Raman sont toujours les mêmes, soit un temps d'acquisition fixé à 5 min (10 × 30 s), les spectres étant toujours acquis en continu tout au long de la culture.

1.1 Modèles de régression pour les cellules CHO

Nous commençons par travailler sur les cultures de cellules CHO. Au total, nous disposons de neuf lots de culture réalisés dans les conditions normales de culture, à savoir à une température de 37°C et un pH de 6,9, une vitesse de rotation de l'hélice pour l'agitation du milieu (nous parlerons de vitesse d'agitation) fixée à 260 rpm et une pO_2 régulée à 40% de saturation. De plus, le système d'acquisition Raman utilisé permet de relier quatre sondes à immersion différentes au spectromètre, ce qui permet alors de réaliser le suivi spectroscopique de quatre cultures différentes simultanément. Les neuf lots utilisés pour le développement chimiométrique sont réalisés dans les conditions dites de référence et permettent donc de développer les modèles de régression PLS intégrant la variabilité qui existe entre les différents lots ainsi qu'entre les différentes sondes voire leurs positionnements. Nous conservons un maximum de *batches* dans l'ensemble de calibration pour de futurs développements. De ce fait, nous ne prendrons que les échantillons d'une seule de ces neuf cultures dans l'ensemble test.

Nous réalisons pour commencer une étude ACP sur les données des neuf cultures soit un total de 428 échantillons. Les spectres sont traités de la même façon que les derniers travaux du chapitre précédent, à savoir une dérivée première (Savitzky-Golay, fenêtre de 15 points et polynôme d'ordre 2) suivie d'une normalisation SNV puis une sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Le nombre d'échantillons pour chacune des cultures est récapitulé dans le Tableau 4.1.

Tableau 4.1 Récapitulatif du nombre d'échantillons pour chaque culture de cellules CHO employée afin de développer un modèle de régression intégrant les variations inter-*batches*.

Numéro de lot	Nombre d'échantillons
#1	57
#2	17
#3	51
#4	48
#5	62
#6	62
#7	45
#8	43
#9	43

Il est important de noter à nouveau que toutes les cultures mises en jeu n'ont pas été acquises au moyen de la même sonde à immersion. Ainsi, non seulement les variabilités inter-*batches* pourront être intégrées aux modèles, mais aussi les variabilités inter-sondes. Les résultats obtenus par l'étude ACP sont représentés Figure 4.1 à travers les graphiques des *scores* et les *loadings* des premières composantes principales obtenues.

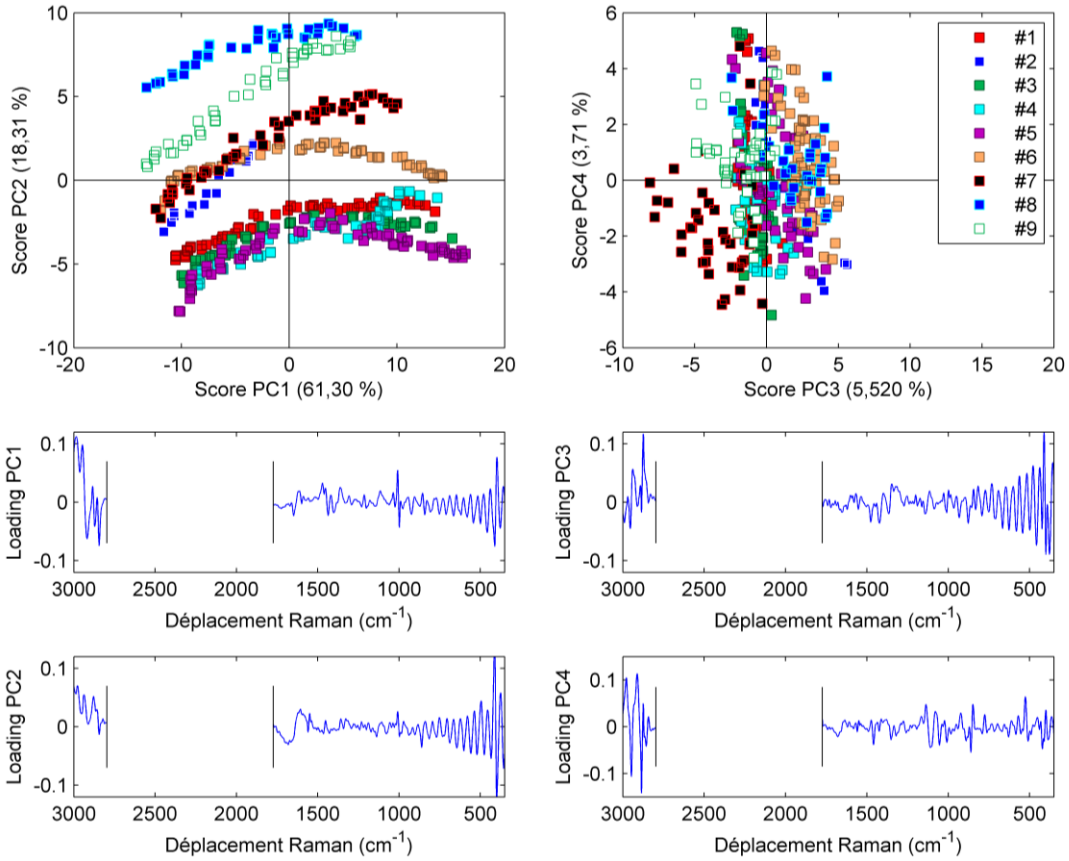

Figure 4.1 Résultats obtenus pour l'analyse ACP réalisée sur un jeu de données cumulant les échantillons provenant de neuf cultures de cellules CHO réalisées dans les conditions normales de culture.

Dans un premier temps, nous pouvons observer sur le premier graphique des *scores* les séparations qui existent entre les différents lots. En effet, bien que le prétraitement appliqué réduise les différences inter-*batches*, nous avons pu voir qu'il n'effaçait pas totalement ces variations, traduites par la PC2 (18,13 %). Nous observons également, toujours sur le graphique des *scores*, que la première composante principale PC1 (61,30 %) représente l'évolution du bioprocédé au cours du temps. Les composantes principales PC3 (5,52 %) et PC4 (3,71 %) montrent également des séparations entre les différents lots de culture, néanmoins moins prononcées que sur la composante PC2. Sur les graphiques des *loadings* de ces composantes principales, nous pouvons souligner la présence de signaux particuliers entre 1000–400 cm^{-1}. Ces-derniers sont principalement dus à des différences d'intensité entre plusieurs cultures dans ces régions spectrales, ce qui provoque ces phénomènes interférents pris en compte par le modèle ACP.

Afin de construire un modèle de régression robuste, nous souhaitons donc intégrer les différences observées entre les cultures en cumulant les lots dans l'ensemble de calibration. Toutefois, il est nécessaire de conserver un certain nombre d'échantillons dans un ensemble test absent de l'ensemble d'apprentissage. Dans le cas des cellules CHO, nous exclurons les échantillons d'une seule culture pour l'évaluation des modèles de régression. En effet, bien que cela ne soit pas totalement satisfaisant, nous devons souligner le fait que les modèles de régression développés ici seront réutilisés par la suite. Nous souhaitons donc intégrer le plus grand nombre de *batches* possible à l'ensemble de calibration. En prenant les résultats obtenus par l'étude ACP, nous sélectionnons une culture ne présentant pas de comportement extrême. Par exemple, les échantillons du lot #8 présentent des *scores* importants sur la composante PC2, Figure 4.1, traduisant donc plus de variations par rapport au reste de la population. Nous éviterons alors de prendre ce type de lot pour éprouver les modèles calculés. Nous sélectionnons finalement le lot #1 en tant que jeu test.

Les modèles de régression sont réalisés sur la base de l'ensemble d'apprentissage et valider une première fois à l'aide d'une validation croisée (*10-fold cross-validation*), ce qui permet de déterminer le nombre optimal de facteurs (variables latentes) puis l'erreur RMSECV, puis à l'aide du lot de test, permettant d'accéder à l'erreur RMSEP. Les résultats obtenus pour les modèles de régression calculés pour chacun des paramètres métaboliques sont représentés Tableau 4.2.

Tableau 4.2 Récapitulatif des performances des modèles de régression développés en cumulant les échantillons des cultures de cellules CHO réalisées dans les conditions normales de culture.

Biomarqueur (unité)	R^2	LV	RMSECV	RMSEP	Gamme
Glucose (mM)	0,883	8	4,25	5,68	2,25 – 62,13
Glutamine (mM)	0,917	5	0,17	0,46	0,00 – 2,72
Glutamate (mM)	0,952	6	0,63	0,41	2,21 – 13,67
Lactate (mM)	0,969	8	1,55	2,52	0,00 – 39,18
Ammonium (mM)	0,833	8	0,85	0,99	0,00 – 9,70
VCD (10^6 cell/mL)	0,945	7	0,99	1,10	0,29 – 15,57
TCD (10^6 cell/mL)	0,942	7	1,05	1,13	0,29 – 15,70

Pour étudier les performances des différents modèles développés à partir des données recueillies sur les cultures de cellules CHO, nous travaillerons suivant la même logique que pour l'étude du temps d'acquisition des spectres Raman lors du chapitre précédent (partie 2.2, Temps d'acquisition pour les cultures de cellules CHO), à savoir en étudiant dans un premier temps les paramètres glucose et lactate, ensuite la glutamine, le glutamate et l'ammonium, et enfin les densités cellulaires TCD et VCD.

1.1.1 Modèles pour les concentrations en glucose et en lactate

Les modèles de régression calculés pour les concentrations en glucose et lactate sont représentés Figure 4.2 à travers les graphiques des concentrations prédites par rapport aux concentrations calculées, accompagnés des coefficients de régression, marqueurs de fiabilité des modèles. Dans un premier temps, nous pouvons remarquer que les erreurs RMSEP des deux modèles sont plus importantes que lors de l'étude du temps d'acquisition des spectres Raman (5,60 mM contre 2,03 mM pour la concentration en glucose et 2,52 mM contre 0,70 mM pour la concentration en lactate). En effet, ceci provient directement de l'ensemble test employé et de toutes les variations présentes dans le lot de calibration. Lors de l'étude du temps (page 89), nous travaillions sur les échantillons d'une seule et même culture, nous n'avions donc aucun phénomène provenant des variations inter-*batches* qui entraient en ligne de compte pour l'erreur RMSEP.

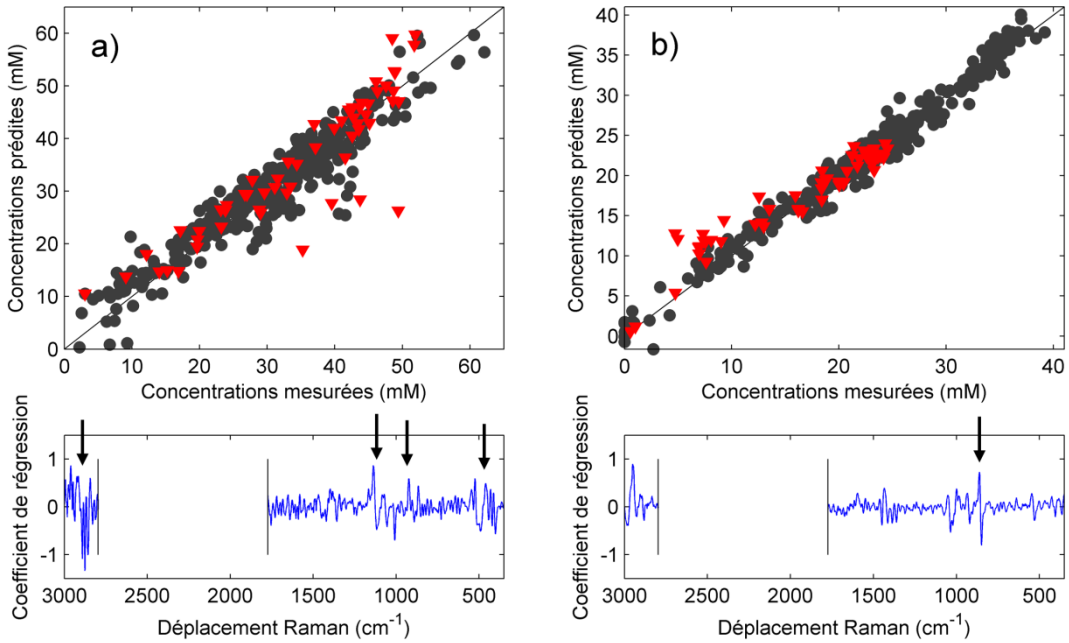

Figure 4.2 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules CHO pour a) le glucose et b) le lactate. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Ici, bien que les variations entre les cultures soient intégrées aux modèles, la prédiction des échantillons d'un nouveau lot reste plus difficile. De plus, pour les deux paramètres, nous pouvons voir que les gammes de concentrations sont plus élevées et présentent plus de variations, passant de 11,54–58,70 mM à 2,25–62,13 mM et 0,00–10,06 mM à 0,00–39,18 mM pour le glucose et le lactate respectivement. Toutefois, comparées aux erreurs RMSECV obtenues en réalisant la validation croisées sur un ensemble d'échantillons provenant de huit cultures de cellules CHO, nous pouvons considérer à partir du Tableau 4.2 que les erreurs obtenues sont acceptables. Enfin, nous pouvons souligner l'augmentation du

nombre de variables latentes employées pour construire les modèles, provenant simplement du fait que nous avons plus de variance à expliquer afin d'obtenir des modèles de régression satisfaisants.

L'étude des coefficients de régression doit une nouvelle fois être réalisée afin de justifier les performances des modèles de régression. Ces derniers sont similaires à ceux obtenus lors du développement de modèles de régression pour l'étude des temps d'acquisition, en ajoutant cette fois la région spectrale 3000–2800 cm^{-1}. Ainsi, nous pouvons voir que les modèles glucose (Figure 4.2a) et lactate (Figure 4.2b) reposent en partie sur les bandes de cette nouvelle région spectrale traduisant les élongations v(C–H) des différentes molécules [146]. Pour le glucose, nous pouvons remarquer que les bandes comprises entre 1200–1000 cm^{-1} sont principalement dues aux élongations v(C–C) et v(C–O) [137] tandis que les bandes autour de 900 cm^{-1} représentent les déformations des bandes δ(COH), δ(CCH) et δ(OCH) du cycle carboné [137]. Enfin, toujours pour les coefficients de régression du modèle PLS glucose, nous pouvons observer des signaux importants entre 550–400 cm^{-1}, caractérisant les déformations exocycliques (autour de 520 cm^{-1}) et endocycliques (autour de 450 cm^{-1}) de la molécule [137]. Pour le lactate, nous pouvons remarquer que le modèle repose majoritairement sur la bande à 860 cm^{-1} traduisant l'élongation de la bande v(C–CO_2^-) du groupement carboxylate, principale fonction de la molécule [138].

Nous pouvons donc considérer que les modèles de régression construits pour les concentrations en glucose et en lactate reposent bien sur les fonctions chimiques des paramètres métaboliques. De plus, à travers les représentations proposées Figure 4.2, nous pouvons avoir un aperçu des capacités prédictives des modèles. Toutefois, afin de représenter de manière plus efficace les prédictions des échantillons provenant de la culture de cellules #1 (culture test), nous pouvons reprendre tous les spectres acquis en continu et y appliquer les modèles de régression PLS. Nous pouvons ainsi obtenir les tendances des concentrations métaboliques prédites tout au long de la culture. Pour la culture de cellules #1, nous avions acquis les spectres en continu pendant plus de 437 h (soit plus de 18 jours), ce qui correspondait finalement à l'acquisition de 4380 spectres de 5 min (10 × 30 s). La Figure 4.3 présente les prédictions des teneurs en glucose et en lactate obtenues pour ces spectres acquis en continu tout au long de la culture.

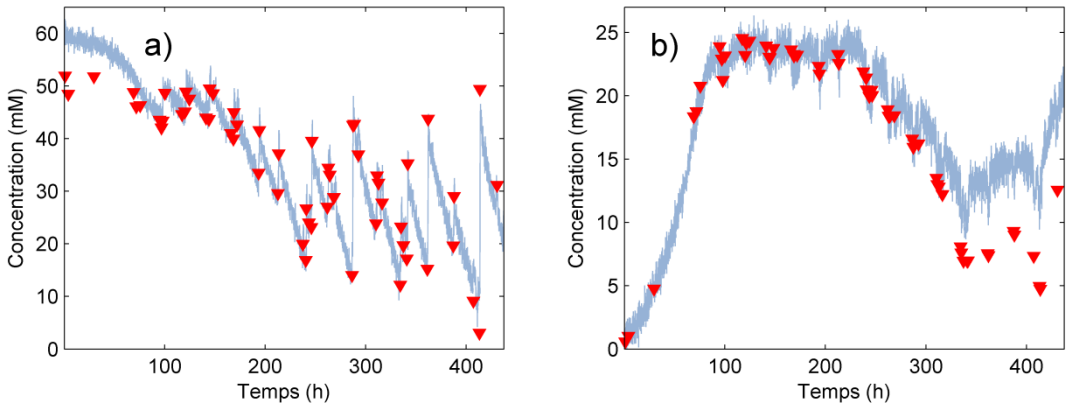

Figure 4.3 Représentations des prédictions réalisées sur les spectres de la culture CHO test acquis en continu (en bleu) pour a) le glucose et b) le lactate. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges.

Nous pouvons alors voir que les modèles de prédiction parviennent à bien suivre les tendances des concentrations métaboliques. En effet, sur la Figure 4.3a, nous pouvons voir que le modèle de régression permet de montrer la consommation du glucose tout en faisant ressortir les *feeds* réalisés pendant la culture (sauts de concentrations à intervalles réguliers). Toutefois, nous pouvons voir un écart sur les prédictions réalisées en début de culture, c'est-à-dire sur les fortes concentrations en glucose, que nous observions déjà sur le graphique des concentrations prédites en fonction des concentrations mesurées (Figure 4.2a). Ceci s'explique avant tout par un manque d'échantillons de référence dans cette zone de concentration. En ce qui concerne les prédictions faites pour le niveau de lactate (Figure 4.3b), nous pouvons voir que nous sommes en mesure d'obtenir la véritable tendance métabolique du paramètre. En effet, nous pouvons observer en premier lieu la forte croissance de la concentration due à la multiplication cellulaire au sein du milieu, produisant donc beaucoup de lactate. La stagnation qui suit annonce la phase de sénescence et donc de mort cellulaire. Les cellules ne produisent donc plus de lactate. Enfin, nous pouvons observer la décroissance de la concentration en lactate, déjà présentée Figure 1.5 (page 31) pour la reformation de pyruvate dans le milieu.

Finalement, nous pouvons conclure que les deux modèles de régression calculés pour les paramètres glucose et lactate permettent de réaliser de manière cohérente les prédictions des concentrations métaboliques tout au long d'une culture de cellules dont les échantillons n'ont pas été pris en compte dans l'ensemble d'étalonnage. De plus, le fait d'acquérir les spectres en continu permet de confirmer que les modèles sont bien capables de suivre correctement l'évolution des concentrations des métabolites lors d'une culture de cellules réalisée dans les conditions normales.

1.1.2 Modèles pour les concentrations en glutamine, en glutamate et en ammonium

Lors de l'étude des paramètres glutamine, glutamate et ammonium dans la partie 2.2.2 du Chapitre 3 (page 93), nous avions pu montrer que les modèles de régression basés sur les molécules de glutamine et glutamate étaient pertinents, tandis que le modèle de régression pour la prédiction des concentrations en ammonium ne s'appuyait pas sur les bandes caractéristiques de la molécule.

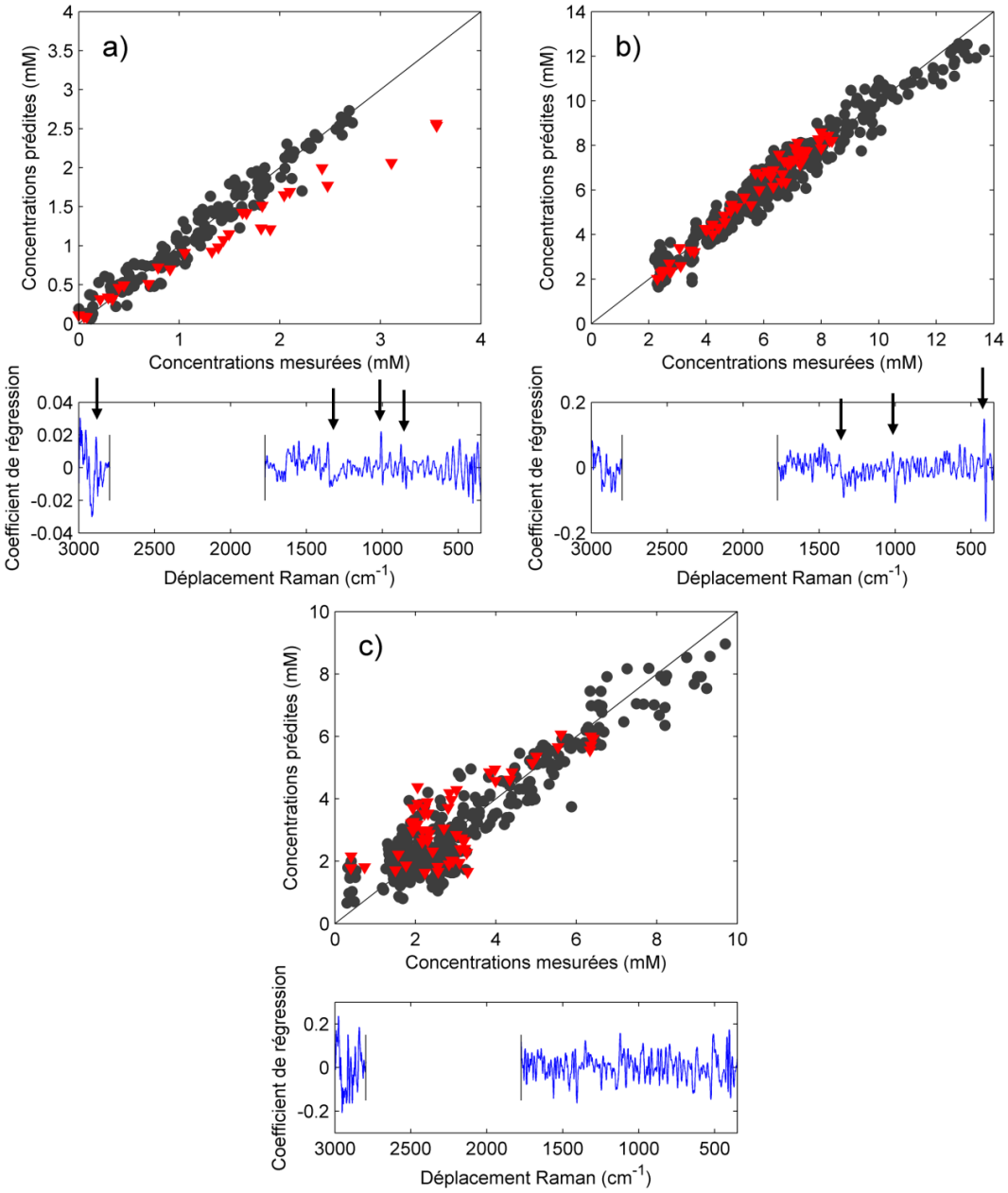

Figure 4.4 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules CHO pour a) la glutamine, b) le glutamate et c) l'ammonium. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Les modèles de régression obtenus en accumulant les données des huit lots de cellules CHO cultivées dans les conditions normales sont représentés Figure 4.4a pour les concentrations en glutamine, Figure 4.4b pour les concentrations en glutamate et Figure 4.4c pour les concentrations en ammonium. En reprenant les résultats obtenus pour chaque métabolite (Tableau 4.2) nous pouvons remarquer que, de manière générale, le nombre de variables latentes est plus important pour construire les modèles ici que pour l'étude du temps d'acquisition (Tableau 3.1, page 89), et ce pour les mêmes raisons que pour les modèles des concentrations en glucose et lactate : le fait d'avoir plus de données induit plus de variance à expliquer lors de la modélisation PLS.

En prenant d'abord le modèle calculé pour les concentrations en glutamine, nous pouvons voir que celui-ci présente une erreur RMSEP importante (0,46 mM) par rapport à l'erreur RMSECV calculée (0,17 mM). La Figure 4.4a permet de montrer que les prédictions réalisées pour les échantillons de la culture test présentent des écarts de plus en plus importants à mesure que la concentration en glutamine croit, ce qui engendre cette erreur RMSEP. Ceci provient directement de la gamme de concentrations obtenue à partir des mesures des prélèvements par la méthode de référence.

Pour la glutamine, les mesures sont arrêtées à partir de 240 h de culture (soit au bout de 10 j) pour la culture test, ce paramètre n'étant plus suivi à partir de ce moment (en routine). De ce fait les concentrations en glutamine présentent des gammes de variations faibles (0,00–2,72 mM) et donc difficilement perceptibles en spectroscopie Raman. Toutefois, les coefficients de régression permettent de mettre en avant des bandes caractéristiques similaires à ce que nous obtenions précédemment. Ainsi, les signaux obtenus à 1020 cm^{-1} et 850 cm^{-1} permettent de traduire les élongations ν(C–C) de la chaîne carbonée [139,140]. Nous pouvons également distinguer plusieurs bandes importantes entre 1500–1350 cm^{-1} représentant les groupements amide des molécules de glutamine (élongations ν(C–N) et déformations δ(N–H)) et les groupements carboxyle et les déformations $\delta(CH_2)$ de la chaîne carbonée [139,140]. Enfin, nous pouvons voir que les bandes dans la région 3000–2800 cm^{-1}, traduisant les élongations ν(C–H) [146], présentent également des coefficients de régression importants. Malgré les observations réalisées sur les prédictions des échantillons de la culture test, nous pouvons considérer que le modèle obtenu est satisfaisant pour suivre la tendance du métabolite, les coefficients de régression étant cohérents.

Tout comme pour le glucose et le lactate, en prédisant les concentrations en glutamine à partir des 4380 spectres acquis en continu pour la culture test, nous pouvons obtenir l'évolution de la teneur en glutamine pendant toute la culture. Cette évolution, représentée

Figure 4.5a, permet de montrer que la tendance du paramètre glutamine est convenablement déterminée malgré les écarts croissants. En plus de montrer l'évolution du biomarqueur, le modèle de régression permet également de faire ressortir les *feeds* réalisés durant la culture qui impactent la concentration du métabolite (de manière moins importante que le glucose, il convient de noter). Nous devons ici noter que les *feeds* réalisés contiennent de la glutamine, raison pour laquelle nous observons des sauts de concentration lors du *feeding*.

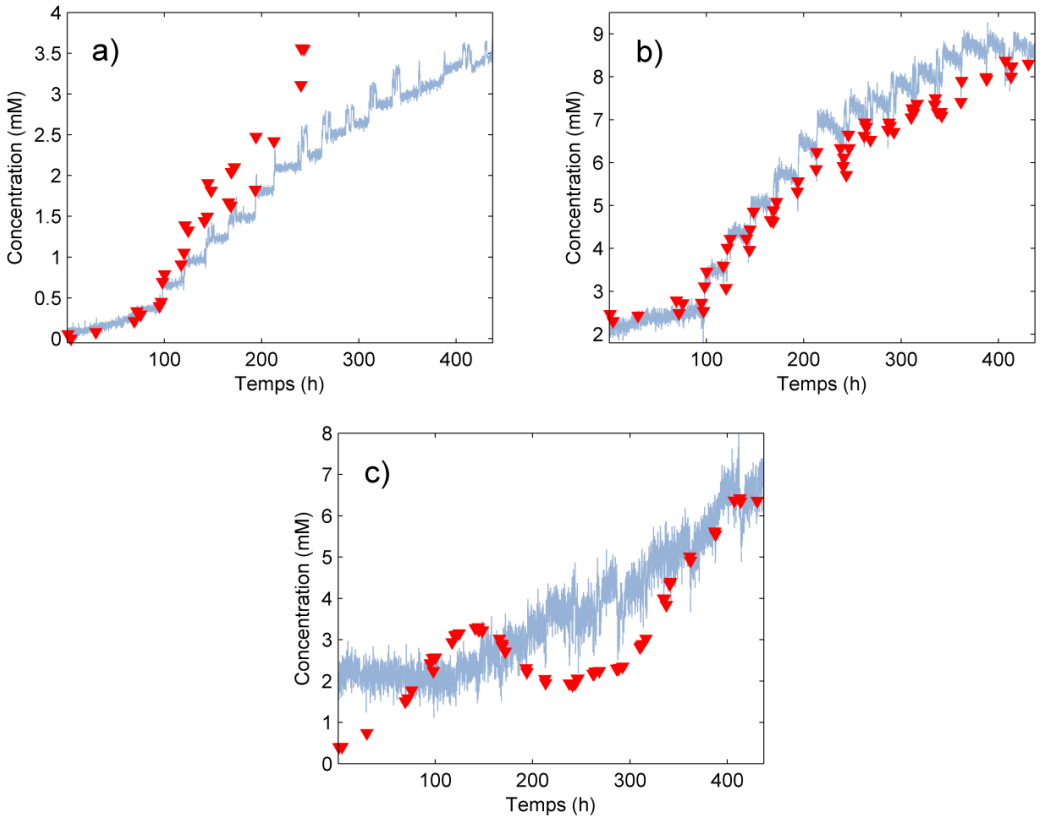

Figure 4.5 Représentations des prédictions réalisées sur les spectres de la culture de cellules CHO test acquis en continu (en bleu) pour a) la glutamine, b) le glutamate et c) l'ammonium. Les concentrations de référence sont représentées par les triangles rouges.

En ce qui concerne le modèle de régression pour les concentrations en glutamate, nous pouvons d'abord remarquer que l'erreur RMSEP (0,41 mM) est moins importante que l'erreur RMSECV (0,63 mM), Tableau 4.2. Ceci traduit simplement le fait que les variations entre les échantillons présents lors de la validation croisée sont plus importantes que les variations qui existent par rapport à l'ensemble test. Donc, puisque plus de variations sont prises en compte durant l'élaboration du modèle, les prédictions réalisées présentent moins d'erreur. Nous pouvons observer cela Figure 4.4b, où les écarts observés entre les concentrations prédites et les concentrations mesurées sont en moyennes plus importants pour les échantillons de calibration (points noirs) que pour les échantillons test (triangles rouges). L'analyse des coefficients de régression permet de montrer que ces derniers reposent principalement sur la bande à 400 cm^{-1} traduisant en partie les vibrations du squelette

carboné du glutamate [140]. La bande à 1020 cm^{-1} est également importante et traduit les élongations v(C–C) de la chaîne carbonée. Enfin, la bande à 1350 cm^{-1} traduit les vibrations v(CO_2^-) du groupement carboxylate de la molécule. En prenant les prédictions réalisées sur les spectres acquis en continu durant toute la durée de la culture test (Figure 4.5b), nous pouvons voir que la tendance de la concentration en glutamate est respectée tout au long de la culture et permet également de mettre en avant la présence de *feeds* puisque ce paramètre, tout comme la glutamine, est impacté par ce phénomène (également de manière moins importante que le glucose, celui-ci étant bien plus concentré dans les *feeds*).

Pour finir, nous pouvons voir que les résultats obtenus pour les concentrations en ammonium présentent des erreurs RMSECV (0,85 mM) et RMSEP (0,99 mM) faibles compte tenu de la gamme d'étalonnage (0,30–9,70 mM). En prenant en compte le graphique des concentrations prédites en fonction des concentrations mesurées, Figure 4.5c, nous pouvons voir qu'il existe des écarts importants entre les prédictions et les mesures de référence. En réalité, en prenant en compte les coefficients de régression, nous pouvons une nouvelle fois voir que ces-derniers ne reposent pas sur les bandes caractéristiques de la molécule d'ammonium, tout comme nous avions pu le montrer lors du chapitre précédent (partie 2.2.2 du Chapitre 3, Modèles pour les niveaux en glutamine, glutamate et ammonium, page 93). Ainsi, le modèle de régression calculé pour la concentration en ammonium en prenant en compte les éléments des huit cultures présentes dans l'ensemble d'apprentissage ne permet pas de développer de modèle de régression cohérent. En prenant les prédictions réalisées sur les 4380 spectres acquis en continu, nous pouvons très bien voir que le profil obtenu pour les prédictions en ammonium ne suit pas de manière convenable l'évolution obtenue à l'aide de la méthode de référence. Il ne s'agit pas ici de simples écarts de prédiction permettant toutefois d'obtenir l'évolution du paramètre biochimique (tel que le modèle de régression pour les concentrations en glutamine par exemple) mais bel et bien d'un manque de corrélation entre les spectres et les mesures obtenus à l'aide de la méthode de référence. Nous pouvons également noter la répétabilité de prédiction moins importante lorsque des spectres consécutifs sont sollicités.

1.1.3 Modèles pour les densités TCD et VCD

Les modèles de régression basés sur la densité en cellules vivantes VCD et la densité totale en cellules TCD sont très similaires. Dans un premier temps, nous pouvons une nouvelle fois souligner le fait que nous utilisons plus de variables latentes pour construire les modèles de régression basés sur les densités cellulaires lors de l'accumulation d'échantillons provenant de plusieurs cultures de cellules (Tableau 4.2) que lors de l'étude du temps d'acquisition (Tableau 3.1, page 89). Nous pouvons ensuite remarquer que les

deux modèles de régression sont très similaires. En effet, les mesures de référence pour les densités TCD et VCD sont très proches, les valeurs obtenues pour ce dernier paramètre étant évidemment inférieures. Ceci joue également sur les coefficients de régression des deux modèles (Figure 4.6).

Nous pouvons voir que les profils obtenus présentent des allures très similaires. Le manque de différence entre les mesures de référence des deux densités ne permet pas de construire de modèle de régression bien distinct entre les paramètres VCD et TCD, ce qui apparait à travers les coefficients de régression similaires Figure 4.6 et les mesures de référence Figure 4.7. Néanmoins, les erreurs de prédiction RMSEP obtenues pour chaque paramètre biochimique ($1{,}10 \cdot 10^6$ cell/mL pour la densité VCD et $1{,}13 \cdot 10^6$ cell/mL pour la densité TCD) soulignent des capacités prédictives satisfaisantes. Cependant, en prenant en compte indépendamment les prédictions réalisées sur les spectres acquis tout au long de la culture, Figure 4.7, nous pouvons voir que les modèles permettent de bien suivre les tendances de chacun des paramètres biochimiques pendant la culture. Nous pouvons également noter l'impact des *feeds*, notamment à partir du milieu de culture pour les deux paramètres. En effet, l'ajout de *feed* entraine une augmentation du volume total du contenu du bioréacteur et donc une diminution des densités en cellules, ce qui est observé sur les traces déterminées, Figure 4.7a pour le paramètre VCD et Figure 4.7b pour le paramètre TCD.

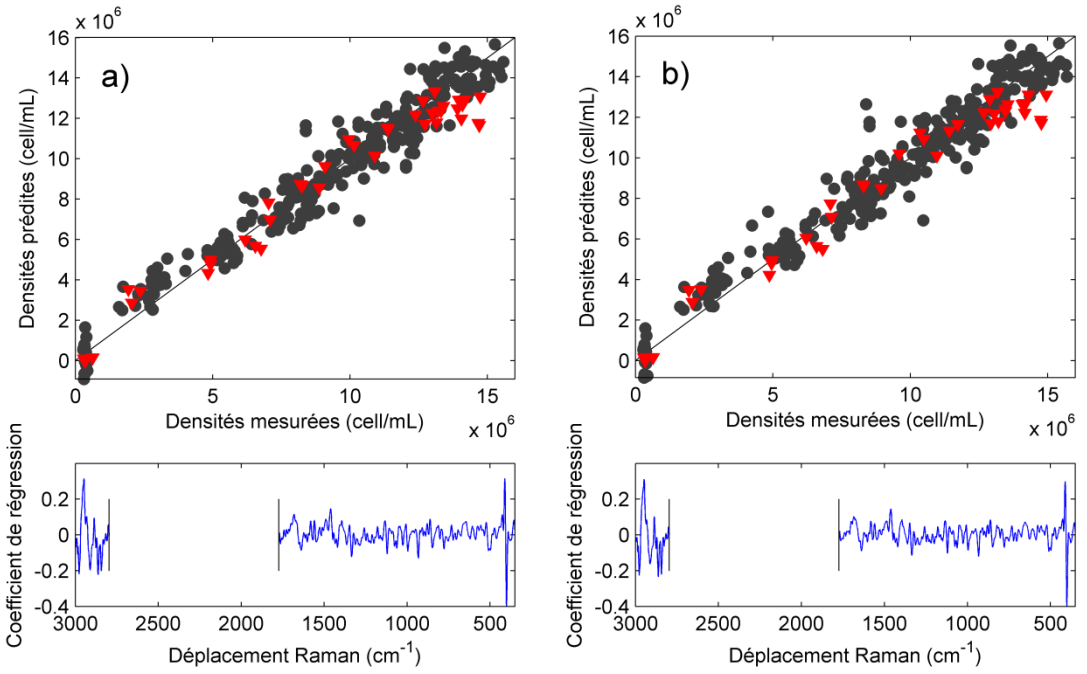

Figure 4.6 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules CHO pour a) la densité en cellules vivantes et b) la densité totale en cellules. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Pour conclure quant aux modèles de régression développés pour les cultures de cellules CHO, nous avons pu voir qu'en prenant en compte dans l'ensemble de calibration les échantillons provenant de plusieurs lots, nous pouvions réduire les variations qui existent entre les différents *batches* en intégrant celles-ci aux modèles de régression. En prenant en considération au sein du jeu test les échantillons d'une culture extérieure à l'ensemble d'apprentissage, nous pouvons alors éprouver les modèles de régression et présenter leurs capacités prédictives pour les différents biomarqueurs engagés. Rappelons que si nous prenons l'erreur calculée lors de la prédiction des densités VCD d'une culture de cellules CHO par un modèle de régression basé sur les échantillons d'une seule culture (résultats obtenus section 3.3 du Chapitre 3, Application du nouveau traitement de données spectrales, page 123), nous obtenions une erreur RMSEP à $1{,}54 \cdot 10^6$ cell/mL. En cumulant les échantillons de huit cultures, nous obtenons ici une erreur RMSEP de $1{,}10 \cdot 10^6$ cell/mL. D'après cette amélioration, nous pouvons en partie conclure que le fait d'accumuler les différentes cultures permet d'améliorer la robustesse du modèle de régression.

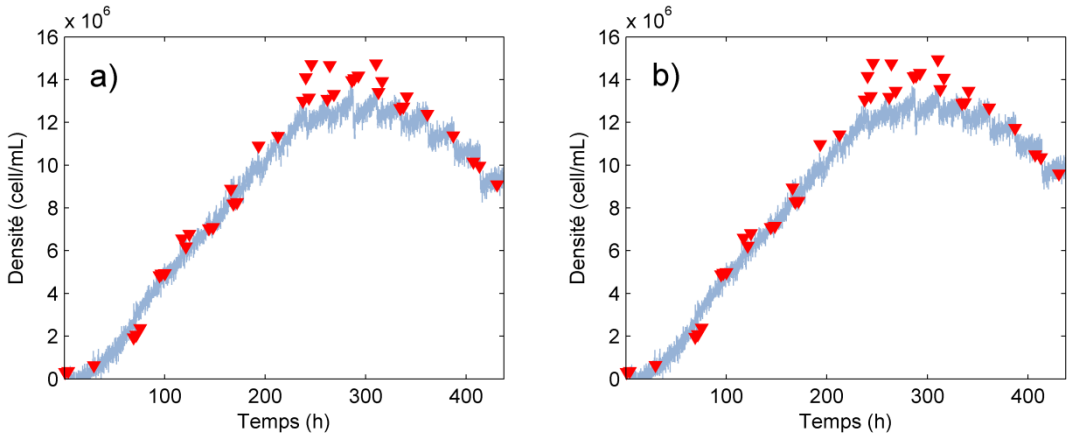

Figure 4.7 Représentations des prédictions réalisées sur les spectres de la culture de cellules CHO test acquis en continu (en bleu) pour a) la densité VCD et b) la densité TCD. Les densités obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges.

Nous réalisons alors des travaux similaires pour les modélisations chimiométriques des paramètres biochimiques des cultures des autres types de cellules mis en jeu au cours de ces travaux de recherche, à savoir les cellules HeLa et les cellules Sf9.

1.2 Modèles de régression pour les cellules HeLa

Pour étudier les cellules HeLa, nous disposons de sept cultures différentes acquises dans les conditions normales comprenant non seulement la phase de croissance cellulaire (phase pré-inoculation virale), mais également la phase de culture virale (phase post-inoculation virale) pour un total de 366 échantillons. Le nombre d'échantillons prélevés pour chaque culture est détaillé au sein du Tableau 4.3.

Tableau 4.3 Récapitulatif du nombre d'échantillons pour chaque culture mettant en jeu les cellules HeLa employée afin de développer les modèles de régression intégrant les variations inter-*batches*, séparant les échantillons acquis avant et après la phase d'inoculation virale. Les lots #4 et #5 mettent en jeu deux sondes différentes pour un même suivi.

Numéro de lot	Nombre d'échantillons		
	Pré*	Post*	*(Total)*
#1	8	6	*(14)*
#2	19	13	*(32)*
#3	40	30	*(70)*
#4-1	39	34	*(73)*
#4-2	39	34	*(73)*
#5-1	7	15	*(22)*
#5-2	7	15	*(22)*
#6	20	14	*(34)*
#7	16	10	*(26)*

** Pré- et Post-inoculation virale*

Nous pouvons noter que les cultures #4 et #5 sont répartis en deux lots distincts. En effet, pour le suivi de ces deux cultures, deux sondes Raman furent immergées au sein du bioréacteur pour suivre le même bioprocédé. Ceci permet notamment d'inclure au sein du jeu de données les variations provenant de différentes sondes. Lorsque rien n'est indiqué (cas des autres cultures), cela signifie que seule la sonde 1 a été mise en jeu pour acquérir les spectres Raman.

Dans un premier temps, une analyse ACP est réalisée sur les 336 spectres reliés aux mesures de référence. Notons une nouvelle fois que les spectres sont traités de la même façon que lors des travaux réalisés pour les échantillons des cultures de cellules CHO (section 1.1 Modèles de régression pour les cellules CHO). Les résultats obtenus pour cette analyse sont représentés Figure 4.8 à travers les graphiques des *scores* des premières composantes principales ainsi que les *loadings* associés.

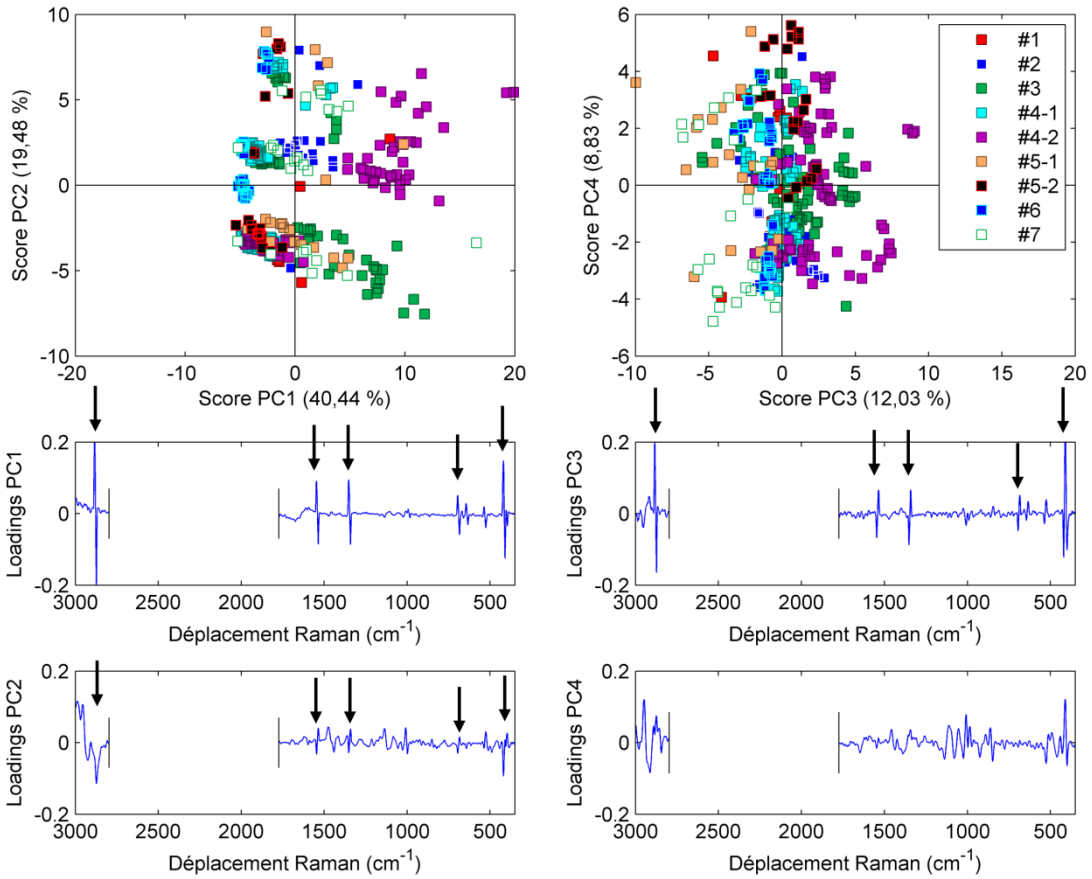

Figure 4.8 Résultats obtenus pour l'analyse ACP réalisée sur un jeu de données cumulant les échantillons provenant de sept cultures de cellules HeLa réalisées dans les conditions normales de culture. Les phases pré- et post-inoculation virale apparaissent sur ces graphiques.

Nous pouvons d'abord remarquer d'après les *loadings* des composantes principales PC1 (40,44 %), PC2 (19,48 %) et PC3 (12,03 %) que ces dernières reposent sur des variables spectrales similaires. Ces contributions indiquent la présence de signaux provenant de la lumière extérieure sur les spectres. Malheureusement, cela traduit une certaine perméabilité des systèmes de protection des bioréacteurs de culture de cellules HeLa. Cependant, il est très difficile de corriger numériquement ce phénomène. De plus, nous ne sommes pas en mesure de discriminer parfaitement les échantillons sur le graphique des *scores* et donc les spectres sujets à cette perturbation, ce qui empêche donc d'exclure cette population. En ce qui concerne les composantes principales PC2 (19,48 %) et PC4 (8,83 %), nous pouvons noter que celles-ci traduisent simplement l'évolution des cultures au cours du temps. En effet, les *scores* de ces composantes évoluent de la même façon que le temps de culture.

Pour le développement des modèles de régression de chacun des métabolites, il convient de scinder l'ensemble de données en deux jeux : un ensemble de calibration et un ensemble test. Dans le cas des cellules HeLa, et contrairement aux travaux menés pour les cellules CHO, nous écartons trois lots afin d'avoir un jeu test suffisamment important pour

éprouver les différents modèles de régression. Rappelons que le choix réalisé pour les cultures de cellules CHO (un seul lot de test) était fait pour conserver un maximum d'information dans l'ensemble d'apprentissage. En effet, les modèles de régression pour ce type de cellule seront éprouvés non seulement par cette unique culture test réalisée dans les conditions normales, mais aussi par d'autres lots différents ultérieurement. Ici, pour les cellules HeLa, nous intégrons les lots #1, #2 et #5 à l'ensemble test. Nous pouvons noter que les échantillons de la culture #5 provenant des différentes sondes seront intégrées conjointement afin d'associer les variations inter-sondes à la fois à l'ensemble test et à l'ensemble de calibration (notamment à travers les échantillons du lot #4). Nous obtenons donc un ensemble d'apprentissage de 276 échantillons pour un jeu test de 90 échantillons provenant de plusieurs cultures impliquant les cellules HeLa et réalisées dans les conditions normales. Les performances des modèles de régression développés pour chaque paramètre biochimique sont indiquées Tableau 4.4. À noter que les modèles sont réalisés de la même façon que pour la partie concernant les cellules CHO (première évaluation des modèles à l'aide d'une validation croisée permettant également de déterminer le nombre de variables latentes à prendre en compte).

Tableau 4.4 Récapitulatif des performances des modèles de régression pour les concentrations métaboliques, développés en cumulant les échantillons des cultures de cellules HeLa réalisées dans les conditions normales de culture, intégrant les phases pré- et post-inoculation virale.

Biomarqueur (unité)	R^2	LV	RMSECV	RMSEP	Gamme
Glucose (g/L)	0,975	6	0,23	0,36	0,00 – 6,20
Glutamine (mM)	0,930	6	0,31	0,45	0,00 – 5,44
Glutamate (mM)	0,766	8	0,31	0,42	0,00 – 2,55
Lactate (g/L)	0,929	6	0,08	0,22	0,00 – 1,89
Ammonium (mM)	0,907	8	0,28	0,42	0,32 – 5,04

L'étude des différents modèles de régression se fera suivant le même ordre que lors de l'étude du temps d'acquisition (section 2.3 du Chapitre 3, Temps d'acquisition pour les cultures de cellules HeLa, page 97).

1.2.1 Modèles pour les concentrations en glucose et en glutamine

Les premiers paramètres biochimiques étudiés sont les concentrations en glucose et en glutamine, ces deux métabolites étant consommés par les cellules HeLa pour le développement cellulaire. La Figure 4.9 permet de représenter les modèles obtenus à travers les graphiques des concentrations prédites en fonction des concentrations mesurées à l'aide des méthodes de référence, accompagnés des coefficients de régression.

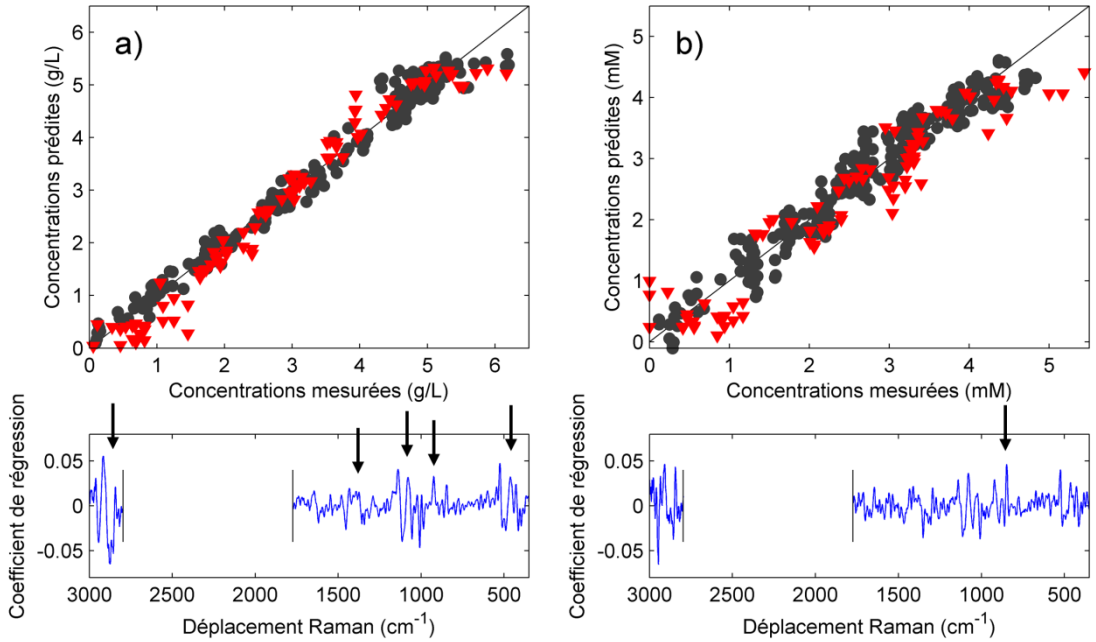

Figure 4.9 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules HeLa pour les concentrations a) en glucose et b) en glutamine. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

En prenant en compte les différentes erreurs calculées (Tableau 4.4), nous pouvons voir que les modèles de régression présentent des capacités prédictives satisfaisantes. En effet, l'erreur RMSEP du modèle PLS pour la concentration en glucose atteint 0,36 g/L pour une gamme de concentration comprise entre 0,00–6,20 g/L. Cependant, en prenant la représentation Figure 4.9a, et en comparant l'erreur RMSEP à l'erreur RMSECV (0,23 g/L), nous pouvons remarquer les sous-évaluations des faibles concentrations et de quelques fortes concentrations (autour de 6,00 g/L). Néanmoins, en étudiant les coefficients de régression, nous pouvons voir que ceux-ci s'appuient bien sur les bandes caractéristiques du glucose. Nous pouvons remarquer la présence de signaux intenses autour de 520 cm^{-1} et 450 cm^{-1} caractérisant respectivement les déformations exocycliques et endocycliques de la molécule de glucose [137]. De plus, les coefficients de régression entre 1200–1000 cm^{-1} traduisent les élongations ν(C–C) et ν(C–O) du cycle benzénique de la molécule [137]. Nous pouvons également noter la présence de signaux importants entre 3000–2800 cm^{-1} dus aux élongations ν(C–H) de la molécule [146]. Enfin, nous pouvons observer les faibles contributions du groupement hydroxyméthyle autour de 1400 cm^{-1} (déformations $\delta(CH_2)$ et $\delta(CH_2OH)$) [137] ainsi que les faibles signaux des déformations δ(COH) et δ(CCH) autour de 900 cm^{-1} [137].

Afin de représenter les prédictions des concentrations en glucose réalisées en continu pour les trois lots test, nous reprenons les spectres acquis en continu pour chacun des lots. Nous avons 15458 spectres acquis en continu pour la culture #1, 14022 spectres pour la

culture #2, ainsi que 1262 spectres et 1263 spectres pour respectivement les sondes 1 et 2 de la culture #5. L'importance des nombres de spectres acquis en continu pour les lots #1 et #2 est due au protocole d'acquisition mis en œuvre. En effet, ces cultures ont été acquises dans les conditions expérimentales mises en œuvre pour l'étude du temps d'acquisition Raman (soit des spectres de 1 × 30 s en continu). Ainsi, en moyennant les spectres artificiellement, nous disposons de près de dix fois plus de spectres Raman que pour un protocole d'acquisition en continu classique de spectres de 10 × 30 s. Les prédictions continues des concentrations en glucose réalisées pour les cultures test sont confrontées aux mesures de référence Figure 4.10.

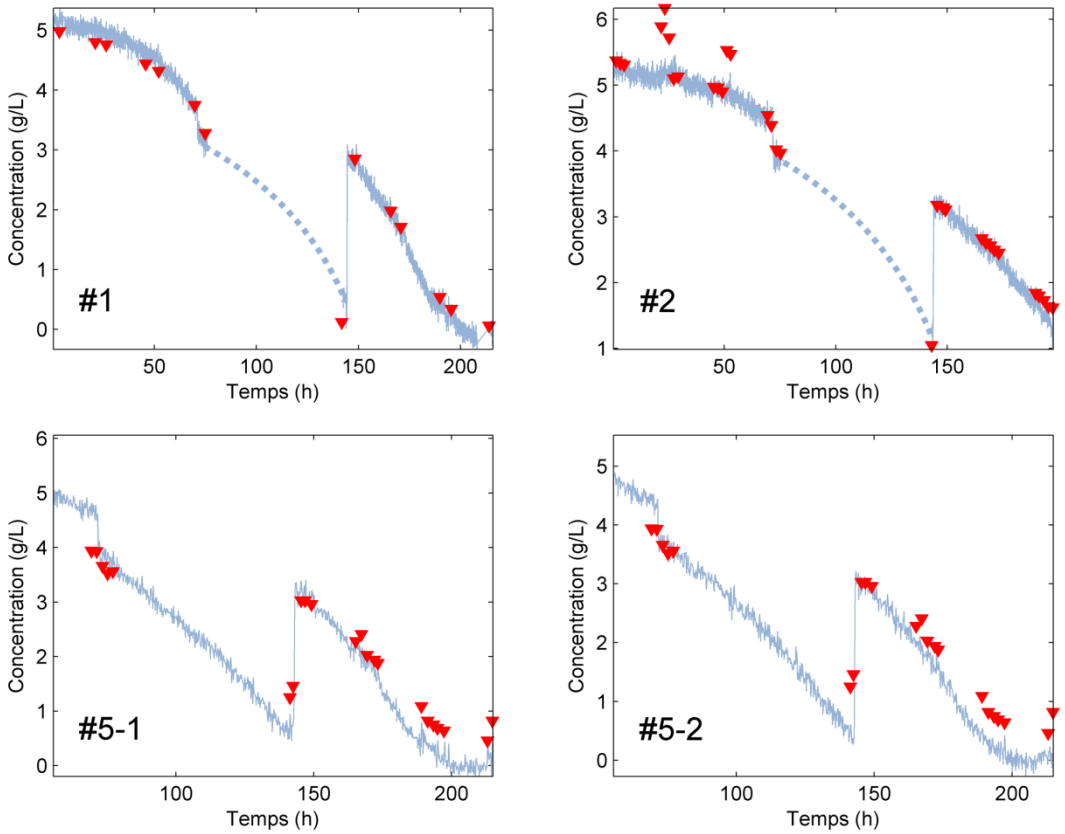

Figure 4.10 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour les concentrations en glucose. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges. Les absences de prédictions entre 76–142 h pour les cultures #1 et #2 et avant 56 h pour la culture #5 sont dues à des soucis techniques lors de l'acquisition Raman.

L'important saut de concentration observé sur chacun des profils traduit l'étape d'inoculation virale réalisée en diluant un quart de la culture cellulaire dans du nouveau milieu. Nous pouvons observer les erreurs pour les fortes concentrations en glucose appartenant au lot #2. Il apparaît alors que ces erreurs de prédiction proviennent en majeure partie des mesures de référence. En effet, nous pouvons observer des variations importantes dans les mesures en glucose entre 0–55 h sans raisons apparentes au sein du bioprocédé. Enfin, les écarts de prédiction observés pour les faibles concentrations

interviennent en fin de culture virale (fin du bioprocédé) et sont certainement dues une nouvelle fois aux mesures des prélèvements puisque ces valeurs se rapprochent de la limite de détection de la méthode de référence. Toutefois, en dehors de ces erreurs de prédiction, nous pouvons voir que le modèle est très satisfaisant et parvient tout-à-fait à capter la tendance décroissante de la concentration en glucose, métabolite consommé tout au long du bioprocédé.

En ce qui concerne les performances du modèle de régression développé pour les concentrations en glutamine, présentées Tableau 4.4, nous obtenons également des erreurs RMSEP (0,45 mM) et RMSECV (0,31 mM) satisfaisantes comparées à la gamme d'étalonnage (0,00–5,44 mM). Toutefois, l'écart existant entre ces deux erreurs indique la présence de certains écarts entre les concentrations prédites et les concentrations mesurées pour les échantillons des lots test. En effet, nous pouvons constater Figure 4.9b que les concentrations en glutamine des échantillons de l'ensemble test sont sous-évaluées, créant ainsi un biais pour les prédictions, responsable de la dégradation de l'erreur RMSEP. En étudiant les coefficients de régression du modèle glutamine, nous pouvons voir que ceux-ci sont similaires à ceux calculés pour le modèle glucose. La différence principale intervient autour de 850 cm^{-1}, bande associée aux élongations v(C–C) de la chaine carbonée de la molécule de glutamine [140]. Cette similarité entre les coefficients de régression avait déjà été observée lors des travaux sur l'optimisation du temps d'acquisition Raman (section 2.3.1 du Chapitre 3, Modèles pour les concentrations en glucose et glutamine, page 99). Celle-ci provient notamment du fait que les évolutions des deux paramètres biochimiques sont très proches au cours des cultures impliquant les cellules HeLa. C'est pourquoi les modèles reposent sur des bandes spectrales analogues.

Néanmoins, en prenant les prédictions en continu pour les cultures test (Figure 4.11), nous pouvons voir que les tendances décroissantes des concentrations en glutamine sont très bien représentées par le modèle PLS. Il apparait finalement que la majeure partie des écarts observés intervient d'une part aux faibles concentrations présentes en particulier en fin de culture, et d'autre part lors de la phase de culture cellulaire de la culture #1. Nous pouvons donc considérer que les concentrations des métabolites consommés lors des cultures mettant en jeu les cellules HeLa peuvent très bien être prédites à l'aide des modèles de régression réalisés sur la base de plusieurs *batches*, permettant donc d'intégrer les variations existantes entre les cultures.

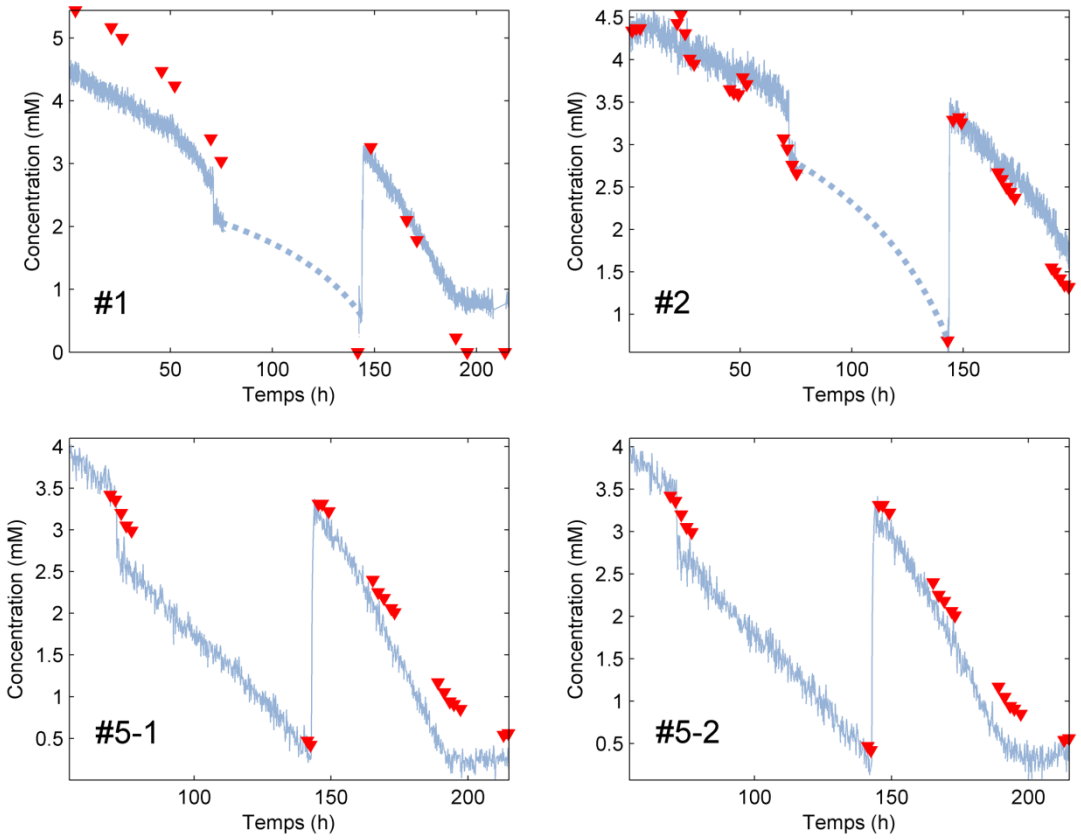

Figure 4.11 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour les concentrations en glutamine. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges. Les absences de prédictions entre 76–142 h pour les cultures #1 et #2 et avant 56 h pour la culture #5 sont dues à des soucis techniques lors de l'acquisition Raman.

1.2.2 Modèles pour les concentrations en glutamate, en lactate et en ammonium

Les paramètres biochimiques glutamate, lactate et ammonium sont des métabolites produits par les cultures de cellules HeLa. Toutefois, ces produits ne sont pas formés à concentrations égales durant le bioprocédé.

En effet, le glutamate est très peu présent durant les cultures. Nous pouvons voir Tableau 4.4 que la gamme de concentration n'évolue qu'entre 0,00–2,55 mM. Les signaux Raman caractéristiques de cette molécule seront donc peu intenses et les variations spectrales plus difficiles à percevoir. Ainsi, nous pouvons voir que la régression PLS parvient difficilement à corréler les concentrations mesurées à partir des méthodes de référence et les spectres Raman (coefficient de détermination R^2 calculé à 0,766, Tableau 4.4). Ceci apparait de manière évidente sur le graphique des concentrations prédites en fonction des concentrations mesurées, Figure 4.12a. En effet, nous pouvons clairement observer le manque de corrélation qui existe entre les mesures de référence et les prédictions réalisées,

ce qui induit une forte erreur de prédiction (RMSEP à 0,42 mM) par rapport à la gamme de concentration.

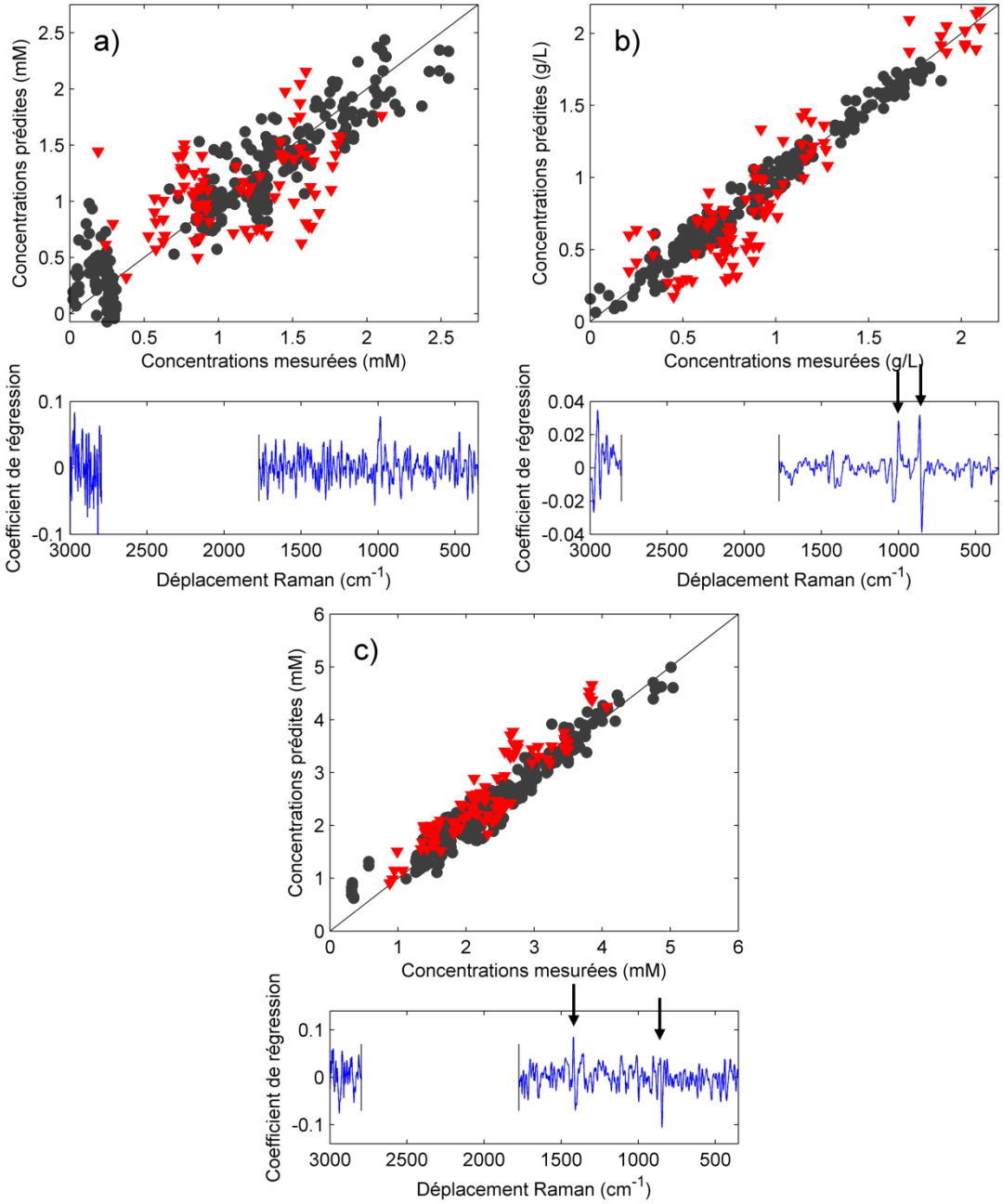

Figure 4.12 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules HeLa pour les concentrations a) en glutamate, b) en lactate et c) en ammonium. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Nous pouvons également noter la structure des coefficients de régression du modèle PLS calculé pour ces concentrations en glutamate. Ils présentent en effet un niveau de bruit très élevé, signifiant qu'il est difficile de ressortir les contributions spectrales dues au glutamate du bruit de fond spectral. En reprenant les spectres acquis en continu des différents lots test, et en y appliquant le modèle PLS calculé, nous obtenons une nouvelle

fois les prédictions réalisées durant toute la durée du bioprocédé. Les prédictions réalisées pour la concentration en glutamate sont disponibles Figure 4.13. Sur ces représentations, nous observons l'influence du bruit présent au sein des coefficients de régression. En effet, les prédictions présentent des variations très importantes d'un échantillon à l'autre, traduisant bien l'instabilité du modèle de régression et la difficulté à prédire la concentration en glutamate dans le milieu.

À propos du niveau de lactate dans le milieu, les performances obtenues sont satisfaisantes compte-tenu de la gamme de concentration mise en jeu. En effet, l'erreur RMSEP est estimée à 0,22 g/L pour une gamme variant entre 0,00 g/L et 1,89 g/L. Néanmoins, nous pouvons voir que l'erreur RMSECV approche les 0,08 g/L, ce qui reste très inférieur à l'erreur RMSEP calculée. Ceci traduit un certain manque de robustesse du modèle de régression à l'égard des échantillons des cultures test. Nous pouvons observer cet effet sur le graphique des concentrations en lactate prédites par rapport aux concentrations mesurées à l'aide de la méthode de référence, Figure 4.12b.

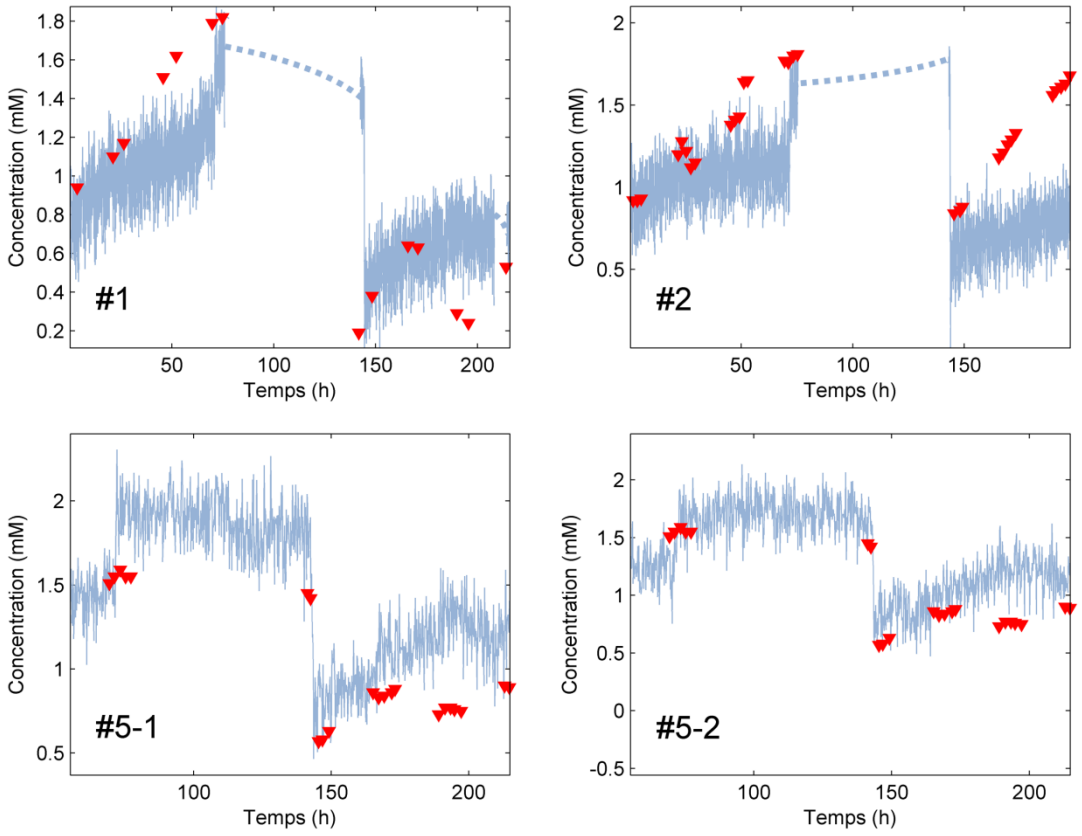

Figure 4.13 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour les concentrations en glutamate. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges. Les absences de prédictions entre 76–142 h pour les cultures #1 et #2 et avant 56 h pour la culture #5 sont dues à des soucis techniques lors de l'acquisition Raman.

Nous remarquons qu'une partie des échantillons test présentent un biais important. Il s'agit des concentrations comprises entre 0,5 g/L et 1 g/L qui correspondent finalement aux échantillons de la culture #2. Ceci apparait nettement sur les prédictions des concentrations en lactate réalisées sur les spectres acquis en continu pour cette culture, Figure 4.14. Nous pouvons voir que les prédictions sont systématiquement inférieures aux mesures de référence. Toutefois, nous notons que malgré ce biais, le modèle est tout-à-fait capable de suivre la tendance de la concentration en lactate tout au long de la culture #2. De plus, nous pouvons noter que les prédictions réalisées en continu pour les autres lots test parviennent très bien à suivre l'évolution particulière du paramètre métabolique lactate.

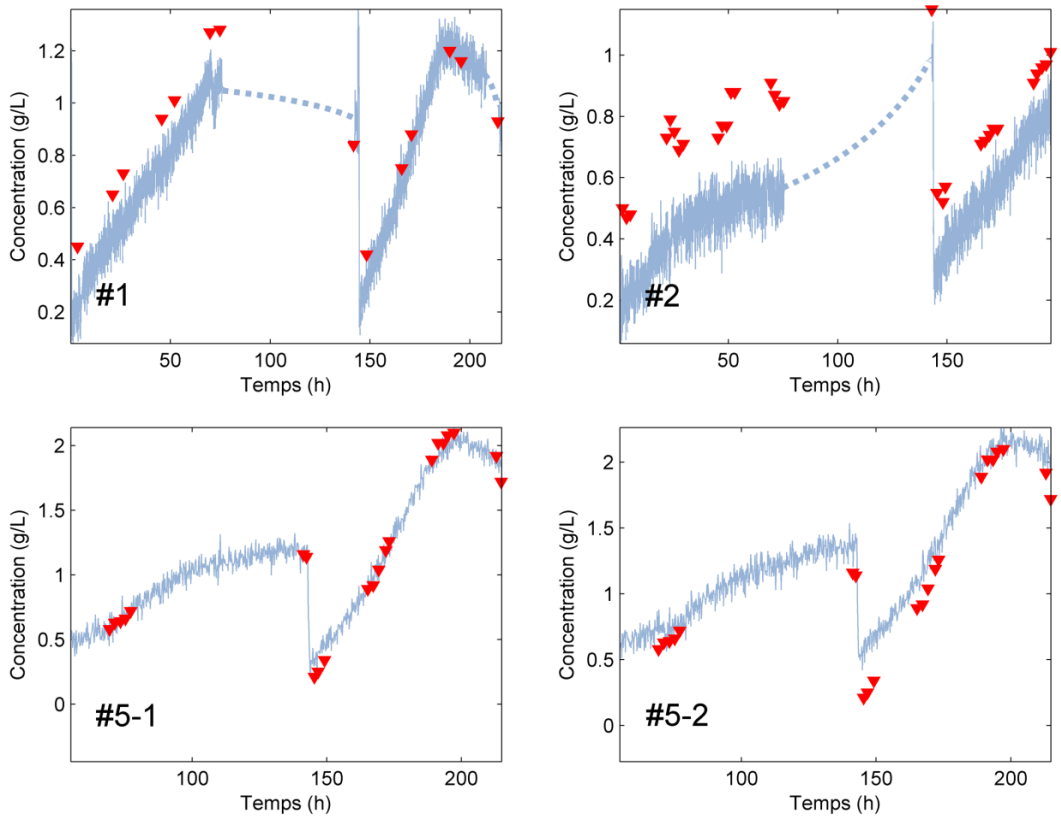

Figure 4.14 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour les concentrations en lactate. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges. Les absences de prédictions entre 76–142 h pour les cultures #1 et #2 et avant 56 h pour la culture #5 sont dues à des soucis techniques lors de l'acquisition Raman.

En reprenant les coefficients de régression du modèle PLS basé sur les concentrations en lactate, Figure 4.12b, deux bandes principales ressortent majoritairement. Tout d'abord, nous pouvons observer un signal important autour de 860 cm^{-1}, caractéristique des élongations $\nu(C–CO_2^-)$ du groupement carboxylate de la molécule, dont la bande Raman est très intense sur les spectres de lactate pur [138]. De plus, nous pouvons observer un second signal caractéristique de la molécule de lactate à travers les bandes autour de 1000 cm^{-1}. Celles-ci traduisent la présence des élongations $\nu(C–C)$ ainsi que les élongations $\nu(C–CH_3)$

du groupement méthyle présent en bout de chaine carbonée [138]. Nous pouvons donc dire que les modèles de régression basés sur les références en lactate reposent sur des bandes spectrales caractéristiques de la molécule mise en jeu.

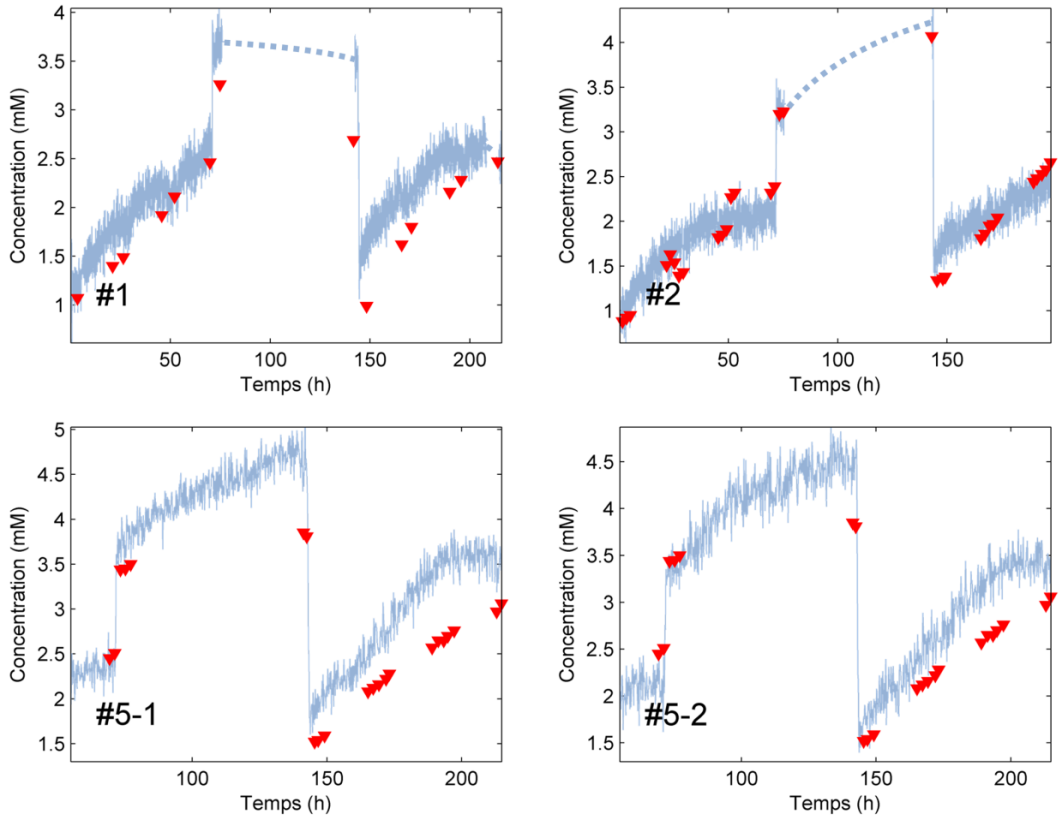

Figure 4.15 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour les concentrations en ammonium. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges. Les absences de prédictions entre 76–142 h pour les cultures #1 et #2 et avant 56 h pour la culture #5 sont dues à des soucis techniques lors de l'acquisition Raman.

En ce qui concerne l'ammonium, le modèle de régression développé présente des performances permettant de réaliser des prédictions satisfaisantes. En effet, l'erreur RMSEP (0,42 mM) reste faible compte-tenu de la gamme de concentration proposée (0,32–5,04 mM). Toutefois, nous pouvons encore noter les différences qui existent entre les erreurs RMSEP et RMSECV qui traduisent une nouvelle fois certaines erreurs lors des prédictions des échantillons des cultures test. En confrontant les représentations graphiques mettant en jeu les prédictions réalisées en continu sur les cultures test (Figure 4.15) et les concentrations prédites en fonction des concentrations mesurées (Figure 4.12c), nous pouvons remarquer que les erreurs de prédiction interviennent en majeure partie sur la fin de la phase de culture cellulaire (soit pour les plus fortes concentrations) et sur les phases de culture virale.

L'analyse des coefficients de régression est similaire à ce que nous avions obtenu lors de l'étude du temps d'acquisition optimal pour les spectres Raman (Chapitre 3, section 2.3.2,

Modèles pour les concentrations en glutamate, lactate et ammonium, page 100). Les signaux présents sur la région 1000–900 cm^{-1} traduisent bien la présence de molécules d'ammonium en solution aqueuse [141]. Toutefois, ces travaux ne permettent pas d'attribuer la bande autour de 1450 cm^{-1} à la molécule d'ammonium. Mais en prenant les prédictions en continu des cultures test Figure 4.15, nous pouvons voir que les modèles permettent de suivre l'évolution du biomarqueur de manière adéquat pendant toute la durée de la culture, que ce soit durant la phase de culture de cellules ou pendant la culture virale. En somme, nous pouvons conclure quant aux modèles de régression construits pour réaliser le suivi des métabolites formés durant les cultures impliquant les cellules HeLa qu'ils permettent de déterminer convenablement les concentrations des produits formés, à condition de présenter suffisamment de variabilité dans les concentrations. En effet, le paramètre glutamate, faiblement produit par les cellules HeLa est difficilement modélisable.

1.2.3 Modèles pour les densités VCD et TCD

Les modèles de régression permettant de prédire la densité en cellules vivantes VCD et la densité totale en cellules TCD sont plus difficiles à mettre en place car ces derniers ne peuvent être construits pour toute la durée du bioprocédé. En effet, il faut noter que les méthodes de référence employées pour mesurer les densités cellulaires diffèrent selon la phase de culture mise en jeu. Il existe ainsi une méthode de mesure pour les densités durant la culture cellulaire et une seconde pour les mesures suite à l'infection virale. Si nous développons un modèle de régression prenant en considération les deux phases de culture (cas de la densité VCD, Figure 4.16), nous observons alors des erreurs importantes dans les prédictions, notamment pour les fortes valeurs de densité.

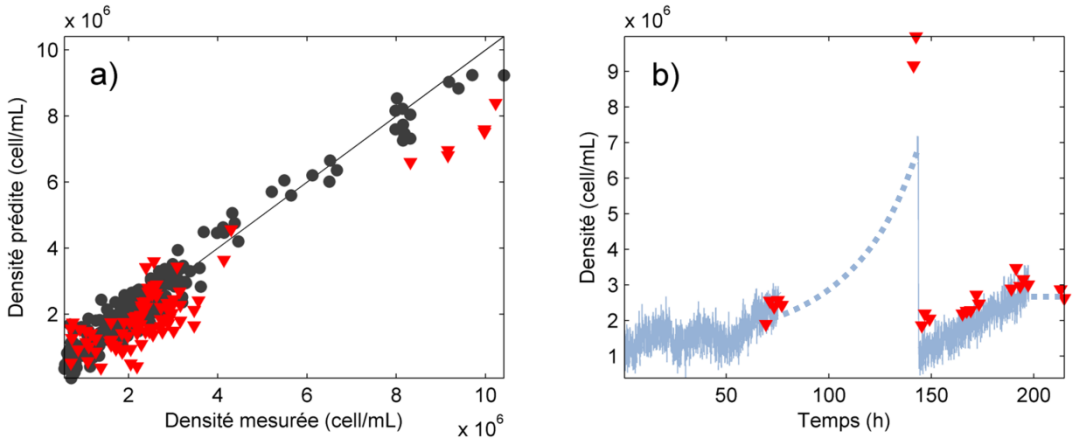

Figure 4.16 Représentations graphiques a) du modèle de régression développé sur l'ensemble des deux phases de culture et b) des prédictions réalisées sur les spectres acquis en continu de la culture #5-1 (à titre d'exemple). Les échantillons de calibration du modèle sont représentés par des points noirs, les échantillons de validation par des triangles rouges et les prédictions faites sur les spectres acquis en continu en bleu.

Il convient donc de séparer les échantillons en deux familles distinctes : pré- et post-inoculation, résultant donc en quatre modèles de régression dont les performances sont récapitulées Tableau 4.5.

Tableau 4.5 Récapitulatif des performances des modèles de régression pour les densités cellulaires, développés en cumulant les échantillons des cultures de cellules HeLa réalisées dans les conditions normales de culture, intégrant les phases pré- et post-inoculation virale.

Biomarqueur (unité)	R^2	LV	RMSECV	RMSEP	Gamme
VCD pré[1] (10^6 cell/mL)	0,984	7	0,49	0,42	0,56 – 10,41
TCD pré[1] (10^6 cell/mL)	0,984	7	0,49	0,42	0,57 – 10,66
VCD post[2] (10^6 cell/mL)	0,729	6	0,35	0,44	1,39 – 4,38
TCD post[2] (10^6 cell/mL)	0,762	6	0,40	0,48	1,58 – 4,85

[1] *Modèles calculés sur les mesures réalisées avant l'inoculation virale*

[2] *Modèles calculés sur les mesures réalisées après l'inoculation virale*

Dans un premier temps, nous étudierons les modèles de régression développés sur les échantillons acquis en phase pré-inoculation. En regardant d'abord les erreurs calculées pour les modèles PLS basés sur les références des biomarqueurs VCD et TCD, nous pouvons très clairement penser que ces modèles de régression sont à même de réaliser des prédictions fiables pour chacun des paramètres. En effet, les erreurs RMSEP des modèles pour chaque densité cellulaire ($0{,}42 \cdot 10^6$ cell/mL pour les deux modèles) sont inférieures aux erreurs RMSECV ($0{,}49 \cdot 10^6$ cell/mL pour les deux modèles également), traduisant seulement les faibles variations du jeu test par rapport au jeu de calibration. De plus, ces erreurs présentent des niveaux relativement faibles vis-à-vis des gammes de densité, variant entre $0{,}56 \cdot 10^6$ cell/mL et $10{,}41 \cdot 10^6$ cell/mL pour les densités VCD et $0{,}57 \cdot 10^6$ cell/mL et $10{,}66 \cdot 10^6$ cell/mL pour les densités TCD. Les représentations des modèles, à travers les graphiques des densités prédites en fonction des densités calculées (Figure 4.17) permettent également d'illustrer les qualités des modèles de régression.

Si nous considérons les coefficients de régression déterminés par les modèles PLS, nous pouvons clairement voir que ceux-ci présentes de très fortes similarités. Ceci s'explique par la très forte corrélation qui existe entre les évolutions de chacun des deux paramètres. En effet, puisque les cultures de cellules HeLa sont stoppées avant la sénescence de façon à obtenir le maximum de cellules potentiellement hôtes pour le virus, la mort cellulaire n'intervient pas durant la phase d'amplification cellulaire. Le niveau de viabilité reste donc élevé durant toute la durée pré-inoculation et les évolutions des biomarqueurs VCD et TCD sont identiques. Sur ces coefficients de régression, nous pouvons noter d'importants signaux autour de 1100 cm^{-1} traduisant les élongations $\nu(PO_2^-)$ des colonnes de macromolécules d'ADN des cellules [144]. Les bandes observées autour de 1000 cm^{-1} représentent les

élongations v(C–C) du cycle benzénique des molécules de phénylalanine présentent dans les cellules [145]. Enfin, les coefficients de régression situés entre 3000–2800 cm^{-1} et autour de 1400 cm^{-1} présentent respectivement les élongations v(C–H) et les déformations $\delta(CH_2)$ de plusieurs molécules différentes [145,146].

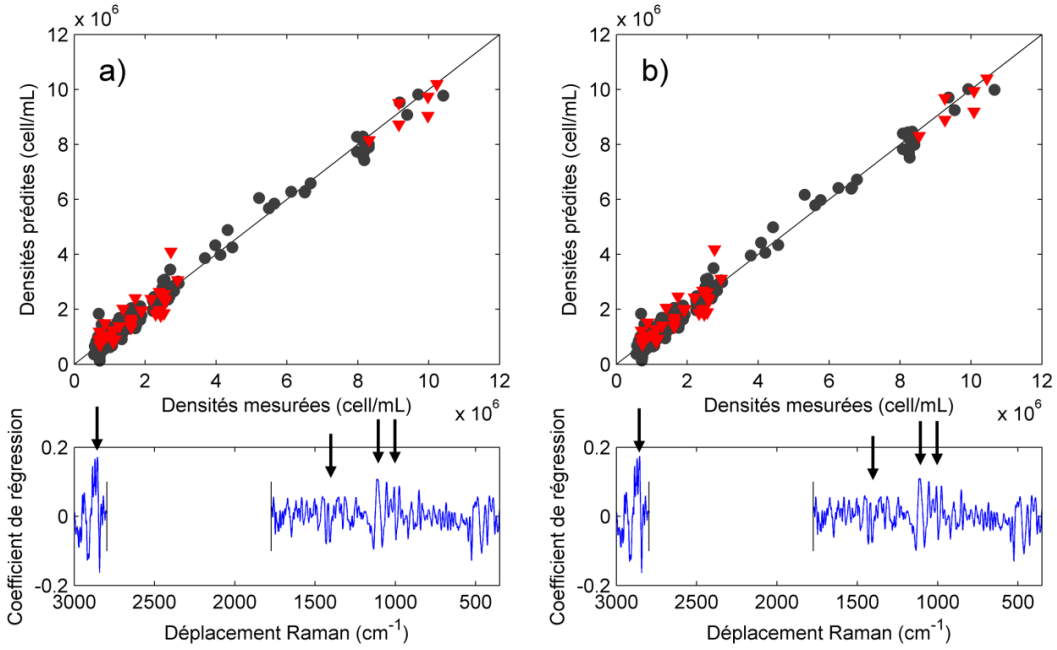

Figure 4.17 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules HeLa pour a) la densité en cellules vivantes et b) la densité en cellules totale avant l'inoculation du virus au sein du bioréacteur. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

En réalisant les prédictions en continu des cultures test, Figure 4.18, nous remarquons que nous sommes en mesure de suivre l'évolution exponentielle de la croissance cellulaire. Il apparait donc que les modèles de régression construits pour les différentes densités cellulaires mesurées avant l'infection virale permettent de réaliser des prédictions satisfaisantes.

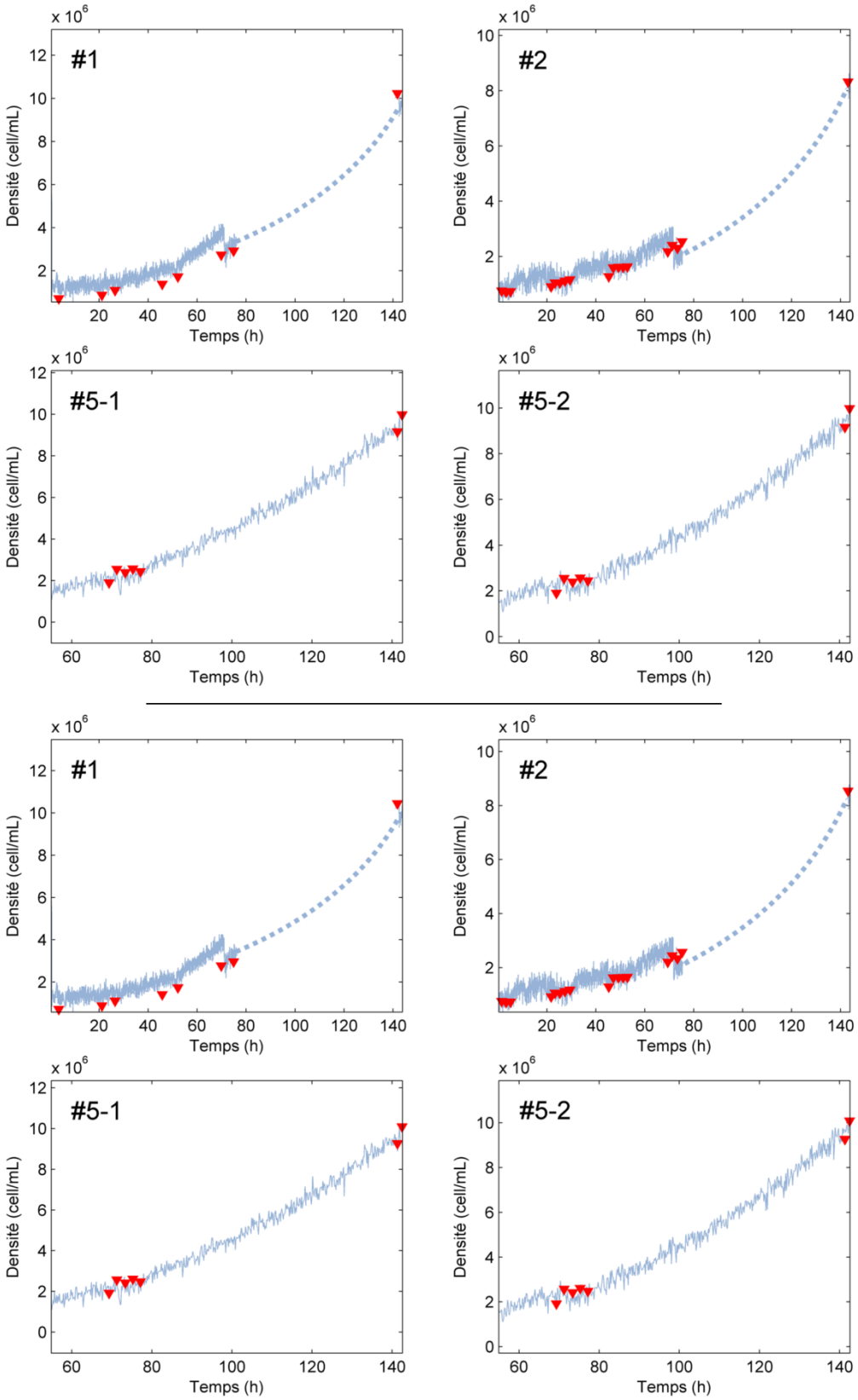

Figure 4.18 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour la densité VCD (en haut) et la densité TCD (en bas) mesurées avant l'inoculation du virus dans le milieu de culture. Les densités obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges. Les absences de prédictions entre 76–142 h pour les cultures #1 et #2 et avant 56 h pour la culture #5 sont dues à des soucis techniques lors de l'acquisition Raman.

En ce qui concerne les densités mesurées suite à l'infection virale, nous pouvons d'abord remarquer que les gammes de variation sont bien moins importantes. En effet, les densités varient entre $1{,}39 \cdot 10^6$ cell/mL et $4{,}38 \cdot 10^6$ cell/mL pour les densités en cellules vivantes VCD et entre $1{,}58 \cdot 10^6$ cell/mL et $4{,}85 \cdot 10^6$ cell/mL, ce qui correspond à moins de la moitié des gammes de variation des densités en phase pré-inoculation. Ceci s'explique par la présence de virus dans la culture induisant un taux de mort cellulaire plus important, inhibant donc la croissance des cellules. Les performances des modèles de régression obtenus sur les densités mesurées durant la culture virale, Tableau 4.5, permettent de montrer qu'il est plus difficile de suivre ces biomarqueurs. En effet, bien que les erreurs de prédiction RMSEP soient mesurées à $0{,}44 \cdot 10^6$ cell/mL pour la densité VCD et $0{,}48 \cdot 10^6$ cell/mL pour la densité TCD, les coefficients de déterminations R^2 ne dépassent pas 0,750, traduisant donc un manque de corrélation entre les spectres et les mesures de référence lors de la modélisation PLS. Les représentations graphiques des densités prédites en fonction des densités mesurées, Figure 4.19, permettent de bien représenter ce manque de corrélation, notamment pour les prédictions des densités mesurées.

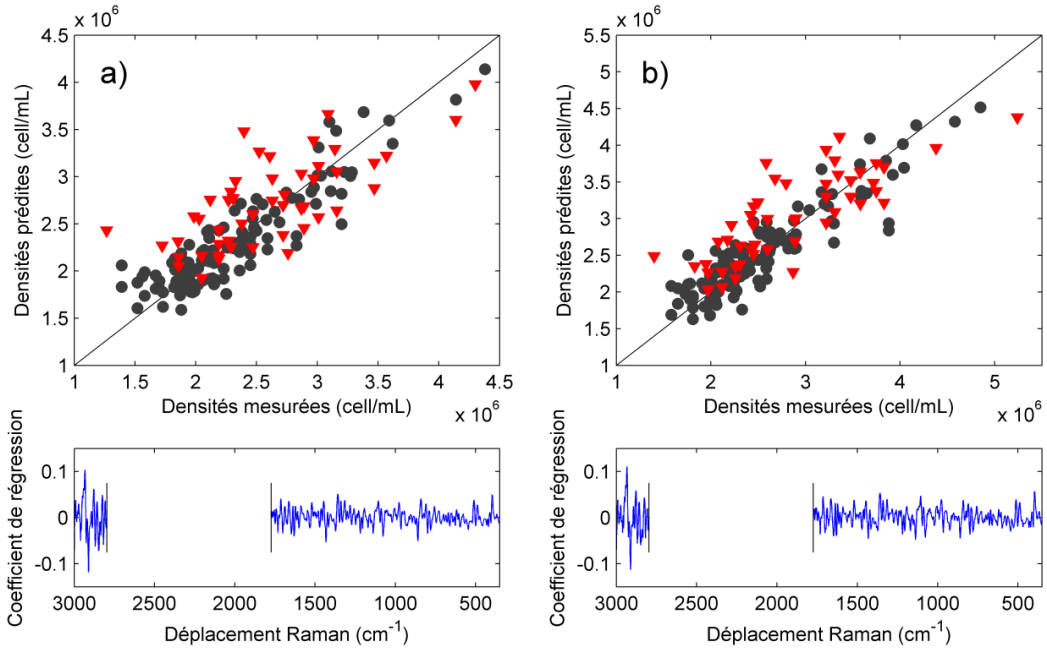

Figure 4.19 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs cultures de cellules HeLa pour a) la densité en cellules vivantes et b) la densité totale en cellules après l'inoculation du virus au sein du bioréacteur. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

À travers les coefficients de régression, nous pouvons observer que leurs structures diffèrent de ce que nous avions pré-inoculation puisque les signaux importants précédemment observés présentent moins d'intensité ici, les structures étant plus bruitées. Notons qu'une nouvelle fois, les coefficients des deux paramètres biochimiques sont similaires. Ceci s'explique à nouveau par les évolutions semblables des deux densités. En

effet, les débris cellulaires résultant de la lyse induite par le virus ne sont pas considérés en tant que cellules à part entière par la méthode de référence. Ainsi, les paramètres VCD et TCD évoluent de façon similaire aux tendances présentées avant l'infection virale, mais sans forte croissance à cause du virus.

Si nous prenons les prédictions réalisées sur les échantillons des cultures test, Figure 4.20, nous pouvons voir que les modèles de régression parviennent à déceler les tendances des niveaux cellulaires. En effet, nous pouvons voir que la croissance des densités cellulaires apparait clairement, présentant même les inflexions des paramètres en fin de culture pour les lots #1 et #5. De plus, sur ces représentations graphiques, nous pouvons noter les fortes variations des mesures de référence de ces cultures, traduisant des difficultés à déterminer les densités en cellules, que ce soit VCD et TCD, lorsque le virus est mis en jeu. Ces variations apparaissent également sur les mesures de référence ayant servi à construire le modèle de régression, ce qui explique le manque de corrélation lors de la régression PLS.

En somme, nous pouvons finalement conclure que les modèles de régression basés sur les densités VCD et TCD présentent des performances satisfaisantes. En ce qui concerne la phase pré-inoculation virale, nous avons pu montrer à travers leurs performances prédictives que les modèles PLS cumulant les échantillons de plusieurs cultures permettaient de bien intégrer les variations inter-*batches*. Quant aux données recueillies sur la phase de culture virale, nous avons pu développer des modèles de régression acceptables malgré les conditions défavorables rencontrées. En effet, les densités cellulaires présentent de fortes erreurs de mesure lors de cette phase du bioprocédé qui influent inévitablement sur les performances des modèles PLS développés. Néanmoins, nous avons pu montrer que les prédictions réalisées lors de cette période de culture virale parvenaient à suivre les tendances croissantes des densités.

Finalement, nous avons montré dans cette section que les modèles de régression PLS développés pour les suivis de paramètres biochimiques en cumulant les échantillons acquis sur plusieurs cultures impliquant les cellules HeLa présentaient des performances prédictives satisfaisantes dans la majeure partie des cas. Le fait d'utiliser plusieurs lots différents au sein du jeu test permet d'appuyer la validité de l'intégration des variations inter-cultures au sein des modèles PLS, présentant donc finalement plus de robustesse que ce que nous avons pu développer lors du chapitre précédent pour l'étude du temps d'acquisition Raman.

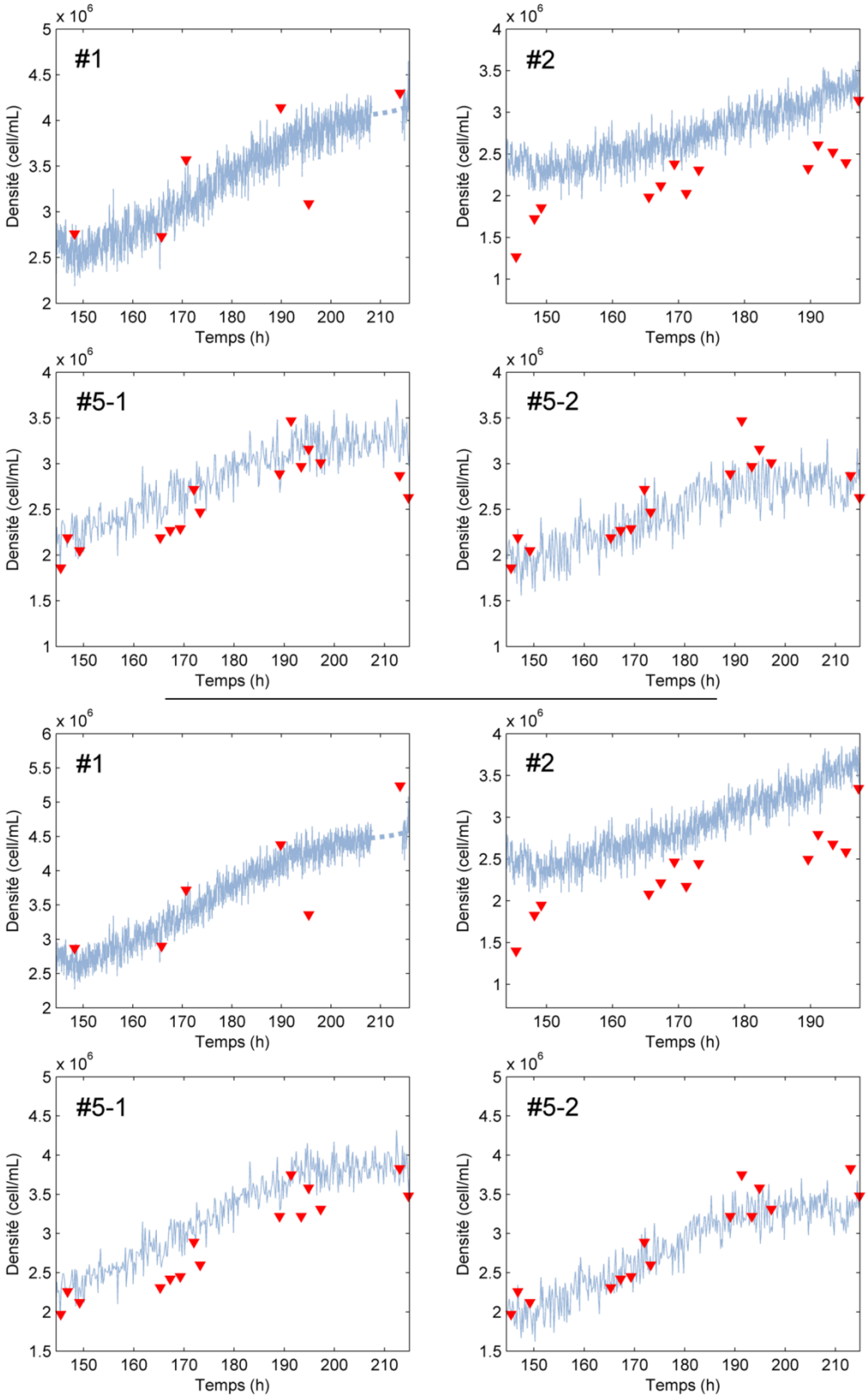

Figure 4.20 Représentations des prédictions réalisées sur les spectres des cultures de cellules HeLa test acquis en continu (en bleu) pour la densité VCD (en haut) et la densité TCD (en bas) mesurées après l'inoculation du virus dans le milieu de culture. Les densités obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges.

1.3 Modèles de régression pour les cellules Sf9

Le principe de fonctionnement des cellules Sf9 est relativement similaire à celui des cellules HeLa, au détail près que les phases d'amplification cellulaire et de culture virale sont réalisées séparément et non pas en continu tels que pour les protocoles de cultures virales employant les cellules HeLa. Nous disposons de six lots impliquant les cellules Sf9 : deux lots d'amplification cellulaire et quatre cultures virales. Le nombre d'échantillons pour chacun des lots est détaillé Tableau 4.6. Nous pouvons noter ici que plusieurs sondes ont été mises en jeu pour les différents suivis, et que tout comme certaines cultures HeLa, les lots #1 et #3 font intervenir deux sondes à immersion différentes pour un seul et même suivi.

Nous comptons finalement un total de 412 échantillons dont 135 pour les amplifications cellulaires et 277 pour les cultures virales. Les spectres de chacun des 412 échantillons sont prétraités de la même façon que pour les études concernant les cultures impliquant les cellules CHO et HeLa, à savoir une dérivée première de Savitzky-Golay suivi d'une normalisation SNV, puis une sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Nous réalisons dans un premier temps l'étude ACP des données disponibles. La Figure 4.21 représente les résultats obtenus pour cette analyse à travers les graphiques des *scores* des quatre premières composantes principales (soit plus de 90 % de la variance totale), accompagnés des représentations des *loadings* associés.

Tableau 4.6 Récapitulatif du nombre d'échantillons pour chaque culture mettant en jeu les cellules Sf9 employée afin de développer les modèles de régression intégrant les variations inter-*batches*. Les lots #1 et #3 mettent en jeu deux sondes différentes pour un même suivi.

Numéro de lot	Nombre d'échantillons	Numéro de sonde
Amplification cellulaire		
#1-1	42	1
#1-2	43	2
#2	50	2
Culture virale		
#3-1	42	3
#3-2	48	4
#4	63	1
#5	61	2
#6	63	2

Sur ces représentations, nous pouvons distinctement observer que malgré les prétraitements réalisés, les échantillons des différents lots sont séparés par les premières composantes principales en différents ensembles. La composante PC1 (70,25 % de la

variance totale) sépare l'ensemble des données en deux groupes : d'une part les sondes n°1 et n°4, d'autres part les sondes n°2 et n°3. En prenant en compte les *loadings* de cette composante, nous pouvons voir que la composante repose principalement sur la bande spectrale située à 400 cm^{-1}. Celle-ci représente la fenêtre de saphir située à l'extrémité de la sonde (voir Figure 2.4, page 59 pour un schéma détaillé d'une sonde à immersion), ce qui traduit simplement des différences physiques entre les différentes sondes employées. Le fait d'associer les échantillons acquis à l'aide des différentes sondes permet donc d'intégrer ces variations aux modèles de régression. Néanmoins, il faut noter que la bande à 400 cm^{-1} peut présenter un intérêt chimique pour la modélisation des biomarqueurs d'intérêt, ce pourquoi nous conservons cette région spectrale dans la sélection pour le développement des modèles de régression.

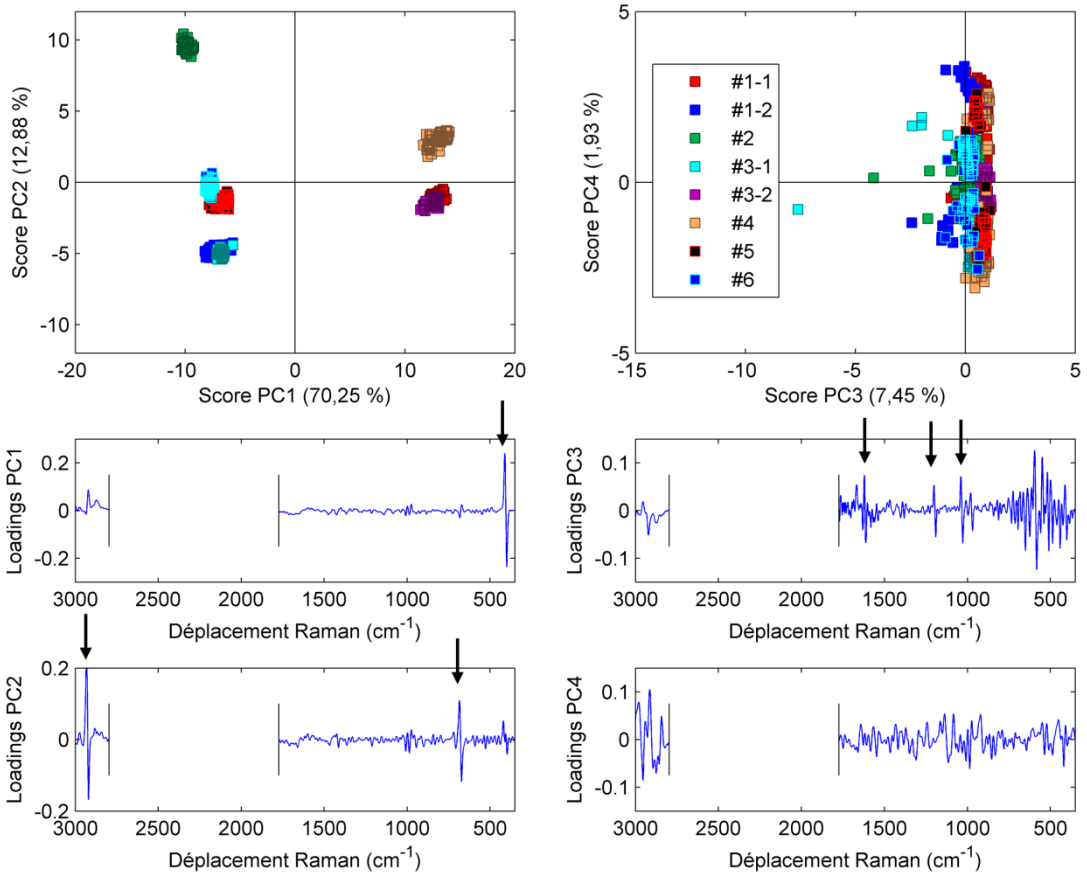

Figure 4.21 Résultats obtenus pour l'analyse ACP réalisée sur un jeu de données cumulant les échantillons provenant de six cultures impliquant les cellules Sf9 réalisées dans les conditions normales de culture. Les phases d'amplification cellulaire et de culture virale apparaissent toutes deux sur ces graphiques.

La deuxième composante principale (12,88 %) sépare les différents lots mis en jeu dans cette analyse en composantes principales. Les *loadings* de cette composante s'appuient majoritairement sur la bande à 2900 cm^{-1} caractérisant les vibrations v(C–H) de différentes molécules présentes en solution [146]. Enfin, la composante PC3 (7,45 %) discrimine

certains échantillons affectés par la lumière extérieure de l'environnement, tandis la composante PC4 (1,93 %) représente l'évolution des différentes cultures au cours du temps.

Afin de développer et valider les modèles de régression pour les cultures impliquant les cellules Sf9, il convient de prendre en compte plusieurs critères pour le développement des ensembles de calibration et test. Tout d'abord, le jeu test doit nécessairement représenter à la fois les cultures de cellules et les cultures de virus. Autre point important, nous ne devons pas dissocier les lots provenant d'une même culture, ce qui signifie que nous ne pourrons pas séparer les lots #1-1 et #1-2 ainsi que les lots #3-1 et #3-2, évitant ainsi de fausser les erreurs de prédiction calculées pour les modèles de régression en intégrant en partie les informations du *batch* dans les différents ensembles (calibration et test). Nous sélectionnons alors le lot #2 pour les amplifications cellulaires et le lot #4 pour les cultures virales dans l'ensemble test. Ainsi, nous conservons 299 échantillons en calibration et 113 échantillons en tant que jeu test. Tout comme pour les travaux réalisés pour les autres types de cellule, les modèles de régression sont validés une première fois (*10-fold cross-validation*) ce qui permet de déterminer les coefficients des modèles, dont notamment l'erreur RMSECV et le nombre de variables latentes. Ces modèles sont ensuite appliqués à l'ensemble test pour évaluer leurs performances prédictives via l'erreur RMSEP. Les performances des modèles PLS développés pour cette section sont récapitulées Tableau 4.7.

Tableau 4.7 Récapitulatif des performances des modèles de régression pour les concentrations métaboliques, développés en cumulant les échantillons des cultures de cellules HeLa réalisées dans les conditions normales de culture, intégrant les phases pré- et post-inoculation virale.

Biomarqueur (unité)	R^2	LV	RMSECV	RMSEP	Gamme
Glucose (g/L)	0,839	5	0,49	0,95	4,10 – 11,80
Glutamine (mM)	0,582	7	0,49	3,57	1,63 – 10,53
Glutamate (mM)	0,333	5	0,58	1,23	9,28 – 13,06

Étant donné que les cellules Sf9 sont très peu productrices de lactate, nous n'étudions pas ce métabolite puisque les concentrations mesurées sont toutes sous la limite de quantification de la méthode de référence. De plus, tout comme lors des travaux sur la détermination du temps optimal d'acquisition Raman (Chapitre 3, section 2.4, Temps d'acquisition pour les cultures de cellules Sf9, page 104), nous n'étudions pas la concentration en ammonium.

1.3.1 Modèles pour les concentrations en glucose, glutamine et glutamate

Nous commençons l'étude des suivis de cultures impliquant les cellules Sf9 en évaluant les modèles de régression s'appuyant sur les concentrations des métabolites principaux, à

commencer par le modèle PLS basé sur les concentrations en glucose. L'erreur RMSEP de ce modèle, évaluée à 0,95 g/L, traduit des capacités prédictives satisfaisantes comparées à la gamme de concentration comprise entre 4,10–10,80 g/L. Nous pouvons toutefois noter que l'évolution du paramètre n'est pas aussi importante que pour les autres types de cellule (soit 2,25–62,13 mM ou 0,41–11,19 g/L pour les cellules CHO, 0,00–6,20 g/L pour les cellules HeLa), notamment parce que les cellules Sf9 requièrent peu de ressources pour favoriser la multiplication cellulaire. Le graphique des concentrations prédites en fonction des concentrations de référence mesurées est représenté Figure 4.22a.

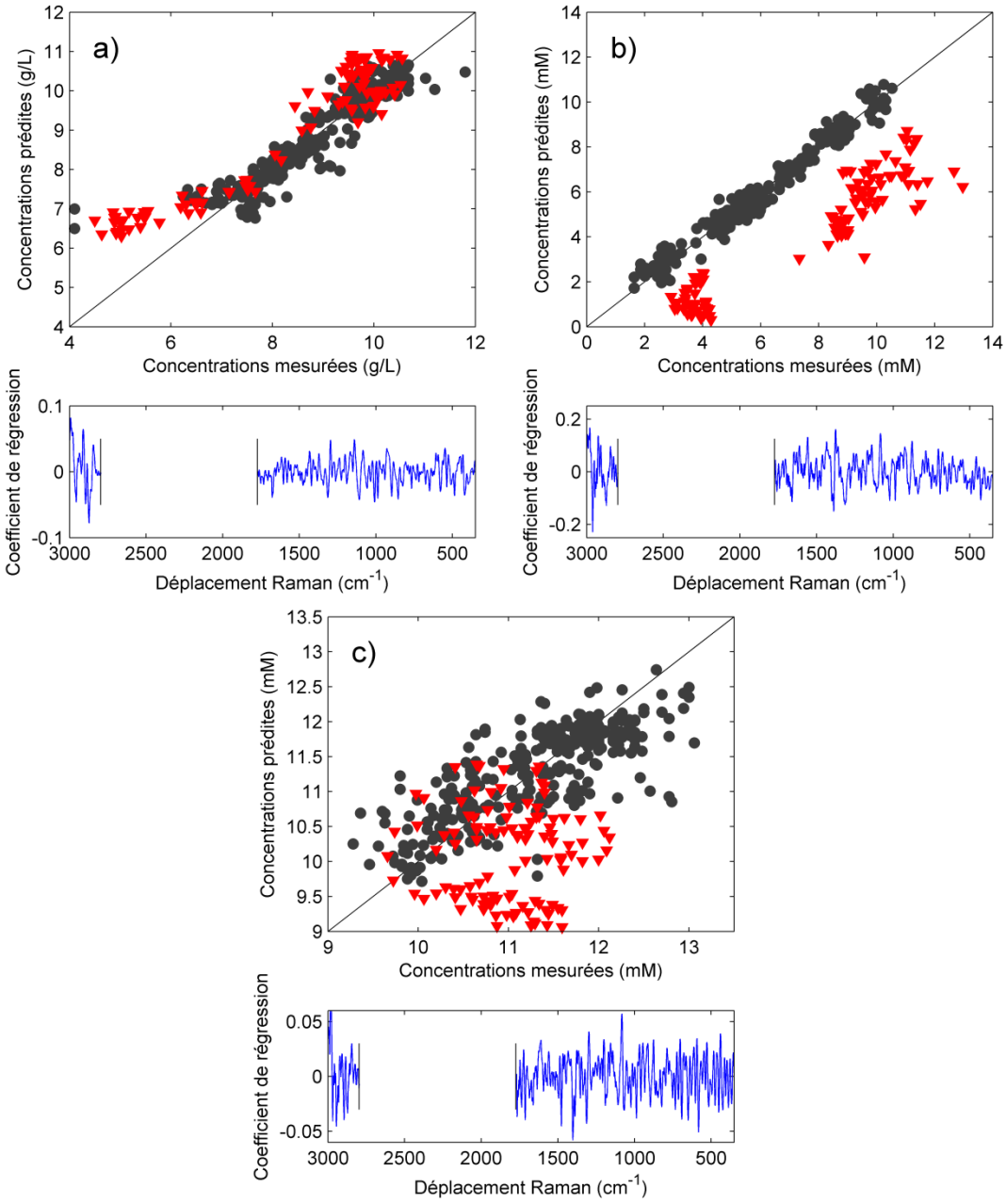

Figure 4.22 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs lots d'amplifications cellulaires et de cultures virales mettant en jeu les cellules Sf9 pour les concentrations a) en glucose, b) en glutamine et c) en glutamate. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Nous pouvons voir qu'en dépit d'une erreur de prédiction RMSEP satisfaisante, le modèle PLS présente de nombreux écarts que ce soit pour les prédictions durant la validation croisée pour les échantillons de calibration ou pour les prédictions des échantillons test, notamment à cause d'un niveau de bruit important sur les coefficients de régression. De ce fait, nous pouvons voir que les prédictions ne sont pas totalement fiables à travers le graphique des concentrations prédites en fonction des mesures de références (Figure 4.22a). Nous pouvons notamment observer des résidus importants sur les concentrations les plus faibles, comprises entre 4,00 g/L et 6,00 g/L. En prenant les spectres acquis en continu pour chaque culture test et en y appliquant le modèle de régression des concentrations en glucose, nous obtenons de nouvelles représentations permettant d'évaluer les capacités prédictives du modèle. Nous disposons de 711 spectres acquis en continu pour la culture #2 (amplification cellulaire) et 1919 spectres acquis en continu pour la culture #4 (culture virale). Les prédictions des concentrations en glucose pour ces spectres acquis en continu sont présentées Figure 4.23a. Tout d'abord, nous pouvons voir que les mesures de référence déterminées pour la culture #2 présentent de fortes variations tout au long de la culture, évolution inattendue pour la teneur en glucose durant une phase d'amplification cellulaire. Au contraire, les prédictions en continu présentent une décroissance plus représentative de l'évolution du biomarqueur. Nous pouvons toutefois noter la présence de variations brusques dans les prédictions, notamment autour de 23 h, 32 h, 48 h et 56 h de culture. Ceci s'explique simplement par la pollution des spectres de ces régions par la lumière extérieure, phénomène montré lors de l'étude ACP sur tous les lots (Figure 4.21, composante PC3). Nous pouvons ensuite noter que les écarts observés aux faibles concentrations sont présents sur les prédictions du lot #4 simplement parce que l'ensemble de calibration ne dispose que de deux mesures de référence pour la concentration en glucose entre 4,00 g/L et 6,00 g/L, ce qui est insuffisant pour être représentatif de cette gamme. Ainsi, bien que le modèle capte la tendance décroissante de la concentration en glucose, les prédictions sont surestimées par rapport aux mesures effectuées par la méthode de référence. En fin de compte, nous pouvons considérer que le modèle calculé pour prédire les concentrations du glucose manque de robustesse puisque celui-ci ne repose que sur seulement quatre cultures dont une seule représente la phase d'amplification cellulaire. Les coefficients de régression présentent donc un niveau de bruit élevé qui pourrait être réduit en accumulant plus de cultures.

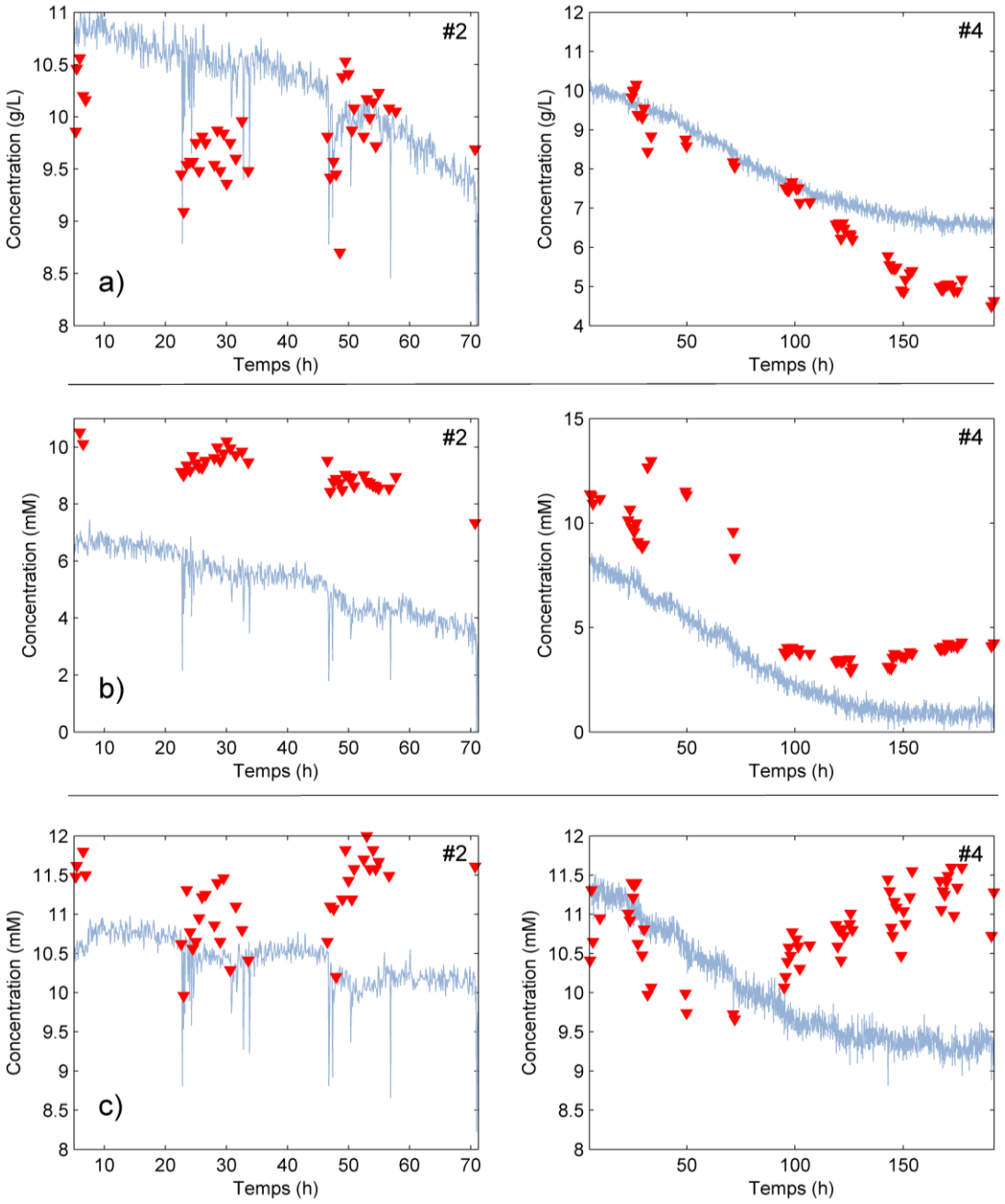

Figure 4.23 Représentations des prédictions réalisées sur les spectres des cultures de cellules Sf9 test acquis en continu (en bleu) pour les concentrations a) en glucose, b) en en glutamine et c) en glutamate. Les concentrations obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges.

En ce qui concerne le modèle de régression basé sur les concentrations en glutamine, nous pouvons directement noter l'erreur RMSEP (3,57 mM) très importante comparée à l'erreur RMSECV (0,49 mM) pour une gamme de variation relativement importante (1,63–10,53 mM), Tableau 4.7. Ceci s'explique simplement en prenant le graphique des concentrations prédites en fonction des concentrations mesurées, Figure 4.22b. Nous pouvons clairement observer un biais de prédiction important pour la totalité des échantillons de l'ensemble test. En effet, le biais calculé ici est de l'ordre de 3,39 mM. Cependant, le fait

que l'erreur RMSECV soit faible et que les prédictions des échantillons de calibration obtenus via la validation croisée soient satisfaisant, nous pouvons penser que le modèle de régression manque seulement de robustesse. Les coefficients de régression (Figure 4.22b) permettent d'appuyer ceci en présentant un fort niveau de bruit, tout comme les coefficients de régression du modèle de régression basé sur les mesures de référence de la concentration en glucose.

En considérant les prédictions obtenues à partir de l'application du modèle de régression sur les spectres acquis en continu pour les cultures test, nous observons pleinement les biais notés sur les graphiques précédents. En effet, bien que les prédictions réalisées suivent une tendance décroissante, attendue pour l'évolution de la concentration en glutamine durant les cultures impliquant les cellules Sf9, nous pouvons voir que les concentrations sont sous-évaluées, que ce soit pour la culture de cellules test ou la culture virale test. Notons une nouvelle fois la présence des brusques variations dans les prédictions réalisées pour la culture #2, qui, pour rappel, traduisent la présence de signaux provenant de la lumière extérieure sur les spectres Raman. En somme, les résultats obtenus par le modèle de régression pour les concentrations en glutamine indiquent que le modèle de régression manque de robustesse. En effet, les observations sont similaires à ce que nous avions pour les concentrations en glucose, présentant tout de même des biais plus importants ici. Ces-derniers sont principalement dus au manque de robustesse des modèles qui ne sont élaborés que sur quatre cultures différentes.

Le dernier métabolite mis en jeu pour les cultures Sf9 est le glutamate. En reprenant les performances obtenues par le modèle de régression, il apparait immédiatement que le modèle de régression ne parvient à relier les spectres aux mesures de référence. En effet, l'erreur de prédiction RMSEP traduisant les capacités prédictives du modèle PLS est importante (1,23 mM) compte tenu de l'erreur RMSECV (0,58 mM). Nous pouvons surtout noter les très faibles changements de la gamme d'étalonnage, entre 9,28–13,06 mM, ce qui ne représente que 3,78 mM de variation. La molécule de glutamate contribue bien aux spectres Raman puisque celle-ci est présente en concentration suffisamment élevée pour être détectée. Cependant, le fait de varier si peu implique un manque de représentativité de la molécule au sein de la variance générale des spectres Raman. En prenant le graphique des concentrations prédites en fonction des concentrations mesurées, Figure 4.22c, nous pouvons observer le manque de corrélation qui existe pour développer le modèle de régression, le coefficient de détermination étant estimé à 0,333.

Si nous considérons les prédictions réalisés par ce modèle PLS sur les spectres acquis en continu pour les cultures test, Figure 4.23c, nous pouvons très bien voir que le modèle ne

parvient pas du tout à capter les tendances de la concentration en glutamate durant les cultures test. Tout d'abord, pour le lot #2, bien que les mesures de référence disposent une nouvelle fois de valeurs instables, les prédictions obtenues présentent une faible décroissance le long de la culture alors que le glutamate est un produit formé par les cellules. Nous devrions obtenir une faible croissance de la concentration. D'autre part, les concentrations prédites pour la culture virale #4 ne suivent pas les mesures de référence réalisées. Le modèle de régression glutamate ne présente donc pas uniquement un manque de robustesse, tels que ceux développés pour les métabolites glucose et glutamine, mais un manque de corrélation entre les spectres et les mesures de référence ayant pour principale cause le manque de variation du métabolite au cours des cultures impliquant les cellules Sf9.

1.3.2 Modèles pour les densités VCD et TCD

Tout comme lors des travaux sur les modèles de régression basés sur les densités cellulaires cumulant les données de plusieurs cultures impliquant les cellules HeLa, les modèles de régression des densités VCD et TCD sont séparés en deux ensembles distincts : amplification cellulaire et culture virale. Les performances des quatre modèles PLS développés pour les densités cellulaires sont répertoriées Tableau 4.8.

Tableau 4.8 Récapitulatif des performances des modèles de régression pour les densités cellulaires, développés en cumulant les échantillons des cultures de cellules HeLa réalisées dans les conditions normales de culture, intégrant les phases pré- et post-inoculation virale.

Biomarqueur (unité)	R^2	LV	RMSECV	RMSEP	Gamme
VCD pré[1] (10^6 cell/mL)	0,579	5	0,36	3,03	0,85 – 8,03
TCD pré[1] (10^6 cell/mL)	0,744	5	0,38	2,39	0,94 – 9,30
VCD post[2] (10^6 cell/mL)	0,810	9	0,34	1,05	0,68 – 3,95
TCD post[2] (10^6 cell/mL)	0,874	7	0,22	1,20	1,41 – 5,39

[1] *Modèles calculés sur les mesures réalisées pour les amplifications cellulaires*

[2] *Modèles calculés sur les mesures réalisées pour les cultures virales*

Puisque les deux phases du protocole de culture impliquant les cellules Sf9 sont séparées, les modèles PLS pour l'amplification cellulaire ne reposent que sur une seule culture, soit la culture #1 (disposant de deux sondes). Nous pouvons alors remarquer directement, d'après les performances des modèles PLS pour les densités VCD et TCD, que ces-derniers présentent de fortes erreurs de prédiction RMSEP ($3{,}03 \cdot 10^6$ cell/mL et $2{,}39 \cdot 10^6$ cell/mL respectivement) par rapport aux gammes de densité étudiées ($0{,}85 \cdot 10^6$–$8{,}03 \cdot 10^6$ cell/mL pour la densité VCD, $0{,}94 \cdot 10^6$–$9{,}30 \cdot 10^6$ cell/mL pour la densité TCD). En prenant les graphiques des densités prédites en fonction des densités mesurées à l'aide des méthodes de référence, Figure 4.24, nous pouvons voir qu'il s'agit de l'existence d'un biais de

prédiction important sur les échantillons des cultures test, biais estimés à $3{,}08 \cdot 10^6$ cell/mL et $2{,}35 \cdot 10^6$ cell/mL pour les paramètres VCD et TCD respectivement.

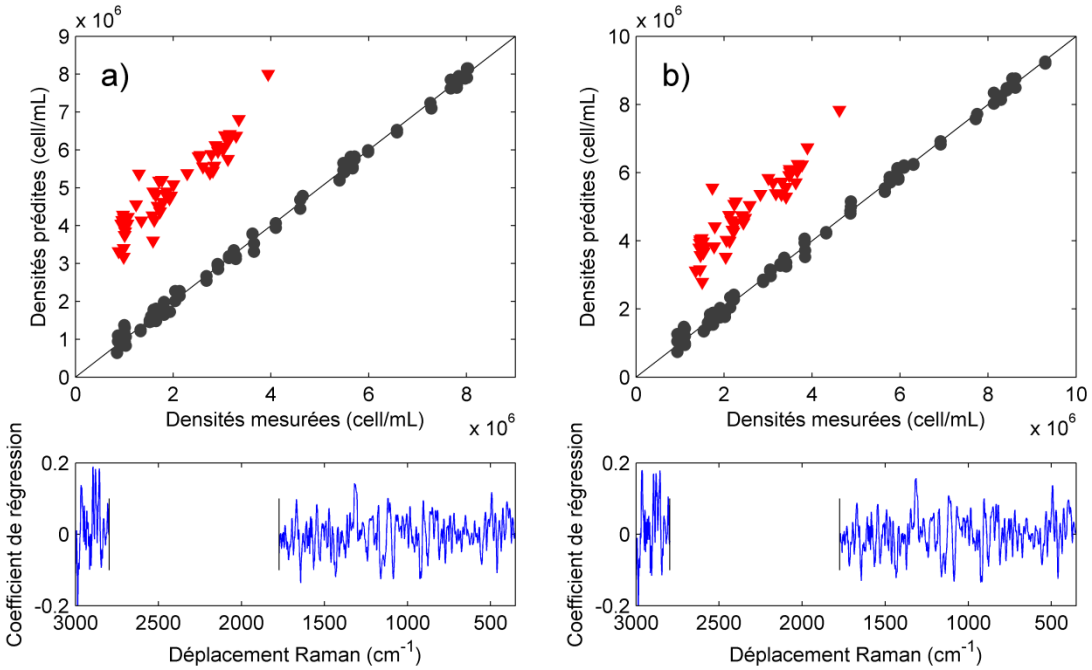

Figure 4.24 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs lots d'amplifications cellulaires mettant en jeu les cellules Sf9 pour a) la densité en cellules vivantes et b) la densité en cellules totale. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

En prenant les prédictions réalisées par ces modèles sur les échantillons de la culture #2, Figure 4.25, nous pouvons observer des tendances analogues pour les paramètres VCD et TCD, ce qui est cohérent durant la phase d'amplification cellulaire avant l'inoculation du virus. De plus, sans compter le biais existant, nous pouvons voir que les prédictions des densités réalisées en continu suivent les mêmes tendances que les mesures réalisées par la méthode de référence.

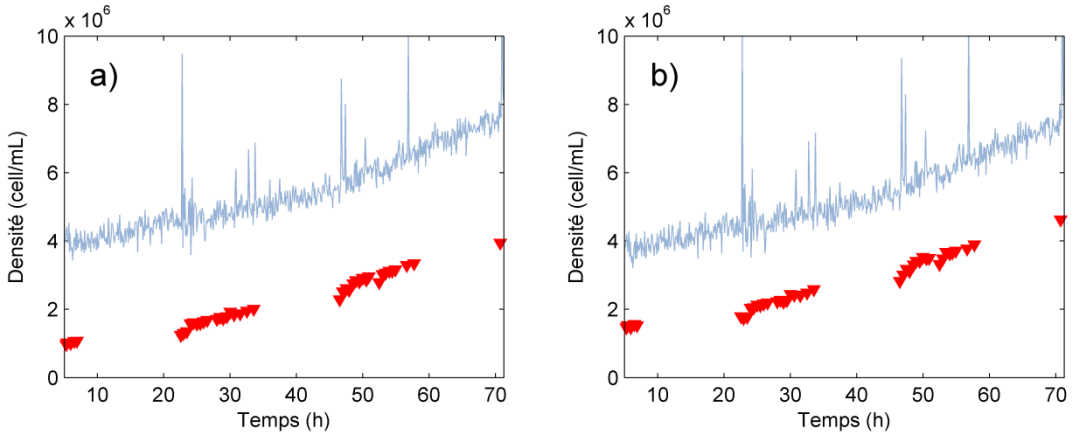

Figure 4.25 Représentations des prédictions réalisées sur les spectres de la culture de cellules Sf9 #2 acquis en continu (en bleu) pour a) la densité VCD et b) la densité TCD. Les densités obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges.

Nous pouvons finalement considérer que les modèles de régression développées sur les densités en cellules vivantes et sur les densités totales en cellules peuvent permettre de réaliser le suivi des paramètres. Néanmoins, le fait de ne disposer que d'une seule culture en calibration et une culture test pour l'évaluation ne nous permet pas de développer des modèles de régression suffisamment robustes pour prédire de façon satisfaisante les niveaux des densités cellulaires sur la phase d'amplification cellulaire.

Les densités mesurées lors de la seconde phase de culture, à savoir la culture virale, présentent quant à elles des gammes de variation moins importantes. En effet, ces-dernières varient entre $0{,}68 \cdot 10^6$–$3{,}95 \cdot 10^6$ cell/mL pour le paramètre VCD et $1{,}41 \cdot 10^6$–$5{,}39 \cdot 10^6$ cell/mL pour le paramètre TCD. À travers les performances des modèles, Tableau 4.8, nous pouvons à nouveau souligner des erreurs de prédiction importantes pour les modèles de régression basés sur les densités VCD et TCD ($1{,}05 \cdot 10^6$ cell/mL et $1{,}20 \cdot 10^6$ cell/mL respectivement) malgré des erreurs RMSECV acceptables ($0{,}34 \cdot 10^6$ cell/mL et $0{,}22 \cdot 10^6$ cell/mL respectivement). Ceci traduit une nouvelle fois la présence d'erreurs dans les prédictions à cause d'un manque de robustesse des modèles PLS. Les représentations disponibles Figure 4.26 permettent de mettre en avant ce phénomène. Néanmoins, il est important de noter qu'ici, les densités cellulaires les plus importantes de la culture test (culture #4) dépassent la gamme de densités mise en jeu dans l'ensemble de calibration, ce qui accroit les erreurs de prédiction puisque les modèles extrapolent les valeurs prédites.

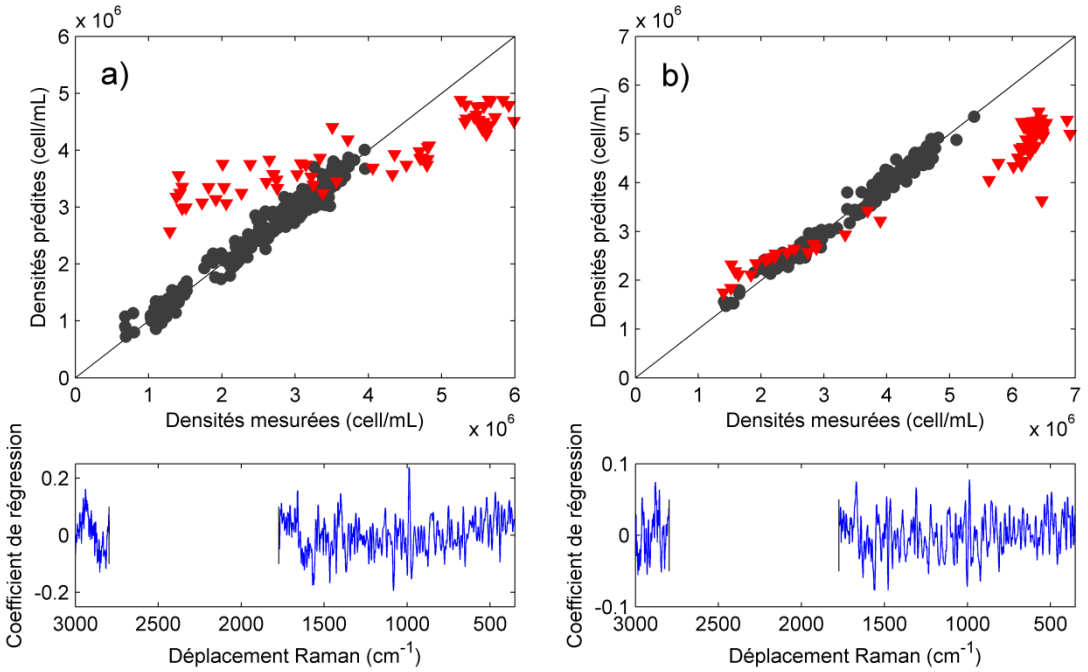

Figure 4.26 Représentations graphiques des modèles calculés à partir de jeux de données cumulant les échantillons de plusieurs lots de cultures virales mettant en jeu les cellules Sf9 pour a) la densité en cellules vivantes et b) la densité en cellules totale. Les points noirs représentent les échantillons de calibration, les triangles rouges les échantillons du jeu test.

Toutefois, en prenant les prédictions réalisées sur les spectres acquis en continu pour la culture #4, Figure 4.27, nous pouvons voir que les prédictions permettent de suivre les tendances des biomarqueurs au cours de la culture virale. Dans le cas des protocoles engagés pour les cultures impliquant les cellules Sf9, les paramètres VCD et TCD se distinguent lors de la phase de culture virale puisque la méthode de référence parvient à mesurer les cellules mortes. Ainsi, la densité VCD tend à décroitre une fois que le virus est très présent dans le milieu, tandis que la densité TCD atteindra un certain palier maximal qu'elle ne pourra pas dépasser à cause du déclin de la viabilité cellulaire dans le milieu.

Finalement, les modèles de régression basés sur les densités cellulaires, que ce soit VCD ou TCD, mesurées pour les cultures mettant en jeu les cellules Sf9 durant les phases d'amplification cellulaire et de culture virale présentent tous des potentiels prédictifs intéressants. Toutefois, le manque de données implique des erreurs de prédiction importantes puisque la variabilité inter-culture n'est pas assez représentée au sein des jeux de données concernés. De ce fait, les performances des modèles PLS restent à un niveau insatisfaisant, nécessitant donc l'apport de données provenant d'autres cultures à l'ensemble d'apprentissage afin de développer des modèles de régression plus robustes.

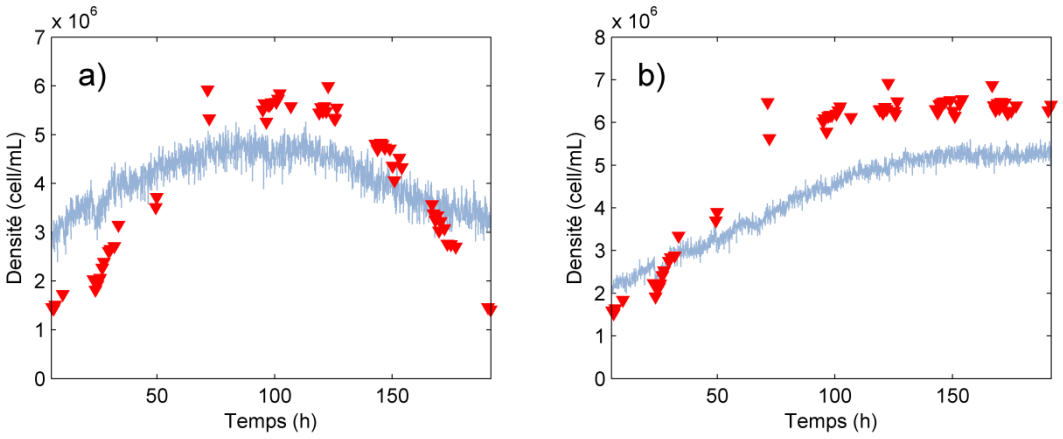

Figure 4.27 Représentations des prédictions réalisées sur les spectres de la culture de cellules Sf9 #4 acquis en continu (en cyan) pour a) la densité VCD et b) la densité TCD durant la phase de culture virale. Les densités obtenues à partir des mesures de référence sur les prélèvements sont représentées par les triangles rouges.

Nous pouvons conclure quant à l'analyse des modèles de régression développés sur la base de plusieurs cultures pour les différents types de cellule (CHO, HeLa, Sf9) que l'intégration de la variabilité qui existe entre les lots permet d'accroitre les performances prédictives des modèles. Néanmoins, nous pouvons noter que jusque-là, nous avons seulement mis en jeu des cultures dont les paramètres physiques restent inchangés dans les conditions normales de culture afin de conserver des cultures semblables d'un point de vue biologique. Cependant, il est important de noter que lors de cultures de cellules, il apparait parfois certaines variations des paramètres physiques telles que des variations de pH ou de

température par exemple. Ces changements influent inévitablement sur le matériel biologique présent dans le milieu de culture, faisant ainsi varier les paramètres biochimiques.

2 Intégration des variations physiques pour la construction d'un modèle de régression robuste

En faisant varier les paramètres physiques de culture, nous pouvons réaliser des lots présentant des développements métaboliques atypiques, telle qu'une production de lactate plus importante ou une faible consommation en glucose. Ainsi, en faisant varier les paramètres de contrôle dans des conditions qui restent viables, nous pouvons obtenir des cultures qui diffèrent de celles réalisées dans les conditions normales et donc apporter plus de robustesse aux modèles de régression.

Au cours de ces travaux de recherche, un certain nombre de paramètres de culture a été étudié afin, tout d'abord, d'analyser l'influence des changements de paramètres métaboliques des cultures de cellules, et donc sur les mesures Raman, pour ensuite intégrer ces variations aux modèles de régression si possible. Toutefois, nous notons ici que ce travail a été réalisé uniquement pour les cultures de cellules CHO par manque de temps.

2.1 Influence des paramètres physiques

Cette étude est réalisée sur plusieurs paramètres physiques critiques pour les cultures de cellules. En effet, ils présentent un intérêt important dans le développement de la culture et nécessitent un contrôle permanent (Chapitre 1, section 4.1, Régulation des paramètres physiques, page 37). Ainsi, nous proposons d'étudier ici l'impact des variations de différents paramètres clés sur les cultures de cellules, notamment le pH, la pO_2 (soit l'apport en oxygène dans le milieu), la vitesse de rotation de l'hélice qui entraine l'agitation du milieu (ou vitesse d'agitation) et la température. Pour ces différents travaux, nous prendrons en considération les modèles de régression élaborés en accumulant les données provenant de plusieurs cultures CHO, modèles présentés section 1.1 de ce chapitre (page 129), en excluant le modèle basé sur les concentrations en ammonium, jugé inadéquat.

2.1.1 Variations du pH

Trois cultures de cellules CHO ont spécialement été effectuées en faisant varier le paramètre pH. La première culture, soit la culture #10, a été amorcée à pH 7,2 (contre 6,9 pour les conditions normales), puis à 6,7 au bout de 100 h de culture. Ce lot comprend un total de 42 mesures de référence pour les niveaux des différents biomarqueurs (sauf la concentration en glutamine qui n'en compte que 12). La deuxième culture (culture #11) a été

réalisée en appliquant un pH de 6,7 du début à la fin du protocole, pour un total de 43 mesures de référence (13 pour la concentration en glutamine). Enfin, la troisième culture utilisée pour étudier l'influence du pH, la culture #12, a été entièrement réalisée à pH 7,2 et compte également 43 mesures de référence (10 pour la concentration en glutamine).

Dans un premier temps, nous vérifions si les modèles de régression développés sur la simple base de cultures acquises dans les conditions normales (pH 6,9) sont capables de prédire de manière satisfaisante les niveaux des biomarqueurs des cultures dont le pH est différent. Ainsi, nous prenons tour à tour les lots #10, #11 et #12 en tant que jeu test et y appliquons les modèles de régression. Les résultats obtenus sont répertoriés Tableau 4.9, reprenant également les erreurs RMSECV des modèles figurant Tableau 4.2 (page 131).

Tableau 4.9 Récapitulatif des performances des modèles de régression appliqués aux jeux test élaborés à partir des échantillons des cultures dont les pH ont été changés par rapport aux conditions normales.

Biomarqueur (unité)	*Lot test*	RMSECV	Gamme (calibration)	RMSEP	Gamme (test)
Glucose (mM)	*#1*	4,25	2,25 – 62,13	5,68	2,95 – 55,59
	#10	–	–	5,36	12,91 – 50,91
	#11	–	–	5,12	45,18 – 91,81
	#12	–	–	5,23	12,82 – 40,93
Glutamine (mM)	*#1*	0,17	0,00 – 2,72	0,46	0,00 – 3,65
	#10	–	–	0,58	0,36 – 1,28
	#11	–	–	0,20	0,11 – 1,47
	#12	–	–	0,20	0,11 – 2,01
Glutamate (mM)	*#1*	0,63	2,21 – 13,67	0,41	2,40 – 9,01
	#10	–	–	0,91	2,35 – 13,80
	#11	–	–	1,76	2,19 – 14,77
	#12	–	–	1,88	2,23 – 14,10
Lactate (mM)	*#1*	1,55	0,00 – 39,18	2,52	0,58 – 24,96
	#10	–	–	2,38	14,04 – 45,07
	#11	–	–	8,90	2,89 – 21,54
	#12	–	–	11,84	1,89 – 66,16
VCD (10^6 cell/mL)	*#1*	0,99	0,29 – 15,57	1,10	0,30 – 14,75
	#10	–	–	1,75	1,87 – 10,00
	#11	–	–	1,30	0,29 – 3,71
	#12	–	–	3,56	0,32 – 6,84
TCD (10^6 cell/mL)	*#1*	1,05	0,29 – 15,70	1,13	0,29 – 15,62
	#10	–	–	1,82	1,92 – 10,08
	#11	–	–	0,92	0,29 – 3,75
	#12	–	–	3,72	0,32 – 6,96

Dans un premier temps, nous pouvons noter l'influence de la variation de pH sur les densités cellulaires. En effet, puisque les conditions optimales de culture ne sont pas prises en compte, nous pouvons souligner une forte diminution des maxima obtenus pour les densités en cellules vivantes et densités totales des nouvelles cultures test. Ceci a donc pour premier effet d'influer sur les consommations et productions des différents métabolites d'intérêt pris en compte dans ces travaux. Nous pouvons en effet remarquer que la consommation en glucose est nettement inférieure pour la culture #11 puisque le niveau maximal pour la gamme de calibration est à 62,13 mM tandis que le lot en question présente un maximum à 91,81 mM. Toutefois, nous pouvons noter des erreurs RMSEP acceptables (5,36 mM, 5,12 mM et 5,23 mM respectivement) pour les trois lots test compte tenu des erreurs RMSECV (4,25 mM) et RMSEP (5,68 mM) obtenues lors du développement du modèle PLS. Néanmoins, ces erreurs traduisent la présence de certains écarts des prédictions réalisées, représentés sur la Figure 4.28.

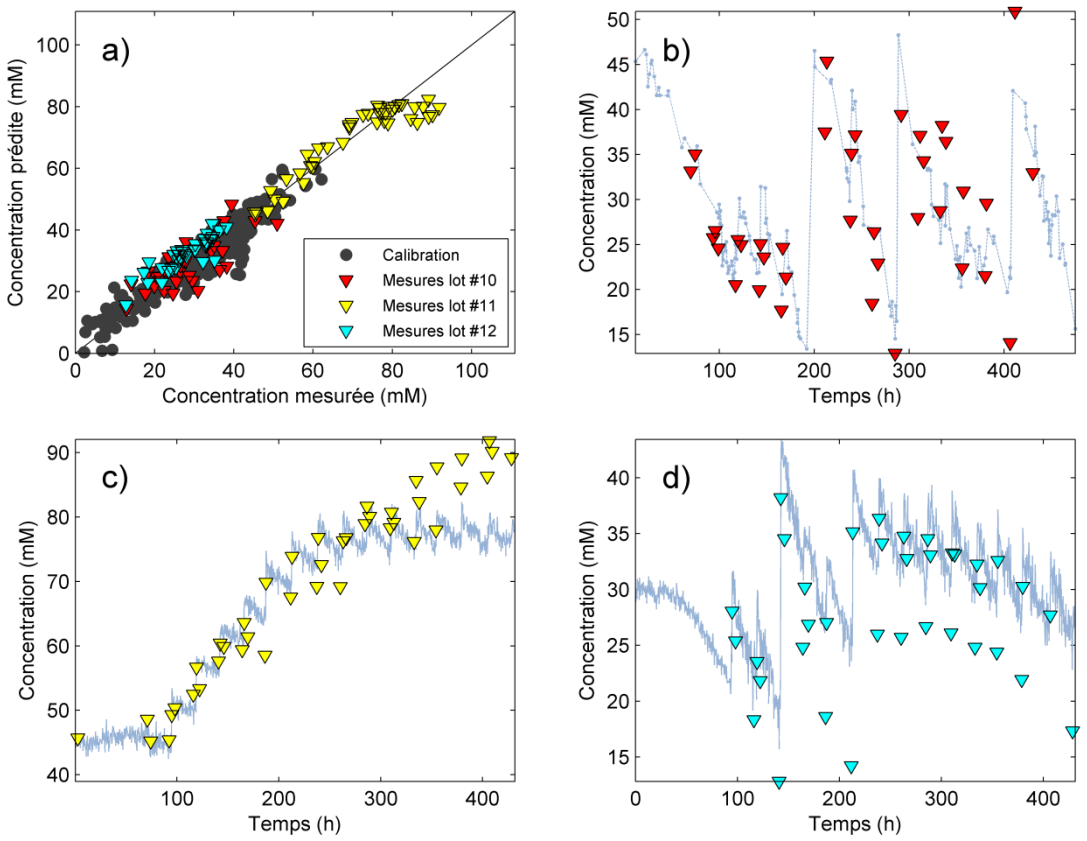

Figure 4.28 Représentations des prédictions des concentrations en glucose obtenues pour les échantillons des cultures réalisées en faisant varier le pH par rapport aux conditions normales, comprenant les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #10, c) des spectres acquis en continu pour le lot #11 et d) des spectres acquis en continu pour le lot #12.

Cependant, en comparant les erreurs RMSEP calculées ici à celle obtenue lors du développement du modèle PLS, à savoir sur un jeu test composé d'une culture de cellules CHO réalisée dans les conditions normales (soit 5,68 mM), nous pouvons voir qu'il n'y a pas

de différence significative entre ces valeurs. Ainsi, nous ne sommes pas en mesure de dire si les biais observés ici sont dus aux variations de pH ou simplement aux variations inter-cultures.

D'autre part, nous pouvons noter les fortes concentrations en lactate lorsque le pH est fixé à 7,2. En effet, nous pouvons voir que le niveau en lactate dépasse les 45,00 mM pour les cultures #10 et #12 tandis que la concentration maximale des cultures acquises dans les conditions normales atteint 39,18 mM. D'un autre côté, nous pouvons noter que la culture test #11, réalisée à pH 6,7 présente une production de lactate inférieure, le niveau maximal étant mesuré à 21,54 mM. En prenant en compte les erreurs RMSEP déterminées pour les échantillons de chacune des trois cultures test, soit respectivement 2,38 mM, 8,90 mM et 11,84 mM (Tableau 4.9), il est évident que les lots #11 et #12 présentent des biais importants à cause des variations de pH. Les prédictions en lactate sont représentées Figure 4.29.

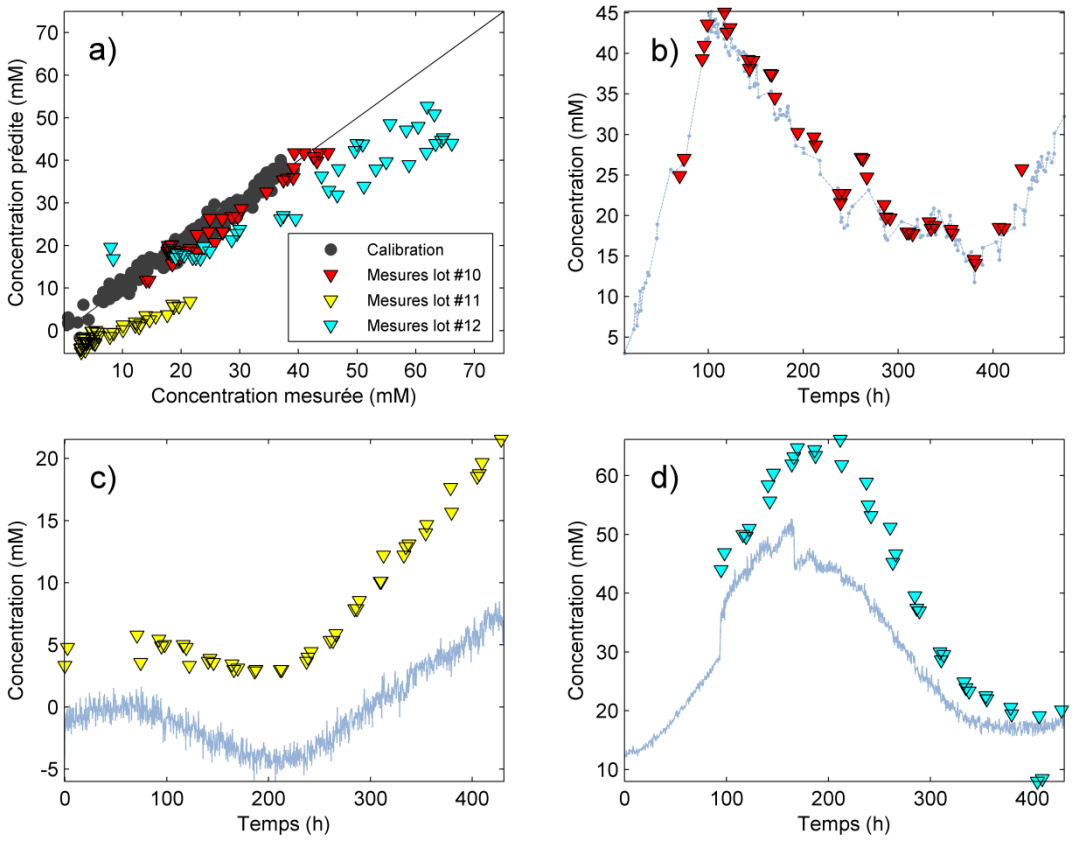

Figure 4.29 Représentations des prédictions des concentrations en lactate obtenues pour les échantillons des cultures réalisées en faisant varier le pH par rapport aux conditions normales, comprenant les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #10, c) des spectres acquis en continu pour le lot #11 et d) des spectres acquis en continu pour le lot #12.

Si nous prenons les prédictions obtenues pour les échantillons acquis lors de la culture #10, Figure 4.29b, nous pouvons voir que le modèle de régression pour les concentrations en lactate permet de très bien prédire les teneurs de ce métabolite tout au long du

bioprocédé avec une faible erreur (2,38 mM contre 2,52 mM lors du développement du modèle). En prenant ensuite les cultures présentant des variations plus prononcées dans les gammes de concentrations, à savoir les lots #11 et #12, les prédictions obtenues présentent indéniablement des biais importants qui apparaissent nettement Figure 4.29c-d. Ceci est directement dû aux gammes d'étalonnage. En effet, pour la culture #11, le niveau de la gamme de concentration est bien plus faible que lors du développement du modèle, rendant donc l'évolution du paramètre moins prononcée et plus difficile à capter sur les spectres Raman. Au contraire, pour la culture #12, nous pouvons voir que la gamme de concentration est bien supérieure à la gamme d'étalonnage. Le modèle de régression PLS doit donc extrapoler les concentrations prédites pour les spectres de cette culture, ce qui accroit alors le biais sur les prédictions. Néanmoins, pour les lots #11 et #12, nous pouvons remarquer que les prédictions obtenues en continu parviennent à suivre les tendances des paramètres durant toute la durée du bioprocédé. Ceci traduit donc bien un biais dans les prédictions et non pas un problème de représentativité des modèles de régression par rapport à la molécule de lactate.

Enfin, si nous prenons en compte les modèles calculés pour les densités cellulaires, nous pouvons remarquer que les erreurs de prédictions RMSEP des cultures #10 et #11 sont satisfaisantes ($1{,}75 \cdot 10^6$ cell/mL et $1{,}30 \cdot 10^6$ cell/mL respectivement pour le paramètre VCD) compte tenu des gammes de densité, notamment pour le lot #11, et de l'erreur RMSEP calculée lors du développement du modèle PLS ($1{,}10 \cdot 10^6$ cell/mL pour la densité VCD). Toutefois, nous pouvons remarquer que l'erreur de prédiction du lot #12 présente un niveau bien plus important ($3{,}56 \cdot 10^6$ cell/mL), traduisant une nouvelle fois la présence d'un biais dans les prédictions, qui peut être facilement représenté en prenant les prédictions de la densité VCD réalisées sur les spectres en continu pour, Figure 4.30d.

De manière générale, nous pouvons finalement dire que le pH a bien un effet non seulement sur les cultures, mais aussi sur les performances des modèles PLS puisque ces-derniers sont directement affectés par les variations des gammes de niveau des différents biomarqueurs. Nous pouvons toutefois remarquer que la culture #10 est moins affectée que les deux autres lots à travers les gammes de densité cellulaire qui parviennent à conserver un niveau similaire à ce que nous pouvons rencontrer lors des cultures réalisées dans les conditions normales. Ainsi, si nous souhaitons améliorer les performances prédictives des modèles en rendant ces-derniers plus robustes, nous devons intégrer la variabilité due aux variations de pH au modèle de régression.

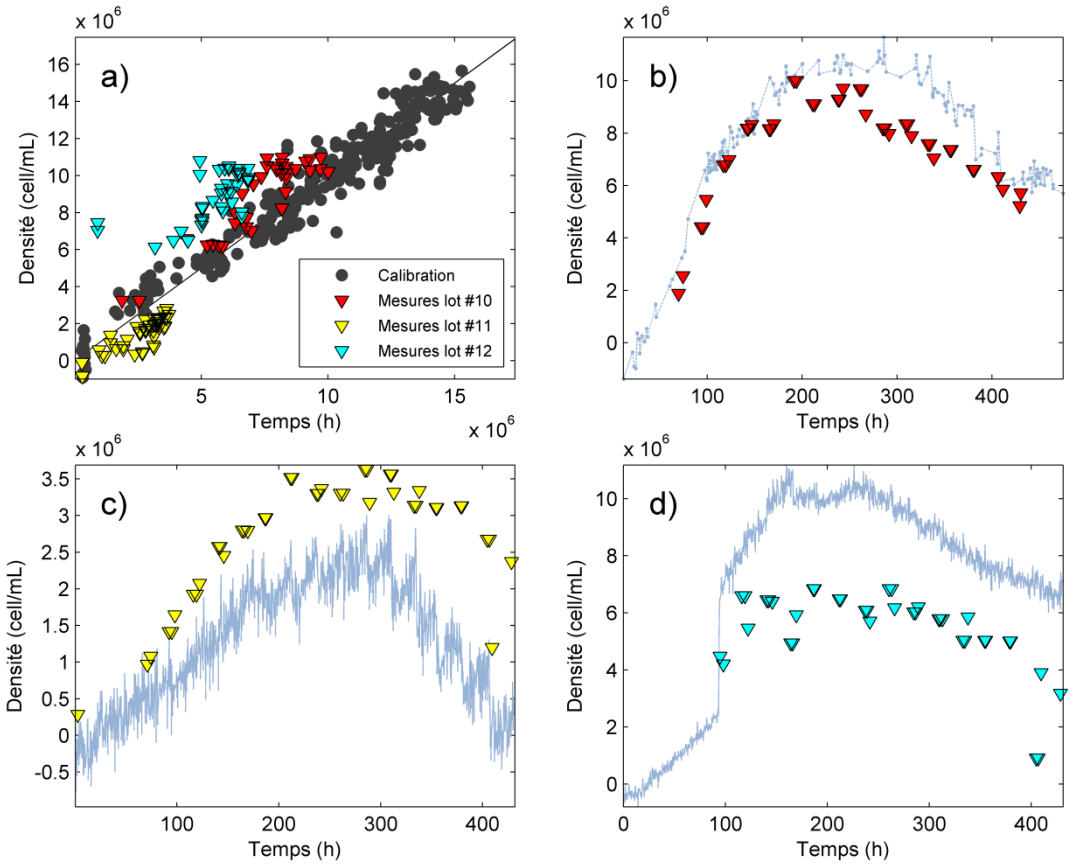

Figure 4.30 Représentations des prédictions des densités VCD obtenues pour les échantillons des cultures réalisées en faisant varier le pH par rapport aux conditions normales, comprenant les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #10, c) des spectres acquis en continu pour le lot #11 et d) des spectres acquis en continu pour le lot #12.

Parmi les trois lots à notre disposition, nous préférons intégrer les échantillons du lot #12 à l'ensemble d'apprentissage. En effet, ce dernier ne présente pas de comportement quasi-normal, tel que le lot #10, et également une gamme de concentration du lactate très supérieure à la normale. Nous aurions également pu prendre les échantillons du lot #11 dans l'ensemble de calibration pour son importante gamme de variation en glucose, mais ce-dernier paramètre présente moins d'erreur que la teneur en lactate sur les prédictions, ce qui le rend moins intéressant pour améliorer la robustesse du modèle. Les résultats obtenus après l'intégration du lot #12 à l'ensemble de calibration sont présentés Tableau 4.10. D'après les résultats obtenus, nous pouvons tout d'abord noter l'évolution des erreurs RMSECV. En effet, l'ensemble d'apprentissage est différent et produit donc de nouvelles erreurs lors de la validation croisée. Nous pouvons alors remarquer que les modèles développés pour les concentrations en glucose, glutamine et glutamate présentent des erreurs plus faibles tandis que celle de la concentration en lactate, ainsi que celles des densités cellulaires présentent de légères hausses. Pour le lactate, ceci s'explique par la gamme de concentration plus importante après l'ajout du lot #12 à l'ensemble.

Tableau 4.10 Récapitulatif des performances des modèles de régression développés après avoir intégré le lot #12 aux cultures acquises dans les conditions normales dans l'ensemble de calibration pour l'étude des variations de pH.

Biomarqueur (unité)	*Lot test*	RMSECV*	Gamme (calibration)	RMSEP	Gamme (test)
Glucose (mM)	*#10*	4,19	2,25 – 62,13	5,17	12,91 – 50,91
	#11	–	–	5,41	45,18 – 91,81
Glutamine (mM)	*#10*	0,17	0,00 – 2,72	0,51	0,36 – 1,28
	#11	–	–	0,22	0,11 – 1,47
Glutamate (mM)	*#10*	0,58	2,21 – 14,10	0,71	2,35 – 13,80
	#11	–	–	0,75	2,19 – 14,77
Lactate (mM)	*#10*	2,34	0,00 – 66,16	3,68	14,04 – 45,07
	#11	–	–	14,90	2,89 – 21,54
VCD (10^6 cell/mL)	*#10*	1,03	0,29 – 15,57	0,99	1,87 – 10,00
	#11	–	–	1,57	0,29 – 3,71
TCD (10^6 cell/mL)	*#10*	1,12	0,29 – 15,70	1,14	1,92 – 10,08
	#11	–	–	1,33	0,29 – 3,75

* *L'erreur RMSECV est calculée à partir d'une* 10-fold cross-validation *réalisée sur le nouvel ensemble d'apprentissage*

D'autre part, nous pouvons remarquer que les erreurs RMSEP ne présentent pas de changement important. En effet, pour le modèle de régression calculé pour le glucose, nous conservons des niveaux d'erreur entre 5,00 mM et 5,50 mM (pour une erreur RMSECV évaluée à 4,19 mM), traduisant une nouvelle fois certains écarts des prédictions par rapport aux mesures de référence, représentés Figure 4.31. Nous faisions la même observation avant l'intégration du lot #12 à l'ensemble de calibration, ce qui signifie que les biais observés précédemment provenaient principalement de différences intrinsèques aux cultures et non pas des variations de pH.

L'amélioration la plus notable concerne le paramètre glutamate. En effet, nous obtenions précédemment une erreur RMSEP de 1,76 mM lors des prédictions des échantillons du lot #11 à l'aide du modèle de régression basé sur les échantillons des lots acquis dans les conditions normales de culture (pour une erreur RMSECV de 0,63 mM). L'intégration des échantillons de la culture cellulaire #12 au sein de l'ensemble de calibration permet de réduire cette erreur RMSEP à 0,75 mM (pour une erreur RMSECV de 0,58 mM), ce qui traduit l'amélioration des prédictions du modèle en réduisant le biais existant, représentée Figure 4.32.

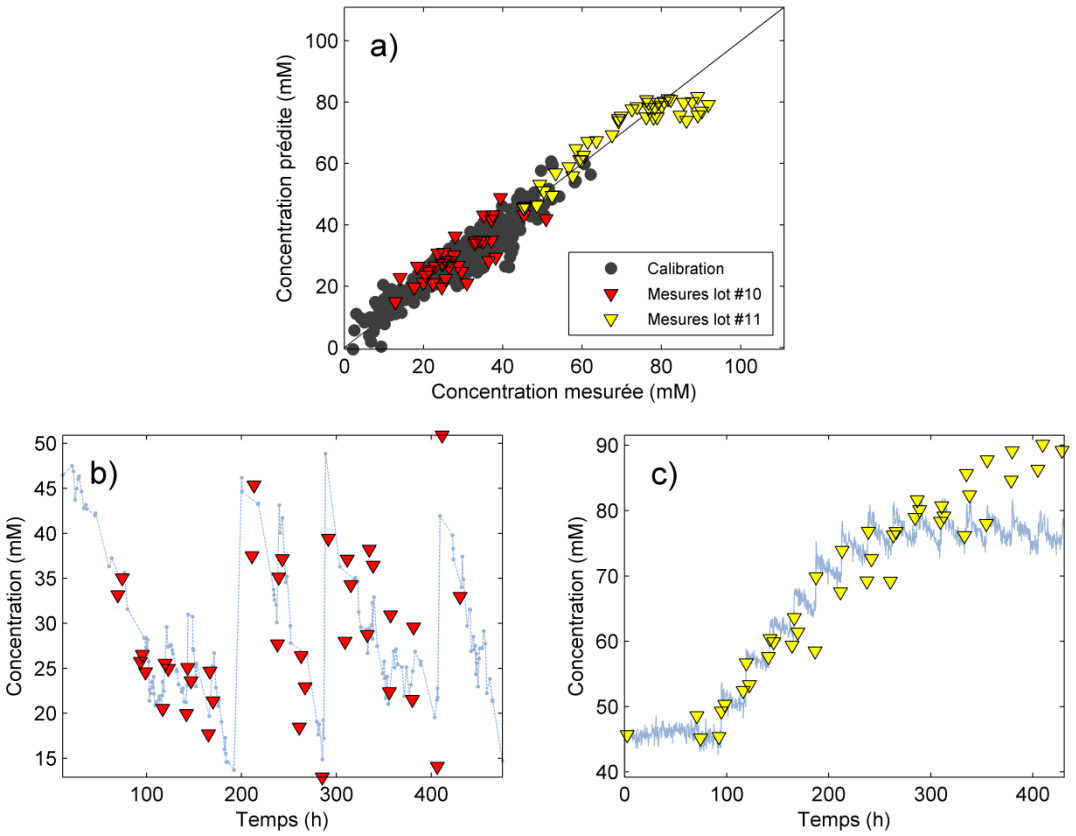

Figure 4.31 Représentations des prédictions des concentrations en glucose obtenues pour les échantillons des cultures #10 et #11 par le modèle PLS intégrant les échantillons du lot #12 dans l'ensemble de calibration. Cette figure comprend les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #10 et c) des spectres acquis en continu pour le lot #11.

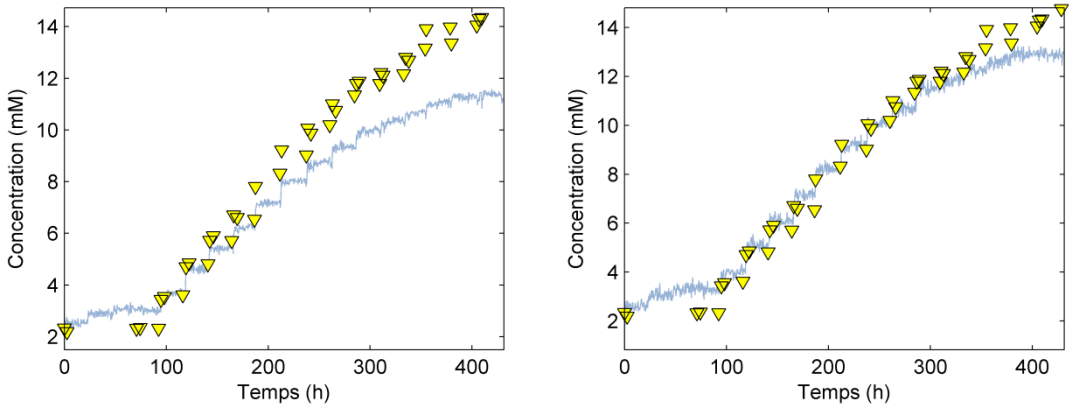

Figure 4.32 Comparaison des prédictions des concentrations en glutamate réalisées sur les spectres acquis en continu de la culture cellulaire #11 avant (à gauche) et après (à droite) l'intégration du lot #12 à l'ensemble de calibration.

Toutefois, nous pouvons observer Tableau 4.10 une très importante erreur RMSEP pour les prédictions des concentrations en lactate pour les échantillons de la culture #11. En effet, une erreur RMSEP de 14,90 mM (pour une erreur RMSECV de 2,34 mM) traduit simplement l'incapacité du modèle à prédire les niveaux en lactate du lot #11. En représentant les résultats du modèle lactate, Figure 4.33, nous pouvons observer l'important biais existant pour les prédictions des échantillons de la culture #11 (Figure 4.33c). Cependant, les

prédictions obtenues pour les échantillons du lot #10 permettent de montrer que le modèle de régression est tout de même capable de réaliser les prédictions des teneurs en lactate. Cela signifie que l'importante erreur obtenue pour le lot #11 provient des différences métaboliques observées, en particulier la faible production de lactate, qui devra donc être intégrée au modèle de régression PLS pour accroitre la robustesse de ce dernier vis-à-vis de variations de pH.

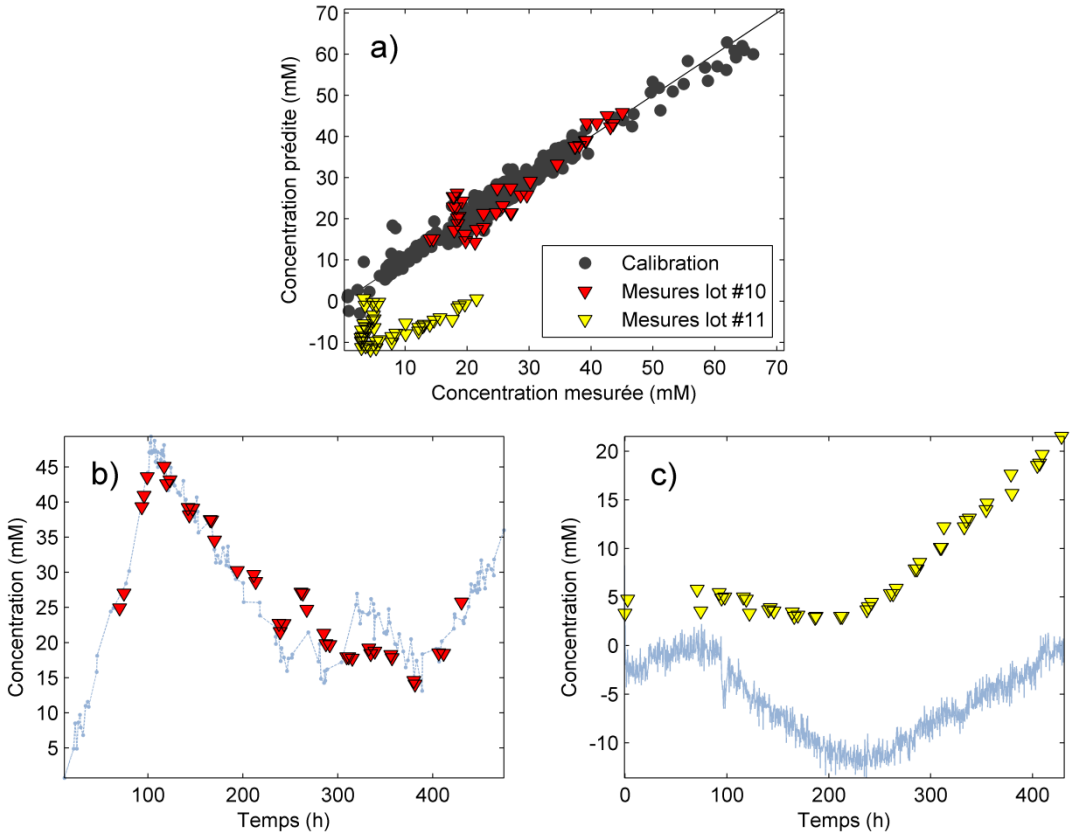

Figure 4.33 Représentations des prédictions des concentrations en lactate obtenues pour les échantillons des cultures #10 et #11 par le modèle PLS intégrant les échantillons du lot #12 dans l'ensemble de calibration. Cette figure comprend les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #10 et c) des spectres acquis en continu pour le lot #11.

Enfin, les résultats obtenus pour les prédictions des densités cellulaires, Tableau 4.10, présentent des caractéristiques similaires à celles réalisées avant l'intégration des échantillons de la culture cellulaire #12 au sein de l'ensemble de calibration, ce qui signifie que l'influence du pH sur ces densités n'a pas d'impact aussi important que ce que nous avons pu observer sur la concentration en lactate par exemple.

Pour conclure quant à l'effet de variations de pH, nous avons pu observer des différences métaboliques pour les cultures cellulaires mises en jeu. Toutefois, les prédictions réalisées sur les biomarqueurs d'intérêt ne présentaient pas de biais trop important, à l'exception de la concentration en lactate pour certaines cultures. Toutefois, l'intégration d'échantillons provenant d'un lot dont le pH présente un niveau différent du paramètre

normal de culture n'améliore pas les différents écarts observés. Ainsi, pour les différents biomarqueurs (excepté le lactate), nous pouvons dire que les variations de pH ne présentent pas plus d'influence que les variations inter-cultures sur les performances prédictives des modèles. Pour les concentrations en lactate, nous pouvons voir que les variations de pH ont une influence qu'il convient d'intégrer à l'ensemble d'apprentissage afin d'accroitre la robustesse des modèles de régression. Finalement, non seulement pour le lactate, mais aussi pour les autres biomarqueurs, l'intégration de cultures réalisées en faisant varier le pH permet de prendre en compte des gammes de concentration atypiques et donc d'accroitre la robustesse des modèles de régression vis-à-vis de changements de pH potentiels lors des cultures de cellules.

2.1.2 Variations de la pO_2

Pour étudier l'influence des variations de ce paramètre, nous disposons de deux lots dont la pO_2 est fixée à des valeurs différentes de celle des conditions normales de culture, à savoir 40 %. Nous avons d'abord une première culture (culture #13) présentant une pO_2 à 20 % pour laquelle 43 prélèvements ont été réalisés pour déterminer les teneurs des différents biomarqueurs (sauf les concentrations en glutamine pour lesquelles 12 mesures ont été faites). Ensuite, nous disposons d'une seconde culture (culture #14) disposant d'une pO_2 différente, fixée à 60 %, pour laquelle 42 mesures de référence ont été réalisées (9 pour les concentrations en glutamine).

Afin d'évaluer l'impact des variations de pO_2, nous travaillons de manière similaire à l'étude effectuée pour les variations de pH : nous prenons d'abord les modèles de régression développés sur la base des échantillons des cultures réalisées dans les conditions normales (section 1.1 de ce chapitre, Modèles de régression pour les cellules CHO, page 129) ainsi que les échantillons des deux cultures présentées ci-dessus en tant que jeu test. Les résultats obtenus pour chaque biomarqueur sont indiqués au sein du Tableau 4.11.

Tout d'abord, nous pouvons noter que les variations de pO_2 ont moins d'impact sur les densités cellulaires que ce que nous avons rencontré pour les variations de pH. En effet, les densités (VCD et TCD) maximales atteintes sont moins importantes que celles rencontrées pour les cultures cellulaires en conditions normales, mais elles se rapprochent d'un comportement normal. De ce fait, les évolutions des concentrations des métabolites suivent des profils quasi-normaux, mis-à-part la très forte production de lactate pour la culture #13, effet le plus notable de la variation de pO_2.

Tableau 4.11 Récapitulatif des performances des modèles de régression appliqués aux jeux test élaborés à partir des échantillons des deux lots dont les pO_2 ont été changées par rapport aux conditions normales de culture.

Biomarqueur (unité)	*Lot test*	RMSECV	Gamme (calibration)	RMSEP	Gamme (test)
	#1	4,25	2,25 – 62,13	5,68	2,95 – 55,59
Glucose (mM)	*#13*	–	–	6,47	10,12 – 45,65
	#14	–	–	6,39	2,00 – 49,78
		0,17	0,00 – 2,72	0,46	0,00 – 3,65
Glutamine (mM)	*#13*	–	–	0,27	0,18 – 2,28
	#14	–	–	0,29	0,45 – 1,34
		0,63	2,21 – 13,67	0,41	2,40 – 9,01
Glutamate (mM)	*#13*	–	–	1,18	2,60 – 9,59
	#14	–	–	1,53	2,45 – 15,22
		1,55	0,00 – 39,18	2,52	0,58 – 24,96
Lactate (mM)	*#13*	–	–	16,93	23,5 – 109,40
	#14	–	–	1,74	2,92 – 34,77
		0,99	0,29 – 15,57	1,10	0,30 – 14,75
VCD (10^6 cell/mL)	*#13*	–	–	1,99	2,28 – 12,16
	#14	–	–	1,21	1,98 – 11,65
		1,05	0,29 – 15,70	1,13	0,29 – 15,62
TCD (10^6 cell/mL)	*#13*	–	–	2,07	2,37 – 12,29
	#14	–	–	1,02	2,04 – 11,73

Si nous prenons les erreurs de prédiction RMSEP obtenues pour les échantillons des cultures #13 et #14, Tableau 4.11, nous pouvons voir que celles-ci sont semblables à celles obtenues lors des travaux sur les variations de pH. À titre d'exemple, nous pouvons prendre le cas du modèle de régression basé sur les concentrations en glucose. Nous obtenons des erreurs RMSEP supérieures à l'erreur RMSECV obtenue lors du développement du modèle de régression (6,47 mM et 6,39 mM pour les cultures #13 et #14 respectivement, contre 4,25 mM pour l'erreur RMSECV), traduisant ainsi quelques écarts lors des prédictions des concentrations en glucose, représentés Figure 4.34. Néanmoins, malgré ces différences, il apparait que le modèle de régression est capable de suivre les profils des concentrations en glucose.

L'erreur RMSEP la plus importante est obtenue pour les prédictions des échantillons de la culture cellulaire #14. En effet, nous pouvons noter une erreur déterminée à 16,93 mM alors que l'erreur RMSECV obtenue lors du développement du modèle atteint 1,55 mM. Ceci provient directement de la gamme de concentration du lactate pour le lot #14 qui s'étend de 23,5 mM jusqu'à 109,4 mM, soit plus de deux fois le niveau maximal rencontré lors de la

calibration du modèle de régression (39,18 mM). En représentant les prédictions réalisées pour les teneurs en lactate, Figure 4.35, nous pouvons très facilement observer les erreurs obtenues pour les prédictions du lot #14. Toutefois, nous pouvons observer que les résultats obtenus pour les prédictions des échantillons de la seconde culture ayant une valeur de pO_2 différente des conditions normales sont satisfaisants. En effet, l'erreur RMSEP obtenue atteint 1,74 mM (contre 2,52 mM lors du développement du modèle PLS).

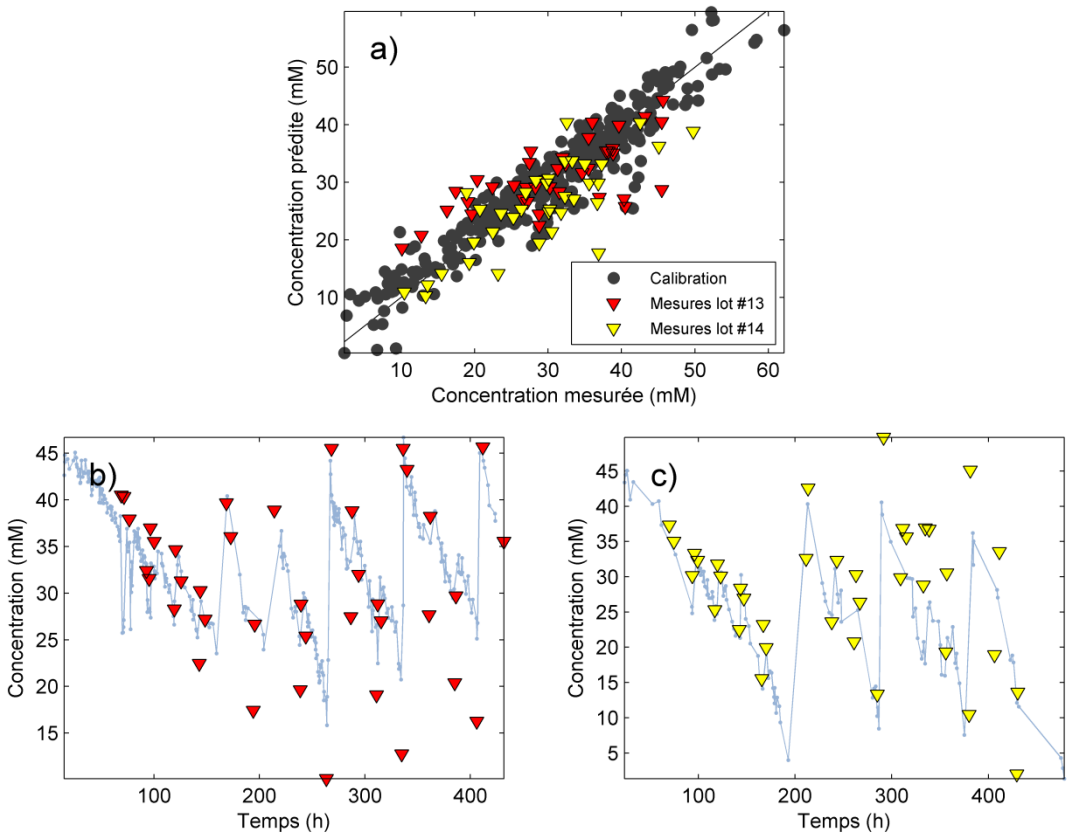

Figure 4.34 Représentations des prédictions des concentrations en glucose obtenues pour les échantillons des lots réalisées en faisant varier la pO_2 par rapport aux conditions normales de culture, comprenant les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #13 et c) des spectres acquis en continu pour le lot #14.

Ainsi, nous montrons ici que les variations de pO_2 ont bien un impact sur les niveaux métaboliques qu'il convient d'intégrer aux modèles de régression, notamment pour permettre à ces-derniers de bénéficier de gammes de concentrations plus importantes. Nous ajoutons alors le lot #13 à l'ensemble de calibration basé uniquement sur des cultures de cellules réalisées dans les conditions normales afin de voir si nous améliorons les performances des modèles pour les échantillons du lot #14. Ces résultats figurent Tableau 4.12.

Tableau 4.12 Récapitulatif des performances des modèles de régression développés après avoir intégré le lot #13 aux cultures acquises dans les conditions normales dans l'ensemble de calibration pour l'étude des variations de pO_2. Le jeu test est uniquement constitué des échantillons de la culture #14.

Biomarqueur (unité)	RMSECV*	Gamme (calibration)	RMSEP	Gamme (test)
Glucose (mM)	4,39	2,25 – 62,13	5,73	2,00 – 49,78
Glutamine (mM)	0,17	0,00 – 2,72	0,30	0,45 – 1,34
Glutamate (mM)	0,53	2,21 – 13,67	1,16	2,45 – 15,22
Lactate (mM)	2,89	0,00 – 109,40	4,51	2,92 – 34,77
VCD (10^6 cell/mL)	0,93	0,29 – 15,57	1,52	1,98 – 11,65
TCD (10^6 cell/mL)	0,91	0,29 – 15,70	1,20	2,04 – 11,73

* *L'erreur RMSECV est calculée à partir d'une* 10-fold cross-validation *réalisée sur le nouvel ensemble d'apprentissage*

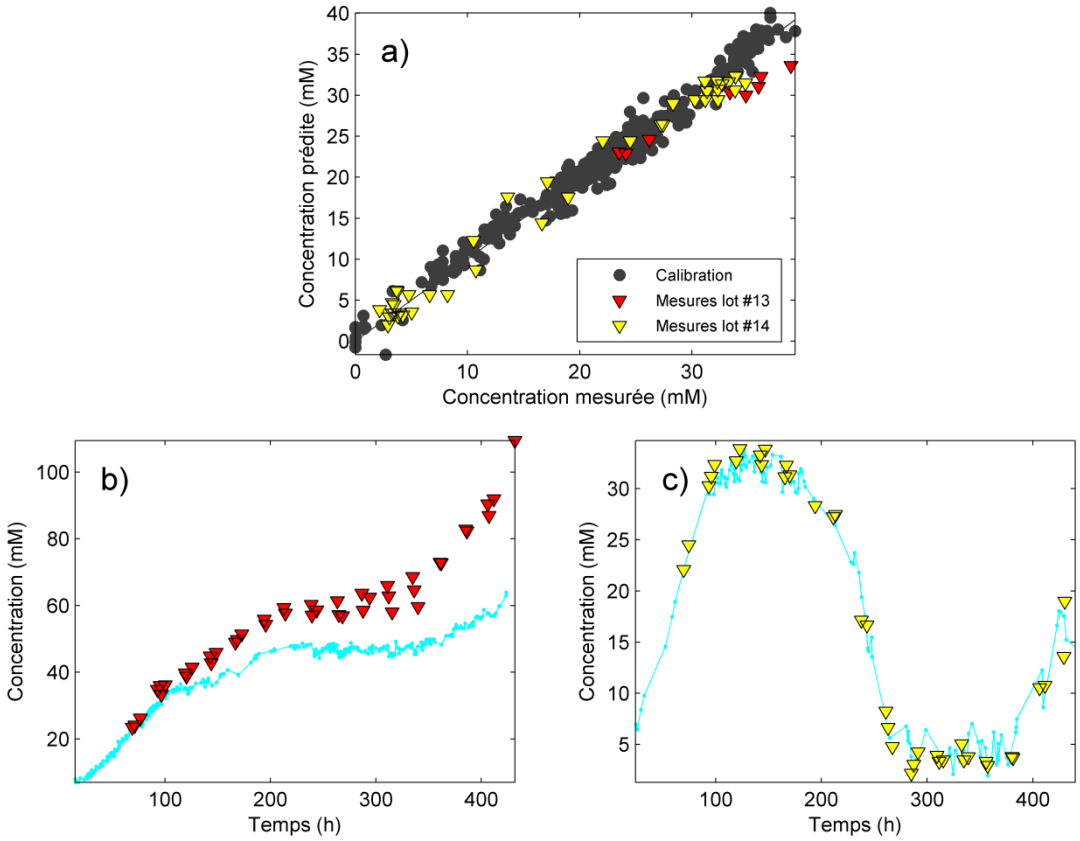

Figure 4.35 Représentations des prédictions des concentrations en lactate obtenues pour les échantillons des lots réalisées en faisant varier la pO_2 par rapport aux conditions normales de culture, comprenant les prédictions a) des échantillons test, b) des spectres acquis en continu pour le lot #13 et c) des spectres acquis en continu pour le lot #14.

Nous pouvons alors remarquer que les performances des modèles de régression vis-à-vis de la culture #14, à travers les erreurs de prédiction, sont du même ordre de grandeur que ce que nous avions précédemment, voire même améliorées (cas du glucose notamment, nous passons d'une erreur RMSEP estimée à 6,39 mM à 5,73 mM). En revanche, l'erreur de prédiction pour le modèle de régression des teneurs en lactate permet

de montrer que le fait d'étendre la gamme de concentration ne permet pas forcément d'améliorer le modèle de calibration. En effet, les échantillons de l'ensemble d'apprentissage possèdent un fort levier lors du développement du modèle et donc beaucoup d'impact du fait de leurs fortes concentrations. Or, le modèle présenté en intégrant la culture #14 ne dispose que d'une seule culture dont les concentrations atteignent ces valeurs. Le manque de représentativité à cette échelle de concentration entraine donc des erreurs de prédiction plus importantes pour les échantillons dont les niveaux en lactate restent normaux. Il sera donc nécessaire d'introduire plusieurs cultures produisant de fortes quantités de lactate au sein du modèle de régression si nous voulons développer un modèle robuste aux variations de pO_2.

2.1.3 Variations de vitesse d'agitation

Contrairement aux études présentées pour les variations de pH ou de pO_2, nous ne disposons pas de plusieurs cultures cellulaires acquises en changeant la vitesse d'agitation du milieu par rapport à son niveau dans les conditions normales. Ici, nous travaillerons donc uniquement sur du milieu de culture, permettant ainsi de déterminer si les variations de vitesse d'agitation influent sur la mesure Raman.

Pour cela, deux séries de spectres Raman ont été acquises avant l'introduction des cellules dans le milieu de culture, une première en fixant la vitesse d'agitation dans le bioréacteur de culture à 294 rpm, une seconde en fixant la vitesse à 299 rpm (la vitesse d'agitation des conditions normales de culture est paramétrée à 260 rpm). Notons que ces valeurs sont choisies afin de respecter les conditions pour obtenir une culture viable, où la vitesse n'est ni trop élevée (risque de cisaillement du matériel biologique), ni trop faible (perte d'homogénéité du contenu du bioréacteur). Nous reprenons ensuite les modèles de régression développés sur la base d'échantillons de lots réalisés dans les conditions normales de culture pour prédire les concentrations des différents métabolites des deux jeux de spectres. Puisque nous travaillons uniquement sur du milieu de culture, nous ne devons pas avoir de différences dans les prédictions réalisées pour les deux lots de spectres, auquel cas nous aurions alors un impact de la vitesse d'agitation sur les acquisitions de spectres Raman.

Pour chacun de deux jeux de spectres, nous en sélectionnons 20 acquis lorsque la vitesse d'agitation était stabilisée. Nous obtenons alors 40 prédictions pour chaque métabolite, 20 pour la vitesse 249 rpm et 20 pour la vitesse 299 rpm. Afin de déterminer s'il existe une différence entre les prédictions obtenues pour chaque groupe, nous procédons alors à une étude statistique permettant d'aboutir sur un test de différence entre les moyennes des données de chaque groupe et donc, les prédictions moyennes obtenues pour un même échantillon soumis à différentes vitesse d'agitation.

La première étape du cheminement statistique consiste à appliquer un test de détection d'*outliers*, échantillons présentant des caractères aberrants comparés au reste de la population. Pour cela, nous appliquons sur les deux lots de prédictions le test de Grubbs simple et le test de Grubbs double [147]. Ces tests comparent les écarts entre les valeurs extrêmes et la moyenne à l'écart-type des ensembles de données. Le test de Grubbs simple prend en compte une seule valeur extrême tandis que le test de Grubbs double tient compte de deux valeurs (soit les deux plus importantes, soit les deux plus faibles). En prenant ensuite des tables statistiques, nous sommes en mesure de détecter les échantillons aberrants pour chaque ensemble.

Ensuite, puisque les tests de différence statistique dépendent de plusieurs critères, il convient de vérifier d'abord la normalité des données. En effet, la normalité ou non des données conduit à différents types de test, tout comme le fait de travailler sur des populations présentant des tailles différentes (soit des nombres différents d'échantillons). Ainsi, nous appliquons d'abord le test de normalité de Shapiro-Wilk [148] sur chaque lot. De plus, nous calculons les coefficients d'asymétrie (*skewness*) et d'aplatissement (*kurtosis*) des deux distributions [149]. Enfin, nous traçons la droite de Henry (ou *Q–Q plot* en anglais) [150] qui permet d'ajuster une distribution gaussienne à nos séries pour vérifier la normalité. Pour les deux séries de données correspondant aux prédictions réalisées sur les spectres acquis sur du milieu de culture présent de le bioréacteur soumis à différentes vitesse d'agitation, nous montrons que quelque soit la concentration métabolique mise en jeu (glucose, glutamine, glutamate ou lactate), les deux distributions suivent statistiquement une loi normale.

Nous procédons ensuite à la dernière vérification statistique avant de réaliser le test de différence. Il s'agit d'un test d'homogénéité des variances. Dans notre cas, pour deux lots de tailles équivalentes et de distributions normales, nous choisissons d'appliquer le test de Fisher [151] qui prend en compte le rapport des variances des deux jeux de données et le compare à une valeur tabulée dépendante du seuil de confiance accordé. Dans notre cas, les résultats quant aux homogénéités des variances diffèrent selon le métabolite pris en compte. De ce fait, les calculs des coefficients mis en jeu dans le test de différence (test de Student [152]), notamment en ce qui concerne les écarts-types, varient selon les prédictions impliquées.

Finalement, les résultats des tests de différence permettent de montrer que les moyennes des deux jeux de données, soit les prédictions moyennes obtenues pour deux états différents du bioréacteur, ne présentent pas de différence significative (au risque de 5 %), quelque soit le métabolite pris en compte. De ce fait, nous concluons finalement que

l'agitation du système de culture n'a pas d'impact sur la mesure Raman. Dans ce cas, si les modèles de régression les plus robustes que nous développons sont basés sur des gammes de concentration importantes par la suite, les prédictions effectuées sur des spectres dont les cultures sont réalisées à vitesse d'agitation différente ne devraient pas présenter de biais.

2.1.4 Variations de température

Pour l'étude des variations de température, nous ne disposons pas de culture test réalisée dans des conditions particulières de culture. Nous procédons alors de la même manière que pour l'étude de l'impact de la vitesse du système d'agitation sur les spectres Raman. Ainsi, nous avons acquis des spectres Raman en continu sur du milieu de culture avant l'introduction des cellules. Deux températures différentes ont alors été soumises au milieu : 33°C et 37°C, températures extrêmes rencontrées durant les cultures de cellules CHO (nous rappelons ici que la température, au même titre que les autres paramètres physiques, est régulée en temps réel).

Deux ensembles de spectres sont alors créés, contenant chacun 20 spectres acquis lorsque la température est stabilisée à l'une ou l'autre des valeurs mises en jeu. Les modèles de régression développés à partir de cultures réalisées dans les conditions normales sont ensuite appliqués aux deux lots afin d'obtenir les prédictions des concentrations de chaque métabolite pour les deux ensembles.

Par la suite, nous procédons aux mêmes études statistiques afin de mener au test de différence des moyennes des deux ensembles pour chaque paramètre métabolique. Ainsi, pour chaque concentration métabolique, nous détectons d'abord les *outliers* protentiels à l'aide des tests de Grubbs (simple et double), ensuite nous vérifions la normalité des deux ensembles de prédictions à l'aide du test de Shapiro-Wilk, des calculs d'asymétrie et d'aplatissement, ainsi que de la droite de Henry, puis nous testons l'homogénéité des variances des deux populations à l'aide du test de Fisher pour enfin finir par un test de différence de Student.

Finalement, nous montrons que, pour chaque métabolite pris en compte, nous n'avons aucune différence significative (au risque de 5 %) entre les moyennes des prédictions réalisées sur les spectres acquis sur du milieu de culture à 33°C et à 37°C. Ainsi, tout comme les variations de vitesse d'agitation, nous pouvons conclure que la température n'a pas d'influence significative sur les acquisitions Raman de cultures cellulaires. De ce fait, si les modèles de régression balayent des gammes de concentration suffisamment étendues, nous ne devrions pas avoir de biais de prédiction provenant de différences induites par la température du contenu du bioréacteur.

Pour conclure quant la modification des paramètres physiques de culture, nous avons pu montrer que finalement, bien qu'il n'y ait pas d'influence sur le système d'acquisition Raman, le fait de faire varier les conditions de culture entraîne inévitablement des changements de gamme de concentration. Ainsi, le phénomène le plus récurrent est une sous-consommation du glucose et une surproduction de lactate entrainée par le faible taux de cellules dans le milieu. En effet, les performances des cultures sont amoindries par les changements apportés sur les conditions normales de cultures, conditions optimales pour favoriser la multiplication cellulaire. À travers ces gammes de concentration élargies, nous pourrons alors accroitre la robustesse des modèles de régression en intégrant les échantillons provenant de cultures cellulaires acquises dans des conditions atypiques.

2.2 Développement du modèle de régression robuste aux variations des paramètres physiques de culture

Suite aux travaux précédents, nous avons montré que l'intégration d'échantillons provenant de lots dont les paramètres physiques varient par rapport aux conditions normales de culture était nécessaire pour améliorer la robustesse des modèles de régression de chaque biomarqueur. Ainsi, nous reprenons d'une part les huit cultures de cellules CHO acquises dans les conditions normales utilisées durant la section 1.1 de ce chapitre (page 129), d'autre part les trois lots utilisés pour étudier les variations de pH (partie 2.1.1, page 170) et les deux lots mis en jeu pour l'étude des variations de pO_2 dans le milieu (partie 2.1.2, page 179). Enfin, une dernière culture, présentant une vitesse d'agitation différente des conditions normales, est ajoutée à l'ensemble d'apprentissage, portant donc le nombre de cultures employées en calibration à quatorze. Nous disposons au total de 685 mesures de références effectuées sur les prélèvements réalisés pour les différentes cultures (241 mesures pour la concentration en glutamine). Afin de comparer les performances prédictives des modèles de régression intégrant le maximum de variabilité (soit sur les échantillons des quatorze cultures présentées) à ceux développés uniquement sur les huit cultures acquises dans les conditions normales, nous prenons en compte le même jeu test, composé des 57 échantillons de la culture #1. Les résultats obtenus sont présentés Tableau 4.13, reprenant également les valeurs du Tableau 4.2 afin de faciliter la comparaison.

De manière générale, nous pouvons noter que les modèles de régression reposent sur un nombre de composantes plus important en intégrant les nouvelles cultures. Ceci s'explique simplement par l'augmentation de la variance du jeu de calibration en intégrant de nouveaux lots, qui plus est, présentant des comportements atypiques. Ceci a également pour effet d'accroitre les erreurs RMSECV obtenues lors des validations croisées réalisées durant le développement des modèles.

Tableau 4.13 Récapitulatif des performances prédictives des modèles de régression développés sur la base d'échantillons de cultures de cellules CHO, d'une part en travaillant uniquement sur huit lots acquis dans les conditions normales de culture, d'autre part en intégrant plusieurs lots présentant des variations dans les paramètres physiques (modèles sur quatorze lots).

Biomarqueur (unité)	*N lots*	R^2	LV	RMSECV	RMSEP	Gamme
Glucose (mM)	*9*	0,883	8	4,25	5,68	2,25 – 62,13
	14	0,895	9	4,65	5,28	2,25 – 91,81
Glutamine (mM)	*9*	0,917	5	0,17	0,46	0,00 – 2,72
	14	0,910	8	0,20	0,37	0,00 – 2,72
Glutamate (mM)	*9*	0,952	6	0,63	0,41	2,21 – 13,67
	14	0,949	10	0,69	1,00	2,19 – 15,22
Lactate (mM)	*9*	0,969	8	1,55	2,52	0,00 – 39,18
	14	0,959	10	3,09	2,47	0,00 – 109,40
VCD (10^6 cell/mL)	*9*	0,945	7	0,99	1,10	0,29 – 15,57
	14	0,938	10	0,97	0,94	0,29 – 16,94
TCD (10^6 cell/mL)	*9*	0,942	7	1,05	1,13	0,29 – 15,70
	14	0,937	10	0,99	0,94	0,29 – 17,72

2.2.1 Modèles pour les concentrations en glucose et en lactate

Tout d'abord, nous étudions les deux paramètres présentant le plus de changements par rapport aux modèles basés sur huit cultures. En effet, le fait d'intégrer les cultures présentant des variations de conditionnement a un impact évident sur les gammes de concentrations de ces deux biomarqueurs. Pour la gamme de concentration du glucose, celle-ci augmente de près de 50 % (de 2,25–62,13 mM à 2,25–91,81 mM) tandis que nous pouvons noter une augmentation de presque 180 % pour la concentration en lactate (de 0,00–39,18 mM à 0,00–109,40 mM).

De plus, nous pouvons également souligner que les erreurs RMSEP calculées pour chacun des modèles présentent des valeurs plus faibles lorsque les jeux de calibration reposent sur quatorze lots que lorsqu'ils sont basés sur huit. Ainsi, nous obtenons une erreur RMSEP de 5,28 mM pour le modèle de régression des concentrations en glucose (contre 5,68 mM précédemment) et une erreur RMSEP de 2,47 mM pour le modèle de régression des concentrations en lactate (contre 2,52 mM précédemment). Les performances prédictives de ces deux modèles sont représentées Figure 4.36.

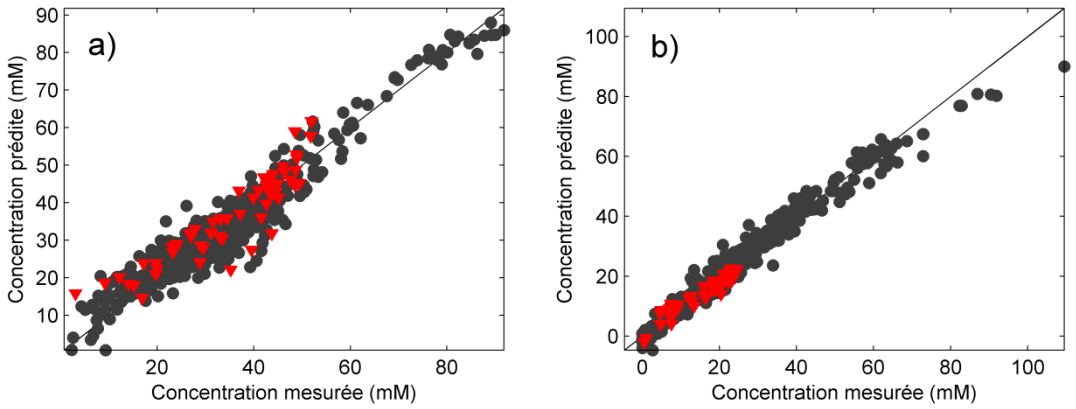

Figure 4.36 Représentations des modèles de régression développés pour a) la concentration en glucose et b) la concentration en lactate, à partir de quatorze cultures de cellules CHO. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Si nous reprenons les spectres acquis en continu pour la culture cellulaire #1, nous pouvons une nouvelle fois obtenir les prédictions faites par les différents modèles de régression tout au long du bioprocédé test. Ces prédictions sont présentées Figure 4.37.

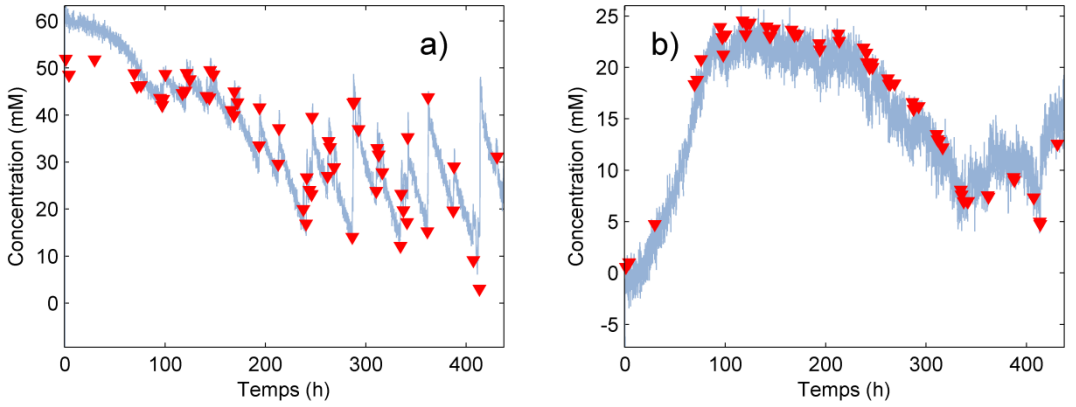

Figure 4.37 Représentations des prédictions des concentrations a) en glucose et b) en lactate réalisées sur les spectres acquis en continu de la culture test #1 par les modèles de régression développés sur les échantillons de quatorze cultures de cellules CHO.

Nous pouvons alors voir que les modèles de régression sont capables de suivre de manière satisfaisante les différentes concentrations. D'une part, le modèle de régression basé sur les concentrations en glucose est capable de suivre l'évolution du biomarqueur pendant la culture. Ces résultats ne diffèrent pas beaucoup de ce que nous avions pu obtenir en construisant un modèle PLS sur les huit cultures réalisées dans les conditions normales. Par contre, les résultats obtenus pour le modèle de régression du paramètre lactate présentent une amélioration de la robustesse du modèle. En effet, lors des prédictions en continu de la concentration en lactate sur les spectres acquis en continu pour la culture #1, nous avions pu observer des résidus importants après 300 h lorsque nous prenions en compte les huit lots acquis dans les conditions normales de culture au sein de l'ensemble de

calibration (Figure 4.3, page 134). Nous pouvons noter que lorsque les prédictions sont faites par le modèle de régression basé sur les quatorze cultures, ces résidus sont réduits, ce qui se traduit finalement par une meilleure erreur RMSEP. En augmentant la variabilité au sein de l'ensemble de calibration, nous avons finalement accru la robustesse du modèle de régression des concentrations en lactate.

2.2.2 Modèles pour les concentrations en glutamine et en glutamate

Le paramètre glutamine ne présente pas de changement de gamme de concentration lorsque nous ajoutons les six cultures atypiques à l'ensemble de calibration. Nous pouvons toutefois remarquer que l'amélioration des performances prédictives du modèle de régression, Tableau 4.13, passant d'une erreur RMSEP de 0,46 mM pour un modèle PLS basé sur huit lots à une erreur RMSEP de 0,37 mM pour un modèle à quatorze cultures. Ceci représente l'amélioration de la robustesse du modèle, représenté Figure 4.38a, non pas en étendant la gamme de variation du paramètre, mais en accumulant les variabilités provenant de différentes cultures au sein de l'ensemble d'apprentissage. Toutefois, en voyant le graphique des concentrations prédites en fonction des concentrations mesurées, il apparait que la majeure partie de l'erreur de prédiction provient de la gamme de concentration du lot test (0,00–3,56 mM) plus importante que la gamme de calibration (0,00–2,72 mM). De ce fait, les prédictions des concentrations les plus importantes sont des extrapolations, ce qui explique l'apparition de certains écarts pour ces valeurs.

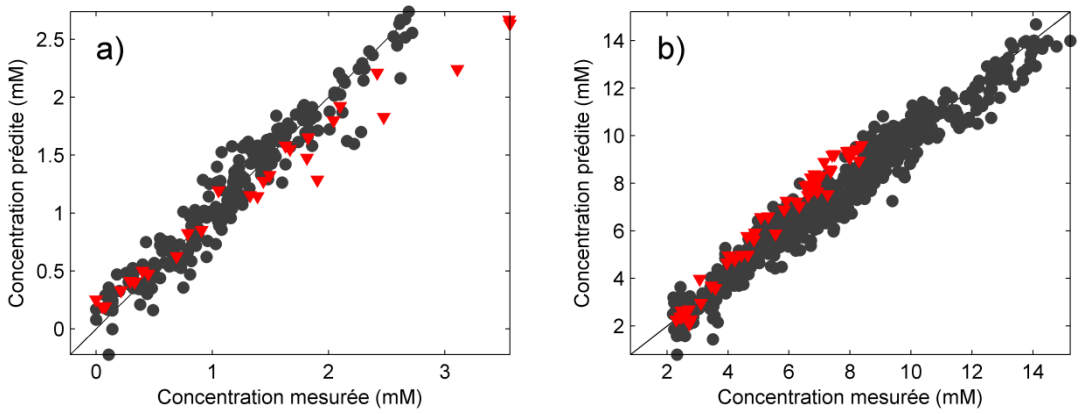

Figure 4.38 Représentations des modèles de régression développés pour a) la concentration en glutamine et b) la concentration en glutamate, à partir de quatorze cultures de cellules CHO. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Nous pouvons observer le biais de prédiction sur les fortes teneurs en glutamine en prenant également les prédictions des concentrations réalisées sur les spectres acquis en continu de la culture test, Figure 4.39a. Bien que la tendance des prédictions suive bien le profil de la concentration en glutamine dans le milieu de culture, une fois atteint les 200 h, la teneur mesurée sort alors de la gamme de calibration mise en jeu lors du développement du

modèle de régression (plus de 2,72 mM mesuré), ce qui traduit le début du biais sur les prédictions. Néanmoins, les prédictions obtenues à partir des spectres acquis en continu sur le début de la culture (lorsque les mesures de référence du lot test sont toujours comprises dans la gamme de calibration) sont très satisfaisantes.

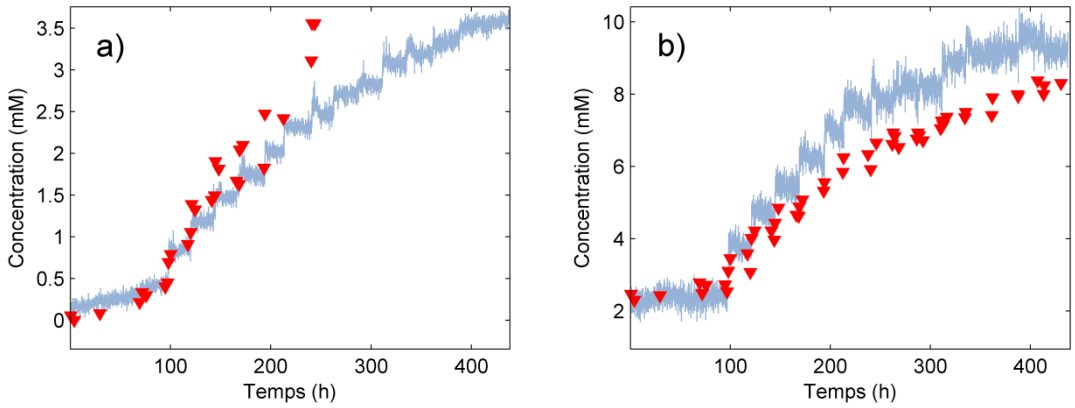

Figure 4.39 Représentations des prédictions des concentrations a) en glutamine et b) en glutamate réalisées sur les spectres acquis en continu de la culture test #1 par les modèles de régression développés sur les échantillons de quatorze cultures de cellules CHO.

Le modèle de régression calculé pour la concentration en glutamate est le seul à présenter une croissance de l'erreur RMSEP lorsque nous ajoutons les six lots réalisés dans des conditions de culture atypiques. En effet, Tableau 4.13, l'erreur de prédiction évolue de 0,41 mM pour le modèle basé sur huit cultures à 1,00 mM lorsque nous avons quatorze lots. Ces variations s'expliquent en partie à cause des changements de gamme d'étalonnage. En effet, nous pouvons constater une augmentation de près de 12 %, passant de 2,21–13,67 mM à 2,19–15,22 mM. Mais l'augmentation de l'erreur de prédiction vient surtout du fait que le modèle de régression développé sur les huit lots acquis dans les conditions normales convenait parfaitement à la prédiction des échantillons de la culture test. Pour rappel, nous avions alors une erreur RMSEP (0,41 mM) inférieure à l'erreur RMSECV (0,63 mM). Ainsi, en ajoutant les échantillons des cultures atypiques, nous changeons le modèle de régression qui n'explique plus parfaitement le lot test. Nous pouvons néanmoins voir, Figure 4.39b, que le modèle de régression parvient à suivre l'évolution du profil métabolique du glutamate malgré la présence de quelques écarts sur les prédictions.

2.2.3 Modèles pour les densités VCD et TCD

Les mesures de densité cellulaire pour les cultures de cellules CHO présentent des profils extrêmement similaires. Ainsi, les comportements des deux modèles sont identiques. Nous pouvons d'abord voir que les erreurs de prédiction sont plus faibles pour les modèles de régression prenant en compte les quatorze cultures de cellules. En effet, les erreurs

RMSEP des densités évoluent de 1,10 mM à 0,94 mM pour la densité VCD et de 1,13 mM à 0,94 mM pour la densité TCD. Les deux modèles obtenus sont représentés Figure 4.40.

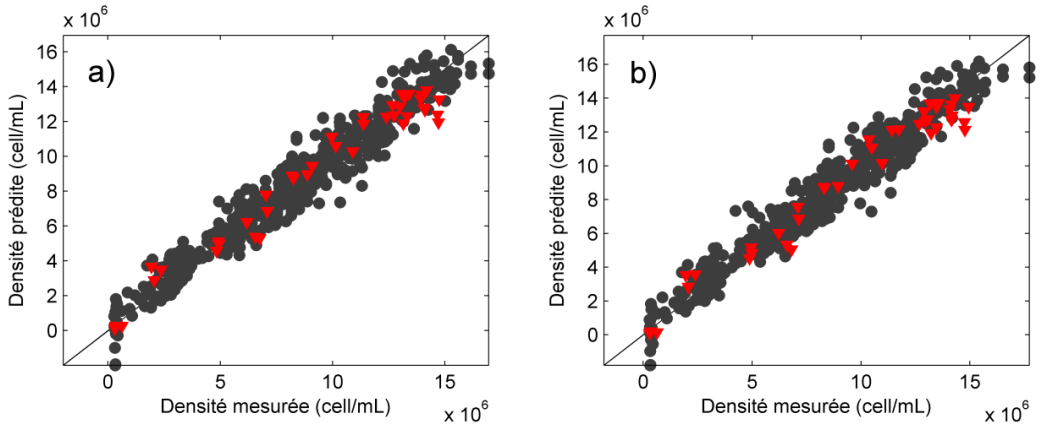

Figure 4.40 Représentations des modèles de régression développés pour a) la densité en cellules vivantes VCD et b) la densité en cellules totale TCD, à partir de quatorze cultures de cellules CHO. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Mais l'amélioration la plus notable pour ces deux modèles de régression apparait sur les prédictions des spectres acquis en continu pour la culture test, présentées Figure 4.41. En comparant ces résultats aux prédictions continues réalisées précédemment à partir des modèles de régression élaborés sur huit lots (Figure 4.7), nous pouvons voir que les prédictions des spectres acquis lorsque la densité cellulaire de la culture atteint la phase de palier présentent moins de biais. Ceci provient directement du fait que nous ayons accumulé plus de cultures avoisinant les fortes densités, ce qui a donc permis d'ajouter aux modèles de régression plus de variabilité dans ces gammes de densité cellulaire.

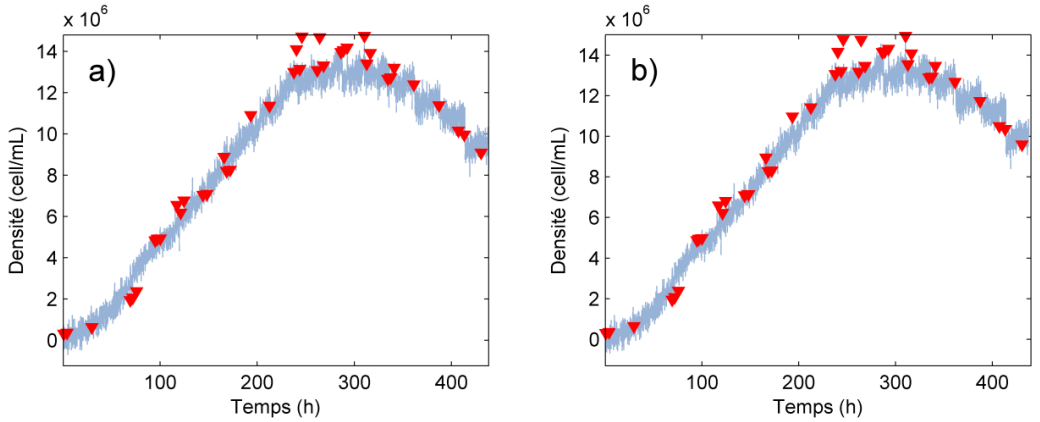

Figure 4.41 Représentations des prédictions des densités a) VCD et b) TCD réalisées sur les spectres acquis en continu de la culture test #1 par les modèles de régression développés sur les échantillons de quatorze cultures de cellules CHO.

Pour conclure quant à l'intégration des données provenant de cultures atypiques, nous pouvons finalement dire que celles-ci présentent un intérêt certain pour accroitre la robustesse des modèles PLS. En effet, puisque les cultures acquises dans des conditions

variant par rapport à la normale présentent des gammes métaboliques différentes, la variabilité apportée à l'ensemble de calibration permet ensuite d'explorer un éventail de concentrations qui n'auraient pas été accessibles lors de cultures cellulaires réalisées dans les conditions normales. Ainsi, ces variations permettent d'une part d'accroitre la robustesse des modèles en ajoutant des données à l'ensemble d'apprentissage, et d'autre part de présenter au modèle de régression des comportements possibles si la culture venait à dévier. Autrement dit, si la culture cellulaire présentait des modifications métaboliques, imperceptibles sans recours aux méthodes de référence, les modèles de régression seraient plus à même de prédire de manière satisfaisante les concentrations métaboliques si nous introduisons au préalable les échantillons des cultures atypiques

3 Génération d'un modèle de régression global pour l'étude de plusieurs types de cellules

Jusqu'ici, nous avons développé les modèles de régression spécifiquement à chaque type de cellule. En effet, les comportements métaboliques étant différents et propres à chaque cellule, il était important de répartir les données provenant des cultures de divers types de cellule afin de respecter les diverses tendances biochimiques et de vérifier séparément la faisabilité du suivi métabolique. Néanmoins, il peut être intéressant de combiner les données provenant de ces différentes cultures, cellulaires et virales, afin d'obtenir des modèles de régression généraux.

3.1 Développement de modèles compatibles pour plusieurs types de cellule

Afin de développer des modèles de régression pour la détermination des paramètres métaboliques de différents types de cellule, il convient donc de mutualiser les données provenant de différentes cultures au sein d'un même ensemble de calibration. Ainsi, nous reprenons les échantillons des cultures présentés dans les sections précédentes, résumés Tableau 4.14. Nous reprenons également les mêmes cultures test que lors des travaux sur les cellules individuelles. Ainsi, l'ensemble d'apprentissage est constitué de 22 cultures cellulaires, soit 14 cultures de cellules CHO, 4 cultures de cellules HeLa et enfin 4 cultures de cellules Sf9 (1 amplification cellulaire et 3 cultures virales). Nous constituons donc un ensemble test composé de 6 cultures cellulaires, soit 1 culture de cellules CHO, 3 cultures de cellules HeLa et 2 cultures de cellules Sf9 (1 amplification cellulaire et 1 culture virale). Le fait de composer l'ensemble test de cultures de différents types de cellule permet de vérifier si les modèles de régression sont applicables aux différentes lignées mises en jeu.

Tableau 4.14 Récapitulatif des données acquises pour toutes les cultures de cellules CHO, HeLa ou Sf9 mises en jeu pour les modélisations chimiométriques. Les cultures test sont grisées.

	Numéro de culture	Nombre d'échantillons		
	Cellules CHO			
	#1	57		
	#2	17		
	#3	51		
	#4	48		
	#5	62		
	#6	62		
	#7	45		
	#8	43		
	#9	43		
pH différent	#10	42		
pH différent	#11	43		
pH différent	#12	43		
pO_2 différente	#13	43		
pO_2 différente	#14	42		
Agitation différente	#15	43		
	Cellules HeLa		*Pré-*	*Post-*
	#1	14	*8*	*6*
	#2	32	*19*	*13*
	#3	70	*40*	*30*
	#4-1	73	*39*	*34*
	#4-2	73	*39*	*34*
	#5-1	22	*7*	*15*
	#5-2	22	*7*	*15*
	#6	34	*20*	*14*
	#7	26	*16*	*10*
	Cellules Sf9			
	Amplification cellulaire			
	#1-1	42		
	#1-2	43		
	#2	50		
	Culture virale			
	#3-1	42		
	#3-2	48		
	#4	63		
	#5	61		
	#6	63		

Les modèles de régression sont développés pour les concentrations en glucose, glutamine, glutamate et lactate, ainsi que pour les densités totales (TCD) et en cellules vivantes (VCD). Tous les spectres Raman disponibles sont traités de la même façon, à savoir une dérivée première (Savitzky-Golay, fenêtre mobile de 15 points et polynôme d'ordre 2), suivie d'une normalisation SNV. Nous sélectionnons ensuite les régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Ensuite, les modèles de régression sont élaborés en réalisant une validation croisée. Nous appliquons ensuite les modèles de régression sur l'ensemble constitué des échantillons des six cultures test afin d'apprécier les performances prédictives des modèles PLS, présentées Tableau 4.15.

Tableau 4.15 Récapitulatif des performances des modèles de régression développés sur la base d'échantillons provenant de cultures de différentes cellules.

Biomarqueur (unité)		**R^2**	**LV**	**RMSECV**	**RMSEP**	**Gamme**
Glucose (g/L)		0,933	13	0,71	1,19	0,09 – 16,54
Glutamine (mM)		0,935	12	0,58	1,33	0,00 – 11,76
Glutamate (mM)		0,964	11	0,71	1,28	0,00 – 15,22
Lactate (g/L)		0,960	11	0,25	0,33	0,00 – 7,82
VCD (10^6 cell/mL)	Ampli. Cell.	*(0,707)** 0,954	13	1,04	*(6,48)** 1,02	0,29 – 16,94
	Cult. Vir.	0,914	13	0,39	0,88	0,68 – 4,38
TCD (10^6 cell/mL)	Ampli. Cell.	*(0,721)** 0,956	13	1,05	*(6,21)** 1,00	0,29 – 17,72
	Cult. Vir.	0,924	10	0,34	0,70	1,41 – 5,39

** Valeurs obtenues sans exclure le lot Sf9 #2 de l'ensemble test.*

Le premier point que nous pouvons soulever à travers ces résultats est le grand nombre de variables latentes prises en compte par les modèles PLS. Ceci est tout-à-fait cohérent avec ce que nous avions pu voir précédemment : plus nous avons de variabilité à prendre en compte, plus l'espace des modèles est élevé. C'est ici le cas puisque nous devons intégrer les variations entre les différents types de cellule, considérant donc différentes évolutions métaboliques et différents milieux de culture.

3.1.1 Modèles pour les concentrations en glucose et en glutamine

Pour commencer l'analyse des modèles PLS obtenus, nous commençons par prendre en compte les paramètres glucose et glutamine puisqu'il s'agit des paramètres

principalement consommés par les cellules. En plus des résultats présentés Tableau 4.15, les deux modèles de régression sont représentés Figure 4.42 à travers les graphiques des concentrations prédites en fonction des concentrations de référence, ainsi que leurs coefficients de régression.

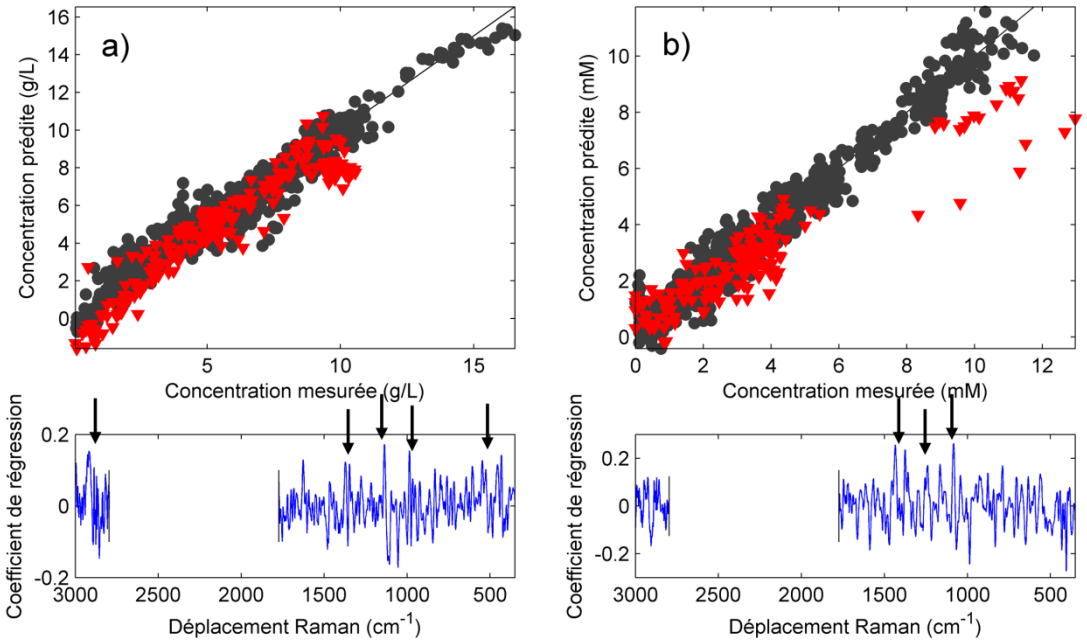

Figure 4.42 Représentations des résultats obtenus pour les modèles PLS basés sur les cultures de cellules CHO, HeLa et Sf9 pour a) la concentration en glucose et b) la concentration en glutamine. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Tout d'abord, en ce qui concerne le modèle de régression PLS calculé pour les concentrations en glucose, nous pouvons noter que l'erreur RMSEP calculée (1,19 g/L) reste satisfaisante compte tenu de la gamme de calibration (0,09–16,54 g/L), malgré un écart par rapport à l'erreur RMSECV (0,71 g/L). Cet écart provient surtout d'un biais important sur la prédiction des échantillons de la culture de cellules Sf9 #2, présenté Figure 4.43e. En effet, nous avions déjà rencontré ce phénomène en ne mettant en jeu que les cellules Sf9.

En voyant les prédictions réalisées sur les spectres acquis en continu, nous pouvons vraisemblablement dire que le modèle de régression est en mesure de prédire la concentration en glucose quelque soit la lignée cellulaire, malgré quelques écarts pour les lignées HeLa (Figure 4.43b-d). De plus, si nous prenons les coefficients de régression (Figure 4.42a), nous retrouvons les marqueurs importants du glucose entre 3000–2800 cm^{-1} (élongations ν(C–H) [146]), autour de 1300 cm^{-1} (déformations δ(CH_2) et δ(CH_2OH) [137]), autour de 1100 cm^{-1} (élongations ν(C–O) et ν(C–C) [137]), entre 1000–900 cm^{-1} (déformations δ(COH), δ(CCH) et δ(OCH) [137]) et enfin entre 550–450 cm^{-1} (déformations exocycliques et endocycliques [137]), qui valident la fiabilité du modèle.

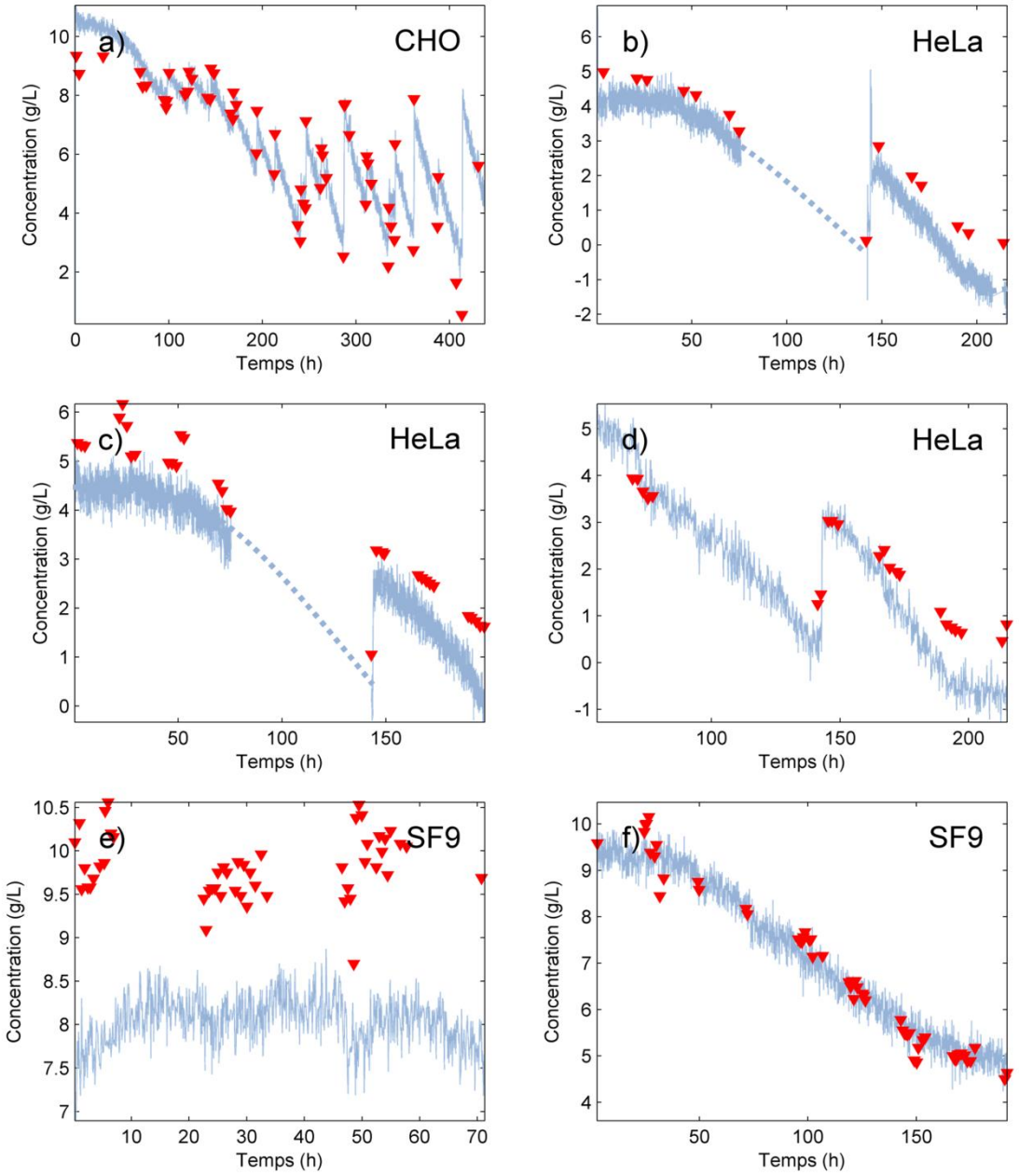

Figure 4.43 Représentations des prédictions réalisées par le modèle de régression des concentrations en glucose développé à partir des échantillons de cultures CHO, HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) CHO #1, b) HeLa #1, c) HeLa #2, d) HeLa #5, e) Sf9 #2 et f) Sf9 #4.

Le modèle de régression pour la concentration en glutamine présente quant à lui quelques différences dans sa conception. En effet, lors des travaux sur les modèles de régression cumulant les échantillons de plusieurs cultures de cellules Sf9, nous avons pu montrer des écarts importants lors des prédictions des concentrations en glutamine qui traduisaient un manque de robustesse (section 1.3.1 de ce chapitre, page 161). Ainsi, afin d'apporter plus de données provenant des cellules HeLa, nous avons intégré le lot test Sf9 #2 à l'ensemble d'apprentissage. En considérant les performances prédictives du modèle ainsi développé (Tableau 4.15) nous pouvons noter une erreur RMSEP (1,33 mM) faible

compte tenu de la gamme de variation comprise entre 0,00 mM et 11,76 mM, mais importante par rapport à l'erreur RMSECV (0,58 mM) obtenue lors de la validation croisée initiale. Nous pouvons noter qu'il existe un biais certain sur les prédictions des échantillons des cultures test employée ici.

Si nous considérons les coefficients de régression du modèle de régression des concentrations en glutamine, Figure 4.42b, nous pouvons noter (tout comme les coefficients de régression du modèle basé sur les concentrations en glucose) un niveau de bruit relativement élevé. Néanmoins, certaines bandes caractéristiques de la molécule de glutamine ressortent, notamment autour de 1400 cm^{-1} et 1090 cm^{-1} caractérisant les bandes d'élongations ν(C–N) des groupements amine [140] et autour de 1300 cm^{-1} représentant les contributions des déformations δ(C–H) de la chaine carbonée [140].

Les prédictions réalisées sur les spectres acquis en continu des cultures du jeu test sont présentées Figure 4.44. Nous pouvons d'abord noter l'importance du bruit observé sur les coefficients de régression qui apparaît sur les prédictions notamment pour les cultures cellulaires présentant peu de variations des niveaux de glutamine, ce qui est notamment le cas des lots CHO #1, HeLa #1, HeLa #2 et HeLa #5 (Figure 4.44a-d). En effet, ces cultures présentent des gammes de variation en glutamine très faibles comparées à la gamme de calibration, ce qui entraine les écarts importants entre les prédictions successives. De plus, nous pouvons souligner des différences dans les profils de prédiction par rapport aux tendances attendues d'après les mesures de référence et aux résultats obtenus en travaillant sur les modèles de chaque type de cellule (Figure 4.39a, page 190, pour les cellules CHO et Figure 4.10b, page 145, pour les cellules HeLa). Par contre, les prédictions des échantillons de la culture test Sf9 #4, Figure 4.44e sont acceptables, malgré certains écarts en début de culture. Ceci vient directement des gammes de concentration mises en jeu pour chaque type de cellule. En effet, pour les cellules CHO et HeLa, les gammes de concentration en glutamine sont de 0,00–2,72 mM et 0,00–5,44 mM respectivement, tandis que la gamme de variation de la concentration en glutamine évolue entre 1,63–10,53 mM pour les cultures de cellules Sf9.

Ceci traduit simplement le fait que le modèle de régression pour les concentrations en glutamine s'appuie principalement sur les échantillons des cultures de cellules Sf9 et n'est donc pas représentatif des autres cultures. Bien que nous possédions plus d'échantillons pour les cellules CHO, le manque de variation de la concentration en glutamine empêche de développer un modèle PLS capable de prédire convenablement les teneurs en glutamine de plusieurs cultures.

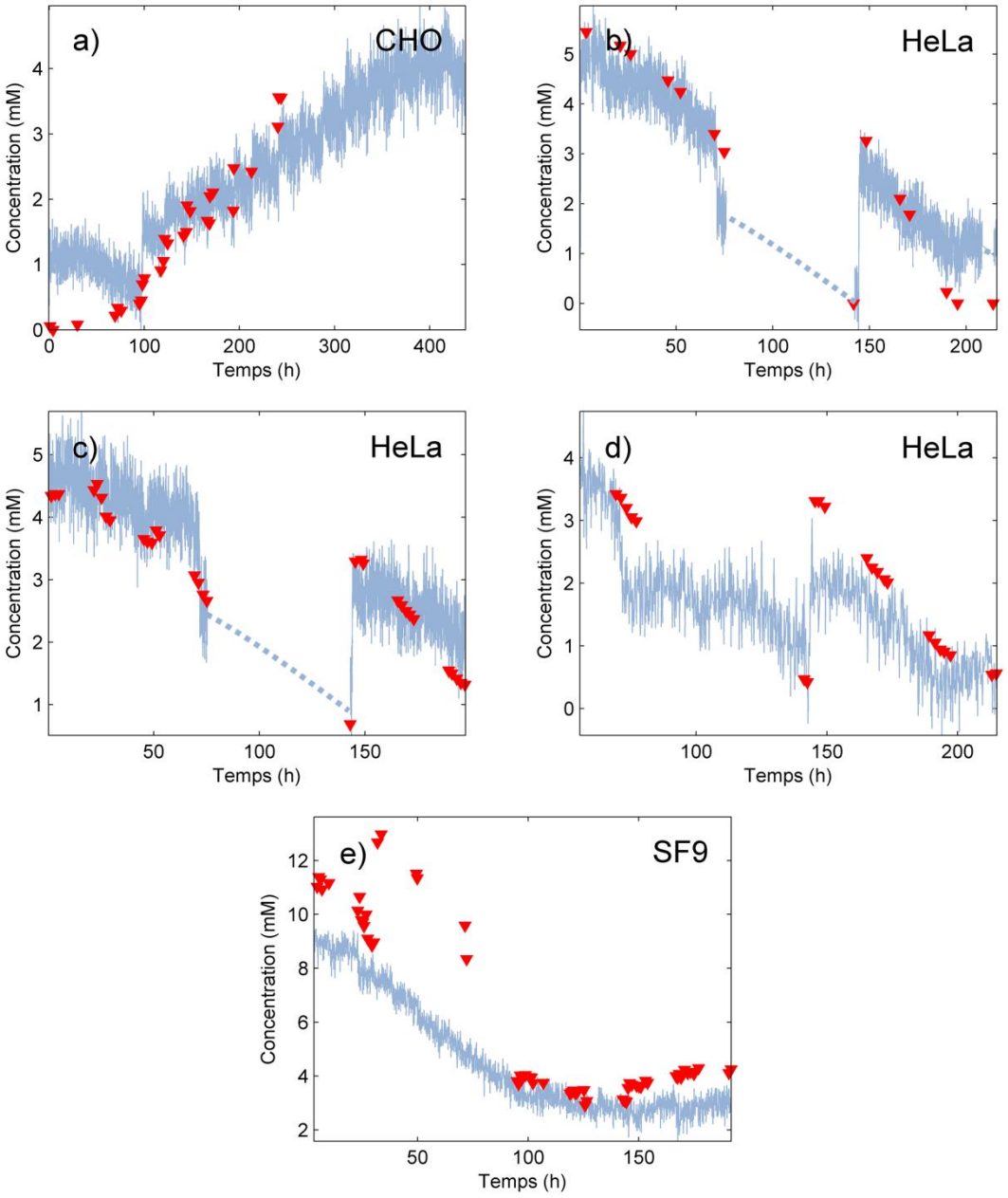

Figure 4.44 Représentations des prédictions réalisées par le modèle de régression des concentrations en glutamine développé à partir des échantillons de cultures CHO, HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) CHO #1, b) HeLa #1, c) HeLa #2, d) HeLa #5 et e) Sf9 #4.

3.1.2 Modèles pour les concentrations en glutamate et en lactate

Les deux paramètres étudiés ici sont les concentrations des deux molécules principalement produites lors des cultures. Nous commençons par regarder les performances des modèles de régression pour les concentrations en glutamate, Figure 4.45a. À première vue, le modèle PLS présente des performances prédictives satisfaisantes, semblables au modèle de régression glutamine, avec une erreur RMSEP évaluée à 1,28 mM

pour une gamme de concentration comprise entre 0,00–15,22 mM. Néanmoins, cette erreur de prédiction reste tout de même importante compte tenu de l'erreur RMSECV, calculée à 0,71 mM, soit presque moitié moins que l'erreur de prédiction.

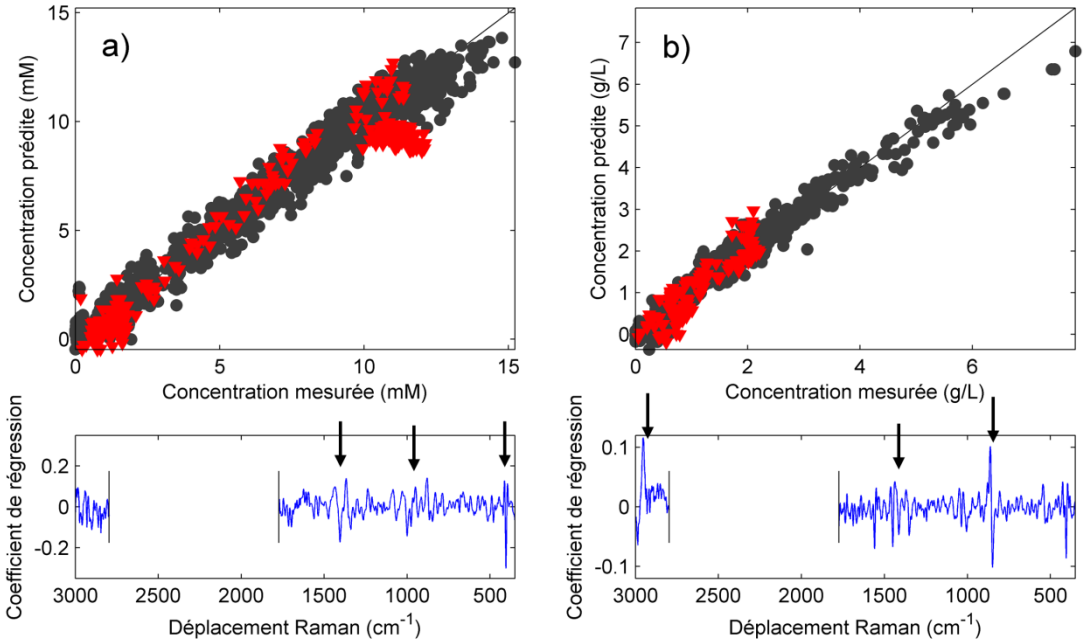

Figure 4.45 Représentations des résultats obtenus pour les modèles PLS basés sur les cultures de cellules CHO, HeLa et Sf9 pour a) la concentration en glutamate et b) la concentration en lactate. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Si nous prenons les coefficients de régression calculés pour ce modèle, nous retrouvons les bandes comprises entre 1500–1350 cm^{-1} caractérisant les déformations δ(CH2) [137] et surtout l'élongation symétrique $v(CO_2^-)$ de la molécule [137]. La région 1000–900 cm^{-1} représente les contributions spectrales apportées par les vibrations δ(O–H) [137] et v(C–N) [140]. Enfin, les signaux autour de 400 cm^{-1} peuvent être attribués aux déformations de la chaine carbonée de la molécule de glutamate [140].

En plus des prédictions sur les jeux test, nous présentons les prédictions réalisées sur les spectres acquis en continu pour les cultures test des trois types de cellule, Figure 4.46. Bien que nous ayons vu les coefficients de régression puissent caractériser la molécule de glutamate, nous pouvons voir que d'importantes erreurs de prédictions pour les lots mettant en jeu des cellules HeLa et Sf9. En effet, nous pouvons noter des profils de prédiction incohérents par rapport aux tendances métaboliques attendues, notamment pour les cultures HeLa (Figure 4.46b-d) et Sf9 (Figure 4.46e-f). En revanche, nous pouvons voir que, malgré un léger biais, les prédictions obtenues pour les spectres de la culture de cellules CHO test sont satisfaisantes.

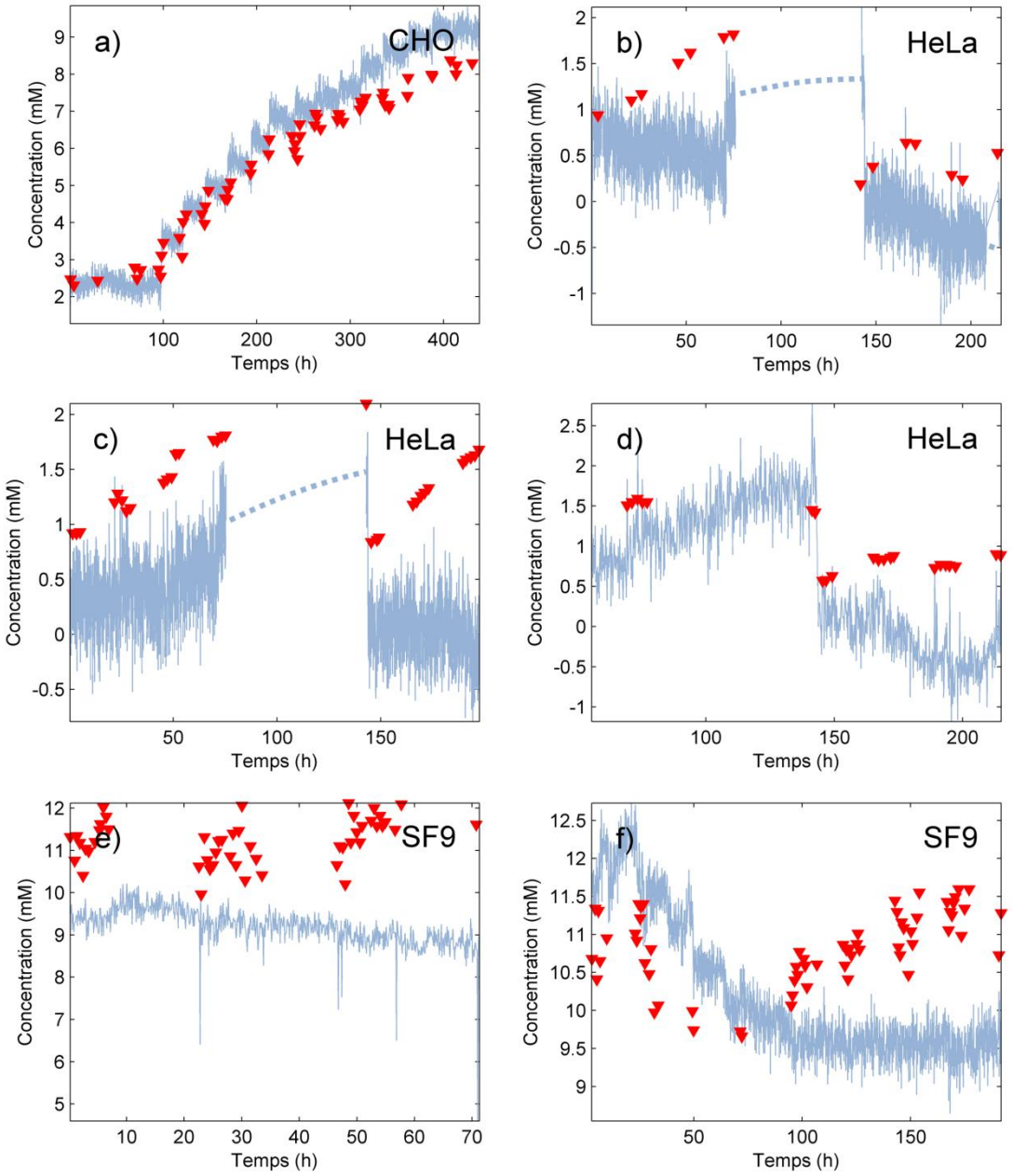

Figure 4.46 Représentations des prédictions réalisées par le modèle de régression des concentrations en glutamate développé à partir des échantillons de cultures CHO, HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) CHO #1, b) HeLa #1, c) HeLa #2, d) HeLa #5, e) Sf9 #2 et f) Sf9 #4.

Dans les faits, les problèmes de prédiction ont la même origine que ceux observés lors de la modélisation du paramètre glutamine. Les gammes de concentration des cultures de cellules HeLa et Sf9 (0,00–2,55 mM et 9,28–13,06 mM respectivement) sont trop faibles comparées à celle des cultures de cellules CHO (2,19–15,22 mM). Nous avons donc, tout comme pour les concentrations en glutamine précédemment, les échantillons des cultures d'un seul type de cellule qui prennent l'ascendant sur le modèle de régression, induisant donc des biais de prédiction importants sur les échantillons des autres cultures.

En ce qui concerne le modèle de prédiction des concentrations en lactate, nous rappelons tout d'abord qu'aucune référence provenant des cultures de cellules Sf9 n'a été prise en compte dans la réalisation des modèles puisque ces-dernières présentent des niveaux en dessous de la limite de quantification. Nous impliquerons tout de même les cultures de cellules Sf9 en réalisant les prédictions des spectres acquis en continu des lots test pour ce type de cellule.

En prenant d'abord en compte les performances prédictives du modèle PLS, Tableau 4.15, nous pouvons voir que celles-ci sont satisfaisantes étant donné que l'erreur RMSEP (évaluée à 0,33 g/L) est proche de l'erreur RMSECV (évaluée à 0,25 g/L) calculée lors de la validation croisée effectuée lors du développement du modèle. Dans le cas des concentrations en lactate, l'avantage de la modélisation réside dans l'homogénéité des gammes de concentration mises en jeu pour les cultures de cellules HeLa et CHO. Toutefois, il est important de noter que malgré cette similitude, les profils d'évolution des concentrations en lactate diffèrent selon les cultures cellulaires et les types de cellule mis en jeu.

Les coefficients de régression obtenus pour le modèle PLS des concentrations en lactate sont présentés Figure 4.45b et permettent de très bien caractériser la molécule. En effet, nous pouvons d'abord noter la présence d'un signal majoritaire présent à 860 cm^{-1} représentant les contributions des bandes d'élongation $\nu(C–CO_2^-)$ du groupement carboxylate du lactate [138]. De plus, nous pouvons noter la présence de signaux dans la région 3000–2800 cm^{-1} correspondant aux bandes de vibration $\nu(CH_3)$ du groupement méthyle [138,146]. Enfin, les signaux compris entre 1550–1350 cm^{-1} comprennent entre autres les élongations $\nu(CO_2^-)$ [138]. Ces coefficients de régression justifient donc les performances prédictives satisfaisantes observées précédemment.

Si nous prenons en compte les prédictions réalisées par ce modèle sur les spectres acquis en continu pour toutes les cultures test (Figure 4.47), nous pouvons voir que nous sommes en mesure de suivre les différents profils de la concentration en lactate suivant le type de culture pris en compte. Tout d'abord, le profil obtenu pour la culture de cellules CHO test est très satisfaisant et permet de bien suivre le profil métabolique du lactate tout au long de la culture. Ensuite, si nous prenons les différentes cultures de cellules HeLa, nous pouvons observer certains écarts des prédictions, notamment pour la culture #2 dues à de très faibles variations de la gamme. Néanmoins, nous observons bien les profils d'évolution du lactate pour les cultures de cellules HeLa.

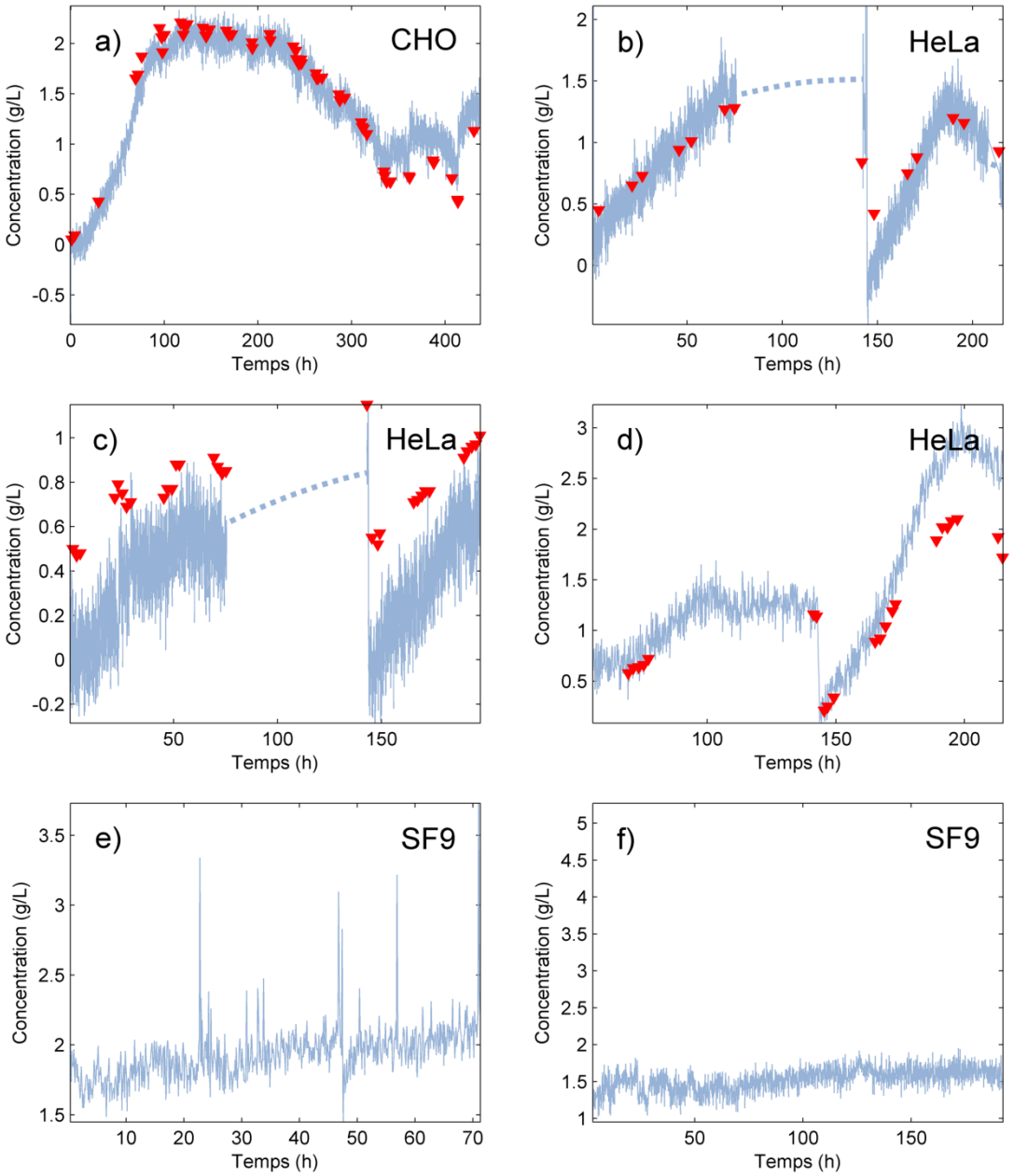

Figure 4.47 Représentations des prédictions réalisées par le modèle de régression des concentrations en lactate développé à partir des échantillons de cultures CHO et HeLa et appliqué sur les spectres acquis en continu pour les lots test a) CHO #1, b) HeLa #1, c) HeLa #2, d) HeLa #5, e) Sf9 #2 et f) Sf9 #4.

Le marqueur le plus significatif des bonnes performances du modèle PLS pour les concentrations en lactate provient surtout des profils obtenus pour les cultures de cellules Sf9, Figure 4.47e-f. En effet, bien que nous obtenions vraisemblablement un biais de prédiction de l'ordre de 1,50 g/L, principalement dû à l'absence de cultures de cellules Sf9 au sein du jeu de calibration, nous pouvons remarquer la faible croissance des profils de prédiction traduisant bien, non pas l'absence de lactate dans le milieu, mais la faible production des cellules Sf9. Nous notons que la limite de quantification de la mesure de la concentration en lactate est estimée à 0,04 g/L.

3.1.3 Modèles pour les densités VCD et TCD

Les modèles PLS des densités cellulaires ont été une nouvelle fois été développés sur la base de deux ensembles. Le premier comprend tous les échantillons acquis durant les cultures cellulaires, soit avant d'introduire du virus au sein des cultures. Nous avons ainsi les cultures de cellules CHO, le début des procédés HeLa ainsi que les phases d'amplification cellulaire des cultures de cellules Sf9. Le second met en jeu les échantillons obtenus durant les cultures virales, à savoir les fins des bioprocédés HeLa et les cultures virales Sf9 (Tableau 4.14).

Sur la phase de croissance cellulaire, nous notons d'abord que les profils des densités cellulaires VCD et TCD sont extrêmement similaires. Ainsi, les observations sont les mêmes, qu'il s'agisse de la densité en cellules vivantes ou la densité cellulaire totale. Les représentations des modèles obtenus pour les modèles PLS des densités avant inoculation du virus sont disponibles Figure 4.48. Nous pouvons noter la présence d'échantillons complètement surévalués qui appartiennent à la culture de cellules Sf9 #2, bien que le modèle comporte certains lots des cultures Sf9. C'est pourquoi nous avons évalué deux erreurs de prédictions différentes, une première (entre parenthèses Tableau 4.15) prenant en compte les prédictions biaisées et une seconde (deuxième valeur Tableau 4.15) excluant donc les échantillons de la culture Sf9 #2.

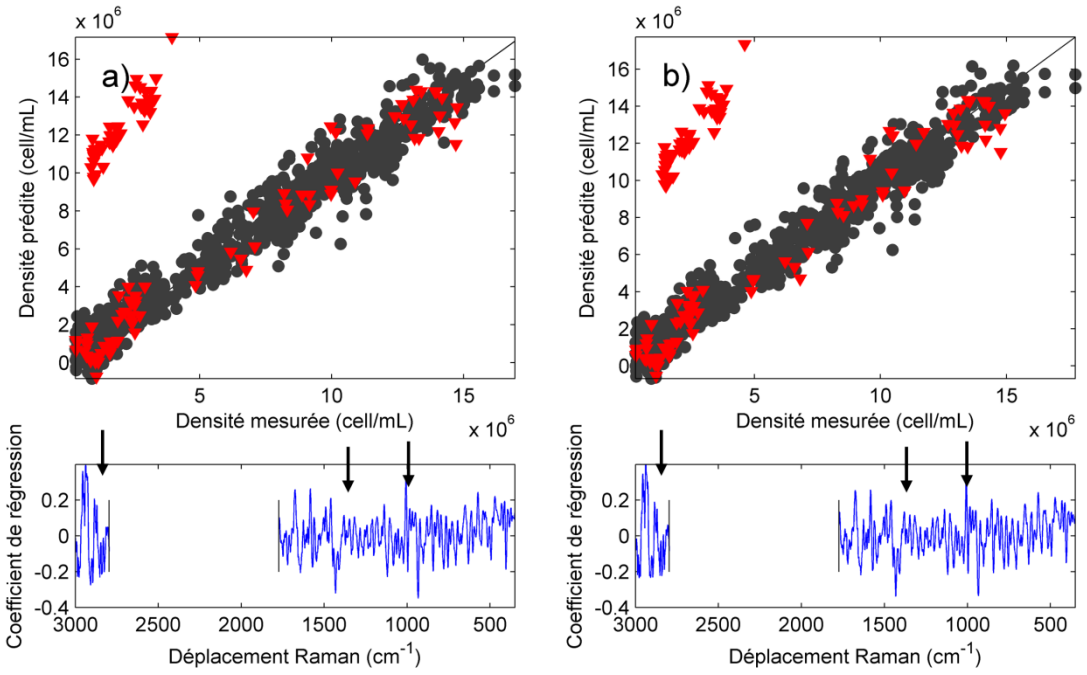

Figure 4.48 Représentations des résultats obtenus pour les modèles PLS basés sur les cultures de cellules CHO, HeLa et Sf9 pour a) la densité en cellules vivantes VCD et b) la densité en cellules totale TCD avant introduction du virus dans le milieu. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Sans prendre en compte les échantillons de la culture test Sf9 exclue, nous pouvons noter que les deux modèles de régression présentent des performances prédictives satisfaisantes, dont une erreur RMSEP de $1{,}02 \cdot 10^6$ cell/mL contre une erreur RMSECV évaluée à $1{,}04 \cdot 10^6$ cell/mL (pour une gamme de densité comprise entre $0{,}29 \cdot 10^6$ cell/mL et $16{,}94 \cdot 10^6$ cell/mL) pour la densité en cellules vivantes VCD, et une erreur RMSEP de $1{,}00 \cdot 10^6$ cell/mL contre une erreur RMSECV calculée à $1{,}05 \cdot 10^6$ cell/mL (pour une gamme comprise entre $0{,}29 \cdot 10^6$ cell/mL et $17{,}72 \cdot 10^6$ cell/mL) pour la densité en cellules totale TCD.

Si nous considérons les coefficients des deux modèles, qui exposent des structures similaires, nous pouvons dégager trois principales régions spectrales. La première se situe entre 3000–2800 cm^{-1}, qui traduit la présence des bandes de vibration v(C–H) de diverses molécules [146]. La deuxième région autour de 1450–1300 cm^{-1} caractérise principalement la respiration des cycles des molécules d'adénine et de guanine constitutives des macromolécules d'ADN [137,142]. Enfin, la bande située autour de 1000 cm^{-1} traduit la présence de la phénylalanine à travers les élongations v(C–C) du noyau benzénique de la molécule [142]. Ces observations appuient les performances prédictives satisfaisantes observées précédemment.

Nous représentons les prédictions des spectres acquis en continu pour les cultures test basées sur les cellules CHO (#1), HeLa (#1, #2 et #5) et Sf9 (#2), Figure 4.49 pour la densité en cellules vivantes et Figure 4.50 pour la densité totale en cellules, nous pouvons voir que les modèles permettent de bien capter les tendances exponentielles de la multiplication cellulaire durant les cultures, notamment pour les types de cellule CHO et HeLa. En effet, nous pouvons voir que les prédictions parviennent à très bien suivre l'évolution des différentes densités pour les cellules CHO à travers la superposition des prédictions et des mesures réalisées à l'aide des méthodes de référence. Pour les cellules HeLa, bien que la tendance soit très bien respectée, nous pouvons observer un léger biais apparent en début de culture, période durant laquelle le taux de cellule a tendance à être sous-évalué, notamment pour les cultures HeLa test #1 et #2. Enfin, pour les cellules Sf9, nous pouvons clairement observer l'important biais de prédiction qui avait été souligné lors des prédictions des échantillons du lot test Sf9 #2.

Nous pouvons donc finalement considérer que le modèle de prédiction développé sur la base des mesures de densités cellulaires permet de prédire de manière satisfaisante les échantillons des cultures CHO et HeLa, mais présentent d'importantes erreurs pour les lots de cellules Sf9. Ceci parait cohérent quant aux natures des cellules mises en jeu. En effet, les cellules HeLa, bien que cancéreuses, sont des cellules mammifères, tout comme les

cellules CHO, tandis que les cellules Sf9 proviennent directement d'insectes et présentent donc des caractéristiques biologiques différentes.

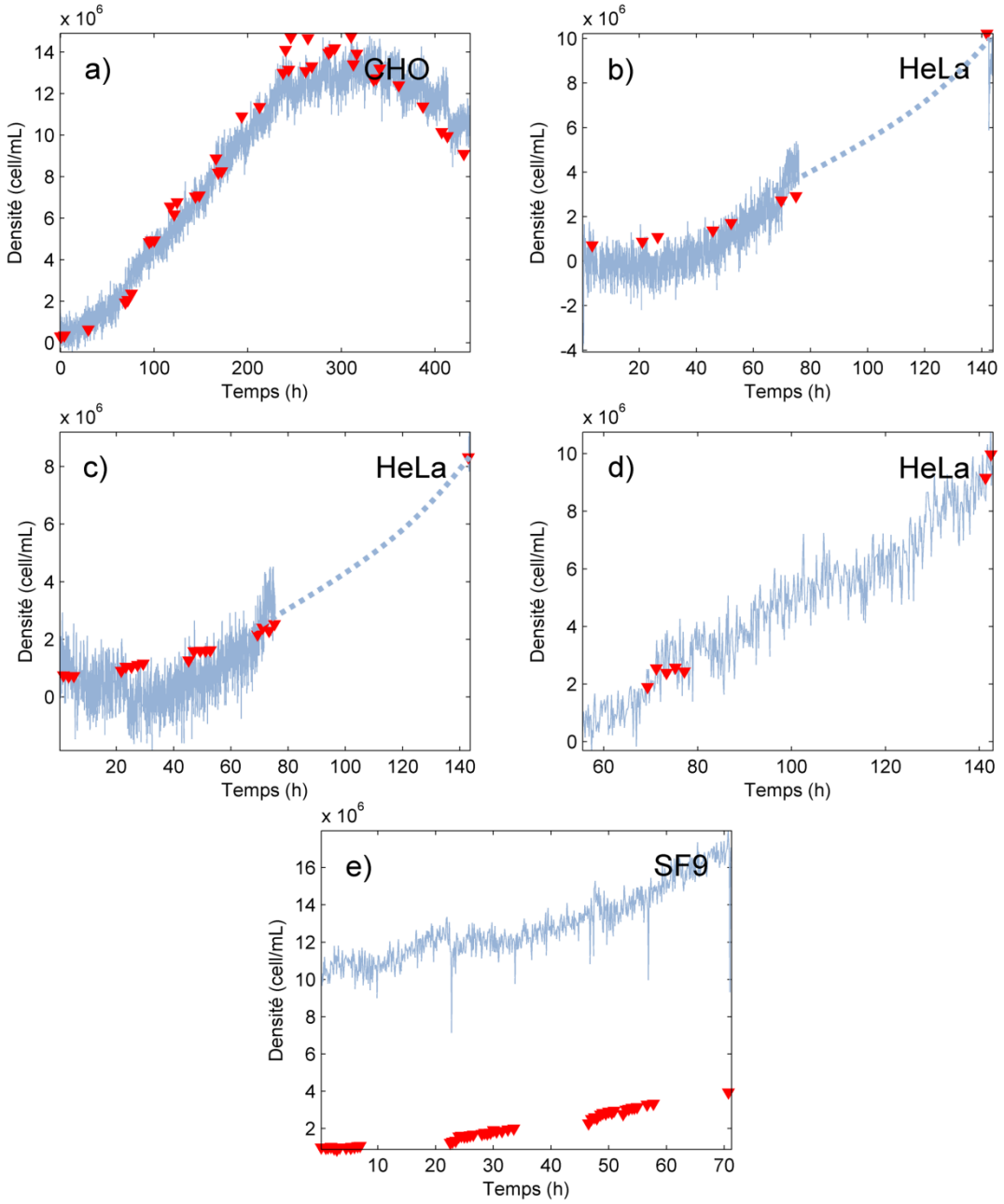

Figure 4.49 Représentations des prédictions réalisées par le modèle de régression densités VCD (durant la phase de culture cellulaire) développé à partir des échantillons de cultures CHO, HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) CHO #1, b) HeLa #1, c) HeLa #2, d) HeLa #5 et e) Sf9 #2.

Nous pouvons ensuite considérer les modèles de régression développés sur les densités VCD et TCD mesurées durant la période de culture virale suite à l'infection des cultures. Pour ces modèles PLS, nous ne pouvons pas prendre en compte les cultures de cellules CHO puisque ces-dernières sont utilisées, non pas pour la culture de virus, mais pour la culture d'anticorps recombinants.

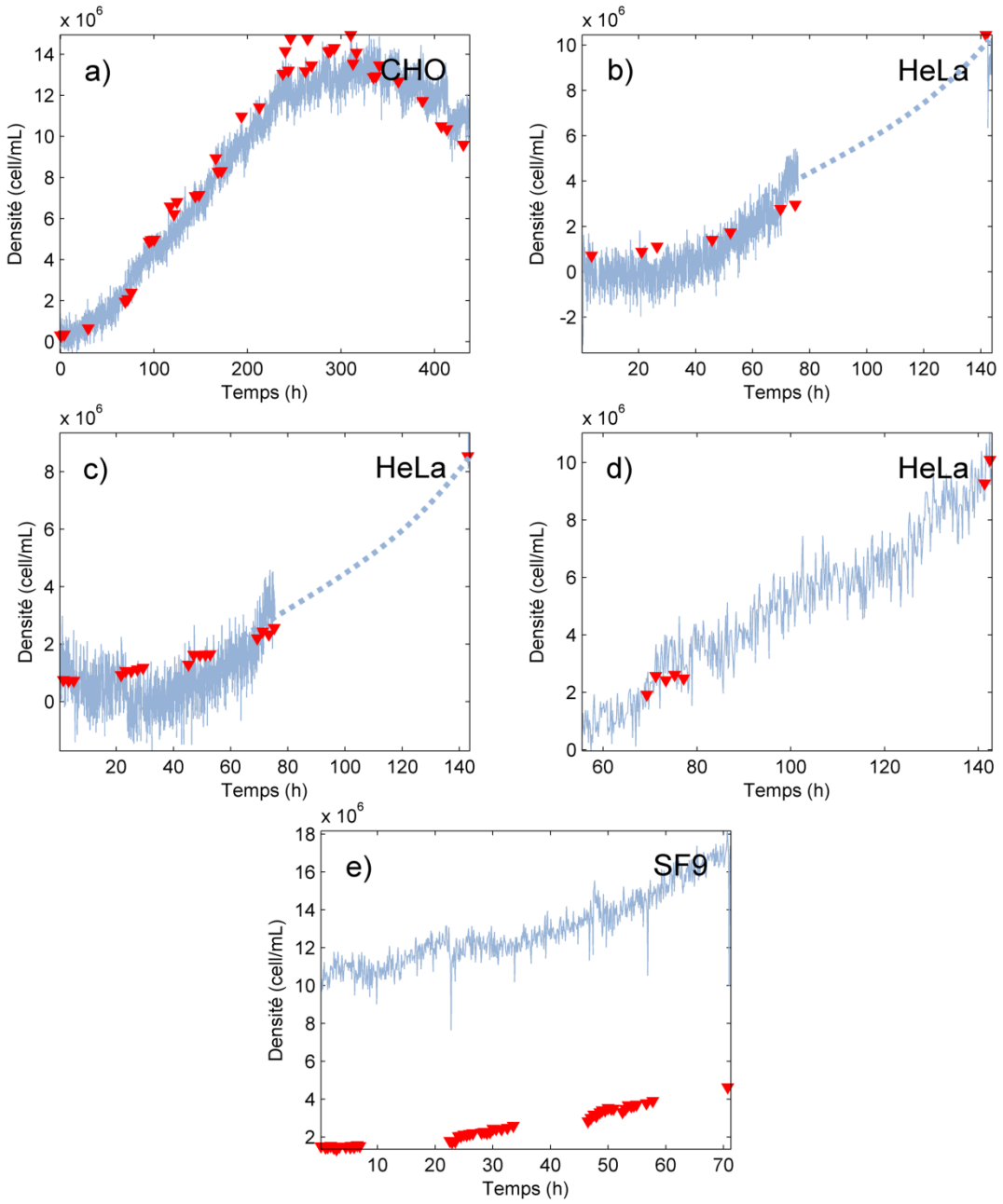

Figure 4.50 Représentations des prédictions réalisées par le modèle de régression densités TCD (durant la phase de culture cellulaire) développé à partir des échantillons de cultures CHO, HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) CHO #1, b) HeLa #1, c) HeLa #2, d) HeLa #5 et e) Sf9 #2.

Si nous prenons en compte les performances obtenues par les modèles de régression présentées Tableau 4.15, nous pouvons voir que les erreurs RMSEP des modèles sont relativement faibles, compte tenu de la gamme de densité. Nous estimons l'erreur RMSEP à $0{,}88 \cdot 10^6$ cell/mL pour une gamme $0{,}68 \cdot 10^6$–$4{,}38 \cdot 10^6$ cell/mL pour la densité en cellules vivantes VCD et à $0{,}70 \cdot 10^6$ cell/mL pour une gamme $1{,}41 \cdot 10^6$–$5{,}39 \cdot 10^6$ cell/mL pour la densité totale en cellules TCD. Néanmoins, ces erreurs de prédiction restent relativement

élevées par rapport aux erreurs RMSECV ($0{,}39 \cdot 10^6$ cell/mL et $0{,}34 \cdot 10^6$ cell/mL respectivement). En fait, ceci provient en partie des gammes de densité de l'ensemble test employées qui sont supérieures à celles de calibration ($1{,}27 \cdot 10^6$–$5{,}99 \cdot 10^6$ cell/mL pour la densité VCD et $1{,}40 \cdot 10^6$–$6{,}92 \cdot 10^6$ cell/mL pour la densité TCD). Nous avons donc une extrapolation des densités les plus importantes, ce qui induit un écart des prédictions que nous pouvons observer sur les représentations graphiques des modèles, Figure 4.51.

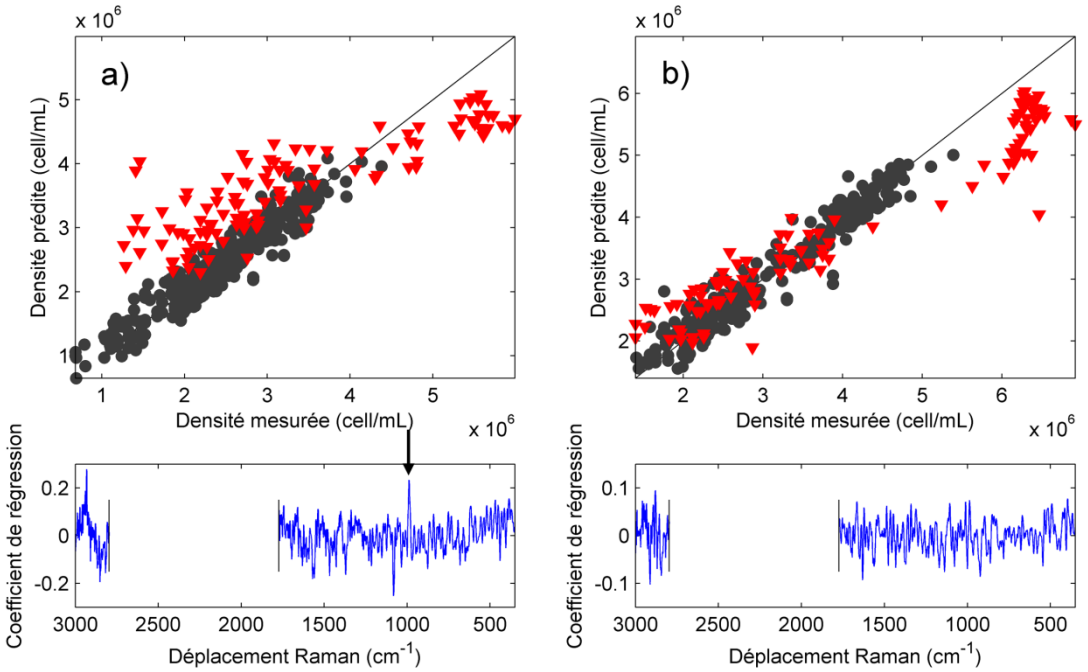

Figure 4.51 Représentations des résultats obtenus pour les modèles PLS basés sur les cultures de cellules CHO, HeLa et Sf9 pour a) la densité en cellules vivantes VCD et b) la densité en cellules totale TCD après introduction du virus dans le milieu. Les points noirs représentent les échantillons de calibration tandis que les triangles rouges représentent les échantillons test.

Si nous prenons les prédictions réalisées sur les spectres acquis en continu pour les cultures test employées ici, à savoir HeLa #1, #2 et #5, ainsi que Sf9 #4, Figure 4.52 et Figure 4.53, nous pouvons observer les différentes erreurs de prédiction observées sur les graphiques des densités prédites en fonction des densités mesurées. Toutefois, il apparait que les modèles de régression sont capables de capter les évolutions des différentes densités cellulaires.

En ce qui concerne les cellules de type HeLa, Figure 4.52a-c pour les prédictions de la densité en cellules vivantes et Figure 4.53a-c pour la densité totale en cellules, nous pouvons noter que les deux paramètres présentent des profils similaires, à savoir des croissances tout au long du bioprocédé. Nous pouvons voir que les modèles PLS sont en mesure de prédire cette tendance dans les deux cas.

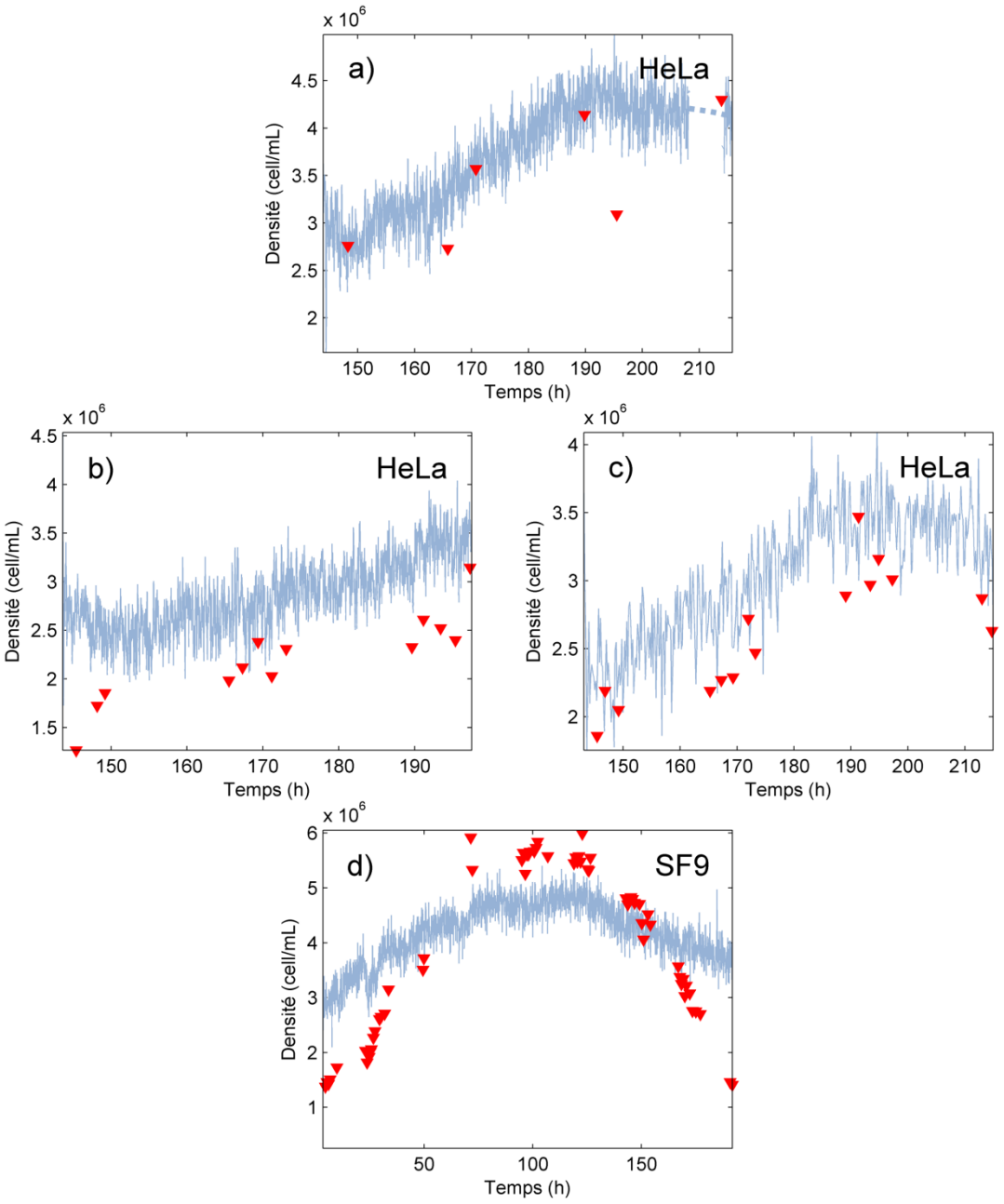

Figure 4.52 Représentations des prédictions réalisées par le modèle de régression densités VCD (durant la phase de culture virale) développé à partir des échantillons de cultures HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) HeLa #1, b) HeLa #2, c) HeLa #5 et d) Sf9 #4.

En ce qui concerne les prédictions en continu pour les paramètres VCD et TCD pour les cellules Sf9, nous pouvons d'abord remarquer que les deux biomarqueurs présentent des profils différents. La densité en cellules vivantes montre bien la croissance cellulaire jusqu'à atteindre le palier seuil avant la phase de sénescence traduisant la mort des cellules, Figure 4.52d, tandis que la densité totale en cellules présente une évolution croissante en fonction de la viabilité du milieu de culture, soit une forte croissance au départ qui ralentit peu à peu jusqu'à atteindre un certain niveau maximal en fin de parcours, Figure 4.53d. Nous pouvons noter sur ces figures que, bien qu'il y ait des écarts dans les prédictions, les modèles de

régression parviennent à capter les différentes évolutions. Ceci est notamment dû aux différences des deux modèles PLS ici développés. En effet, contrairement à ce que nous avons observé lors de la phase de culture cellulaire, les coefficients de régression des modèles PLS pour la phase de culture virale présentent une principale différence autour de 1000 cm^{-1}, traduisant la présence de phénylalanine pour les cellules vivantes, qui permet de différencier les deux modèles de régression. Nous sommes donc en mesure de prédire de façon distincte la densité en cellules vivantes de la densité totale en cellules.

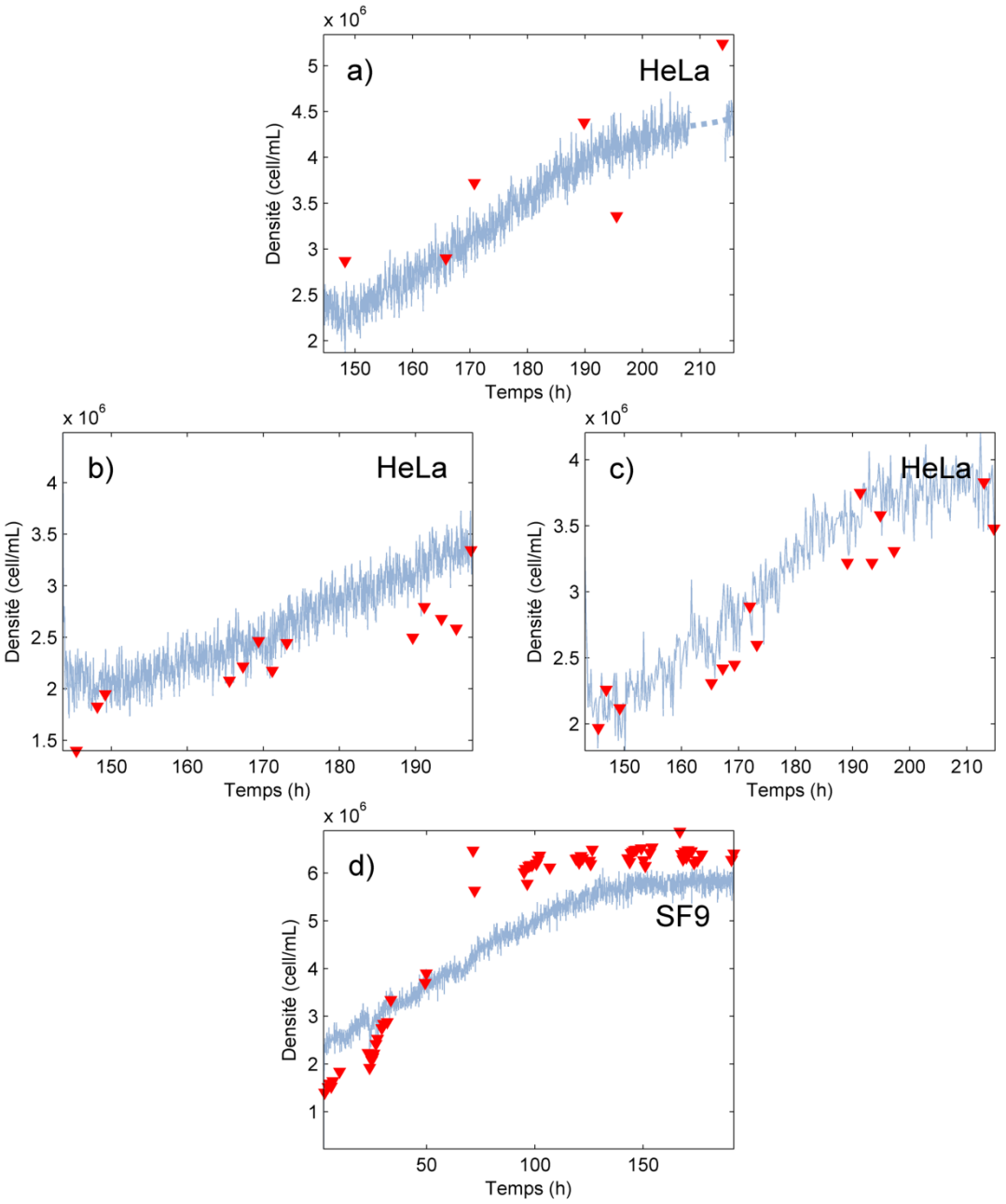

Figure 4.53 Représentations des prédictions réalisées par le modèle de régression densités TCD (durant la phase de culture virale) développé à partir des échantillons de cultures HeLa et Sf9 et appliqué sur les spectres acquis en continu pour les lots test a) HeLa #1, b) HeLa #2, c) HeLa #5 et d) Sf9 #4.

Si nous considérons tous les résultats obtenus, soit les modèles de régression développés pour les concentrations des différents métabolites et les densités cellulaires, nous pouvons souligner plusieurs points. Tout d'abord, nous avons pu montrer que nous étions en mesure de mutualiser l'information détenue par les échantillons acquis sur des cultures de différents types de cellule afin de développer des modèles de régression permettant de prédire les teneurs des métabolites glucose et lactate de n'importe quelle lignée engagée.

En revanche, nous avons également montré qu'il était difficile de construire des modèles de régression capables de parvenir aux mêmes résultats pour les concentrations en glutamine et glutamate. En effet, nous disposons de gammes de concentrations trop hétérogènes. Les modèles PLS ne se reposent donc que sur les échantillons des cultures dont l'évolution du paramètre en question est la plus forte, ce qui provoque d'importantes erreurs de prédiction pour les échantillons des autres types de cellule.

Enfin, nous avons obtenu des résultats mitigés en ce qui concerne les densités cellulaires. En effet, d'une part, les modèles de régression développés avant l'inoculation virale présentent des écarts importants pour prédire les densités en cellules Sf9, contrairement aux densités en cellules HeLa et CHO. D'autre part, les résultats obtenus sur les phases de culture virale sont encourageants puisque nous sommes en mesure de différencier les densités VCD et TCD, notamment grâce aux références disponibles pour les densités cellulaires des cultures Sf9. Toutefois, ces-derniers présentent encore trop peu de robustesse.

3.2 Validation sur un nouveau type de cellule : HEK

Afin de valider les modèles de régression mettant en jeu les trois types de cellule (CHO, HeLa et Sf9), nous introduisons ici un quatrième type de cellule, en l'occurrence les cellules HEK (présentées lors de la section 3.3.4 du Chapitre 1, Cellules , page 34). Les échantillons provenant des cultures de ces cellules n'apparaissant pas au sein des jeux de calibration, il nous est donc possible de les employer afin de voir si les derniers modèles PLS peuvent être exportés à d'autres types de cellule.

Au cours de ces travaux, les cellules HEK sont employées afin de cultiver du virus. Ainsi, les cultures de cellules HEK présentent le même profil que les cultures de cellules HeLa ou Sf9 avec une phase de multiplication cellulaire et une phase de culture virale. Mais tout comme les bioprocédés mettant en jeu les cellules HeLa, les deux étapes sont réalisées directement à la suite. Nous disposons ici de trois lots différents acquis sur 263 h et comprenant chacun 37 échantillons ainsi que près de 910 spectres Raman acquis en

continu. Pour le premier lot, aucune infection n'est réalisée tandis que pour les deux suivants, l'infection est réalisée à 165 h de culture. Le Tableau 4.16 présente un bref récapitulatif des données disponibles pour les cellules HEK.

Tableau 4.16 Récapitulatif du nombre d'échantillons pour chaque culture mettant en jeu les cellules HEK employée afin de tester les modèles de régression développés à partir des échantillons des cultures de trois types de cellule (CHO, HeLa, HEK), séparant les échantillons acquis avant et après la phase d'inoculation virale.

Numéro de lot	Nombre d'échantillons		
	Pré-*	Post-*	*(Total)*
#1	37	-	*(37)*
#2	16	21	*(37)*
#3	16	21	*(37)*

** Pré- et Post-inoculation virale*

Pour valider les modèles de régression construits précédemment (section 3.1), nous prenons l'ensemble des échantillons acquis pour les cultures de cellules HEK en jeu test, sauf pour les modèles des densités cellulaires, scindés en parties pré- et post-inoculation.

3.2.1 Prédiction des échantillons des cultures de cellules HEK

Les prédictions obtenues sur chaque biomarqueur pour les échantillons des cultures de cellules HEK sont représentées Figure 4.54 et permettent d'apprécier les performances des modèles PLS à prédire les teneurs de différents métabolites ou les densités cellulaires d'une nouvelle lignée cellulaire.

Tout d'abord, nous pouvons voir que les modèles élaborés pour prédire les concentrations en glucose et en lactate, Figure 4.54a-b, présentent des capacités prédictives satisfaisantes. En effet, bien que nous observions des biais dans les prédictions, les erreurs RMSEP, calculées à hauteur de 0,85 g/L et 0,94 g/L pour les modèles glucose et lactate respectivement, sont satisfaisantes compte tenu des gammes de concentration mises en jeu. En reprenant les spectres acquis en continu pour chacune des trois cultures de cellules HEK, nous pouvons obtenir les profils de concentrations prédites tout au long des bioprocédés. Ces prédictions, représentées Figure 4.55 et Figure 4.56, permettent de suivre de manière satisfaisante les tendances métaboliques du glucose et du lactate respectivement. Cependant, nous observons un écart des prédictions par rapport aux mesures de référence traduisant la présence de biais évalués à -0,65 g/L pour les concentrations en glucose et à -0,90 g/L pour les prédictions des concentrations en lactate. Ces phénomènes sont directement liés à l'absence d'échantillon de cultures de cellules HEK

dans l'ensemble de calibration, privant donc les modèles de régression de la variabilité spectrale qui existe entre ces cultures et celles de l'ensemble d'apprentissage.

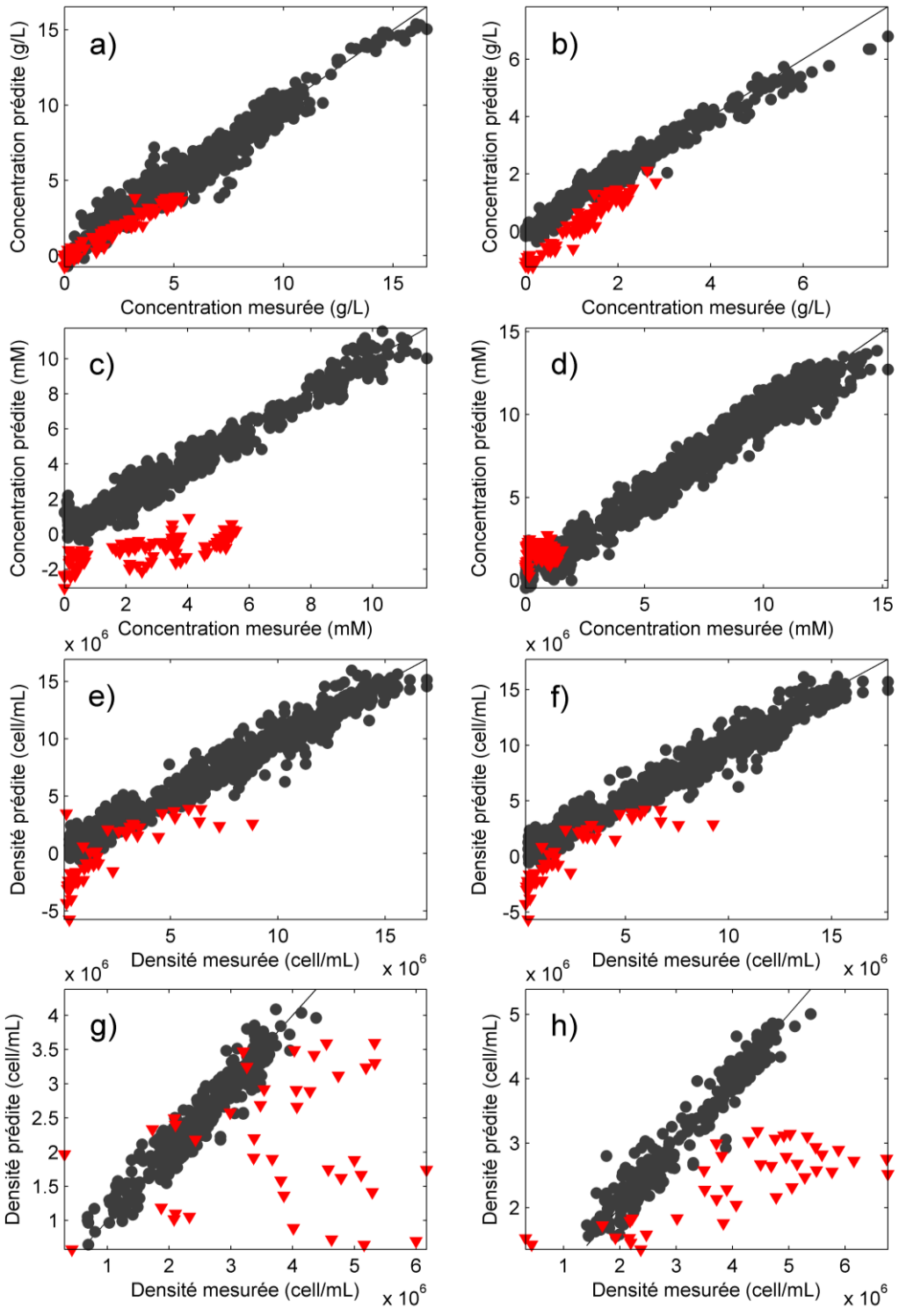

Figure 4.54 Représentations des prédictions réalisées sur les échantillons des cultures HEK (triangles rouges) à partir des échantillons acquis sur les cultures de cellules CHO, HeLa et Sf9 (points noirs) pour a) la concentration en glucose, b) la concentration en glutamine, c) la concentration en glutamate, d) la concentration en lactate, e) la densité VCD pré-inoculation, f) la densité TCD pré-inoculation, g) la densité VCD post-inoculation et h) la densité TCD post-inoculation.

Ensuite, nous pouvons observer les prédictions réalisées par les modèles de régression développés pour les concentrations en glutamine et glutamate, Figure 4.54c-d. Lors de la conception de ces modèles de régression, nous avions montré que les différences de gammes de concentration entrainaient des problèmes pour la prédiction de toutes les lignées cellulaires. En effet, le modèle pour les concentrations en glutamine reposait essentiellement sur les échantillons des cultures de cellules Sf9 tandis que le modèle pour les concentrations en glutamate s'appuyait en majorité sur les échantillons des cultures de cellules CHO. Nous pouvons voir ici que les prédictions réalisées pour les teneurs en glutamine, Figure 4.54c, présente des écarts évidents ce qui signifie que le modèle glutamine est bel et bien inapproprié pour les prédictions de spectres acquis à partir de cultures de cellules différentes des cellules Sf9. Pour les concentrations en glutamate, nous pouvons voir que les gammes de concentration des cultures de cellules HEK présentent de faibles variations (0,00–1,54 mM). Ainsi, nous nous retrouvons dans le même cas que les autres types de cellules lors du développement du modèle, présentant donc trop peu de variations sur les bandes spectrales caractéristiques de la molécule de glutamate pour en ressortir les teneurs de manière satisfaisante.

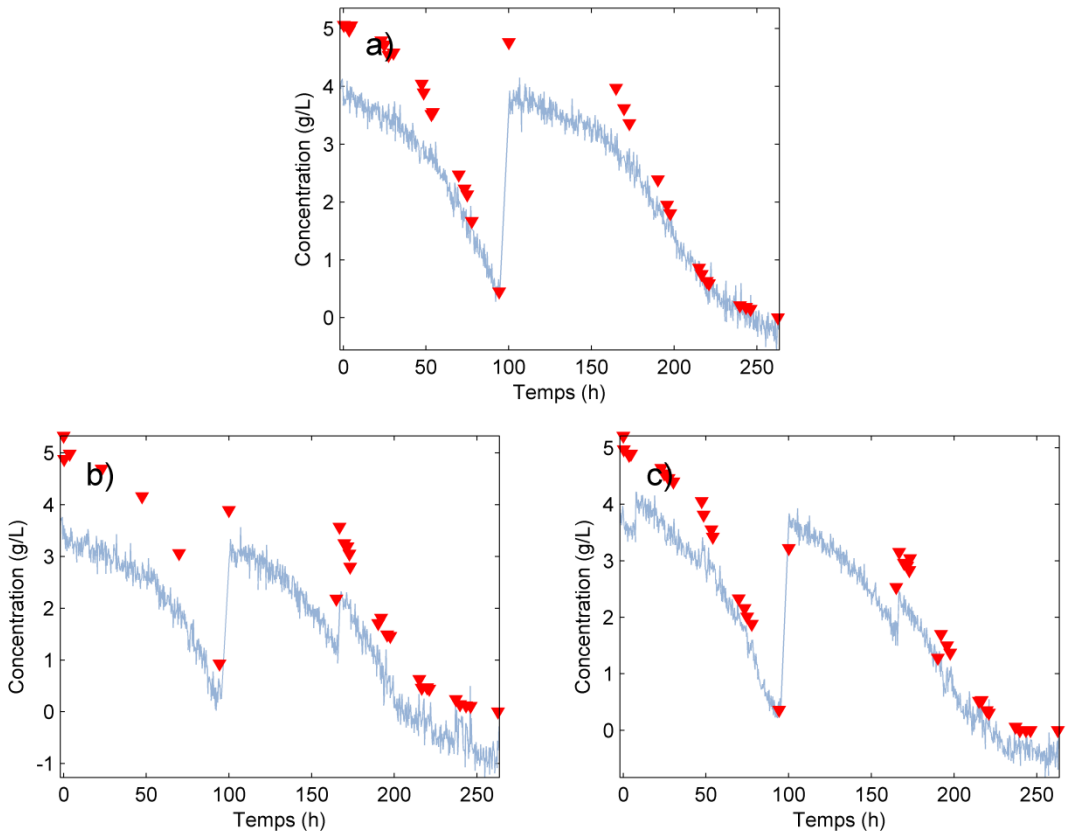

Figure 4.55 Représentations des prédictions de la concentration en glucose réalisées sur les spectres acquis en continu pour les cultures mettant en jeu les cellules HEK a) #1 (culture de cellules uniquement), b) #2 (culture virale, infection à 165 h) et c) #3 (culture virale, infection à 165 h).

Enfin, nous pouvons voir que les modèles de régression développés pour les densités VCD et TCD exposent des prédictions présentant des biais importants pour la phase pré-infection, Figure 4.54e-f. En effet, nous pouvons voir que malgré la présence de quelques échantillons bien prédits en milieu de culture, le modèle PLS ne parvient pas à capter convenablement l'évolution des niveaux cellulaires. En représentant l'évolution de la densité en cellules vivantes VCD pour la culture HEK #1, Figure 4.57, nous pouvons clairement voir les écarts qui existent entre les prédictions et les mesures de référence. À noter également que les observations sont les mêmes pour la densité TCD sur la phase pré-infection.

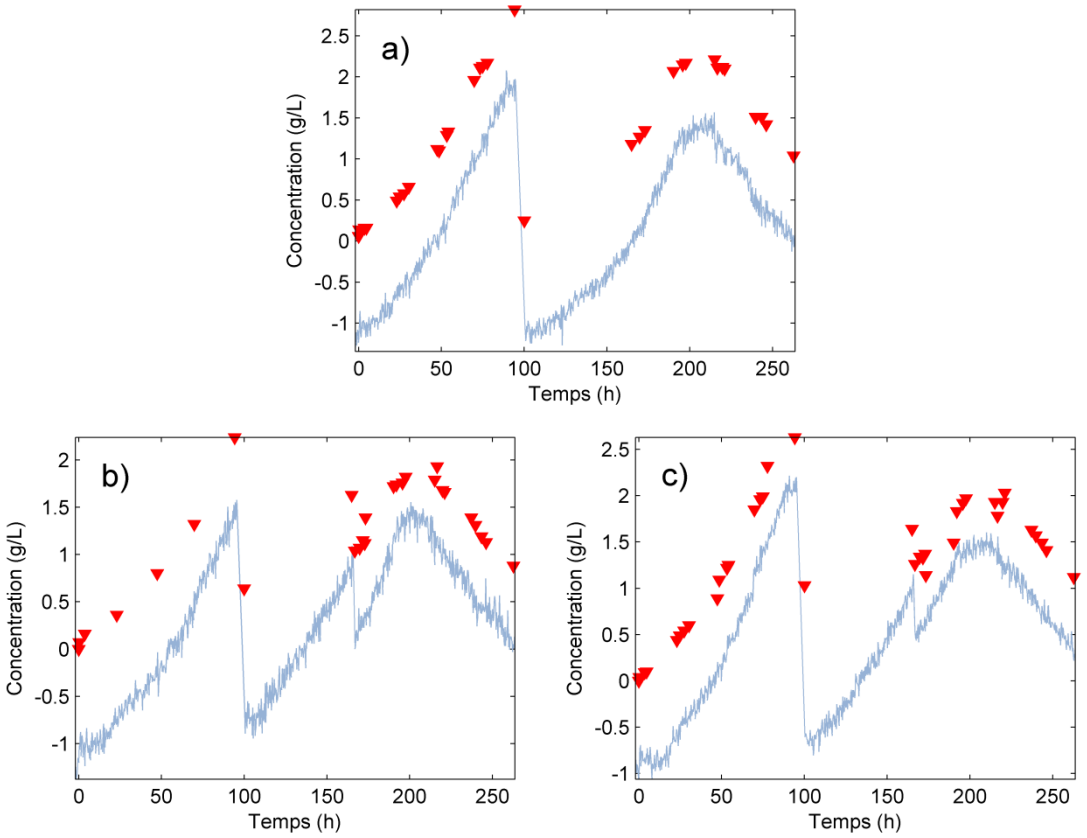

Figure 4.56 Représentations des prédictions de la concentration en lactate réalisées sur les spectres acquis en continu pour les cultures mettant en jeu les cellules HEK a) #1 (culture de cellules uniquement), b) #2 (culture virale, infection à 165 h) et c) #3 (culture virale, infection à 165 h).

Quant aux prédictions de densités VCD et TCD sur les échantillons des cultures virales mettant en jeu les cellules HEK (Figure 4.54g-h), nous pouvons très clairement voir que le modèle n'est pas du tout en mesure de prédire les densités cellulaires. Nous pouvons donc constater qu'il est très difficile d'élaborer un modèle de régression robuste aux changements de type de cellule. En effet, le matériel biologique étant propre à chaque cellule, les modèles de régression ne parviennent pas à trouver des bandes spectrales significatives sur lesquelles s'appuyer afin d'être en mesure de prédire les densités cellulaires de chaque lignée.

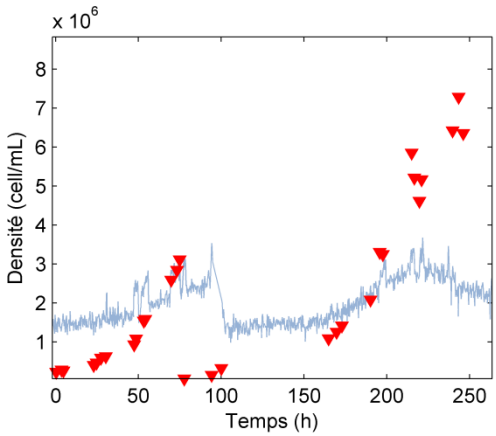

Figure 4.57 Représentations des prédictions des densités en cellules vivantes VCD pour les échantillons de la culture en cellules HEK #1.

Ainsi, il apparait finalement que seuls les modèles de régression pour les concentrations en glucose et en lactate proposent des niveaux d'erreurs satisfaisants, bien que nous observions un écart sur les prédictions des échantillons des cultures de cellules HEK. Néanmoins, ces biais s'expliquent simplement par l'absence d'échantillon provenant des cultures de cellules HEK dans l'ensemble d'apprentissage.

3.2.2 Intégration d'échantillons de cultures HEK aux modèles PLS

Nous ajoutons les échantillons de la culture de cellules HEK #2 à l'ensemble de calibration afin d'intégrer la variabilité qui existe entre les différentes cultures à l'ensemble d'apprentissage, tout en utilisant les échantillons des cultures de cellules HEK #1 et #3 en tant que lot test. Nous prenons la culture HEK #2 (tout comme nous aurions pu prendre la culture HEK #3) puisque celle-ci présente l'étape d'infection virale, contrairement à la culture HEK #1. Les résultats obtenus pour les modèles PLS des concentrations en glucose et lactate sont disponibles Figure 4.58.

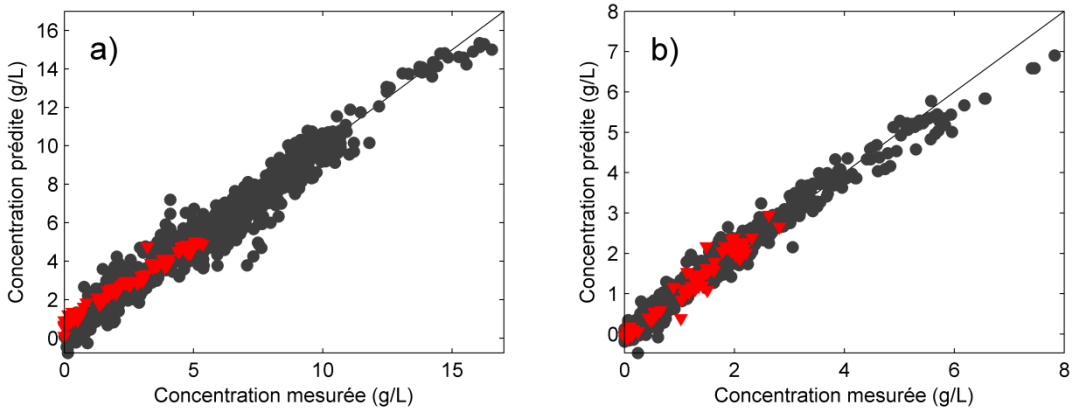

Figure 4.58 Représentations des prédictions réalisées sur les échantillons des cultures HEK #1 et #3 (triangles rouges) à partir des échantillons acquis sur les cultures de cellules CHO, HeLa et Sf9 (points noirs) auxquelles sont ajoutés les échantillons de la culture de cellules HEK #2 pour a) la concentration en glucose, b) la concentration en lactate.

L'erreur RMSEP du modèle de régression pour les concentrations en glucose est estimée à 0,53 g/L, valeur inférieure à ce que nous obtenions précédemment (0,85 g/L). Bien que nous ayons une culture de moins dans l'ensemble test, ce qui pourrait expliquer en partie l'amélioration des performances prédictives, nous pouvons observer une diminution du biais de prédiction, évoluant ainsi de -0,65 g/L à 0,34 g/L. Nous pouvons faire les mêmes remarques pour le modèle de régression pour les concentrations en lactate, présentant une erreur RMSEP de 0,20 g/L (contre 0,94 g/L précédemment) et un biais de prédiction estimé à 0,01 g/L (contre -0,90 g/L avant). En prédisant les concentrations en glucose et en lactate des spectres acquis en continu pour les cultures de cellules HEK #1 et #3, Figure 4.59, nous pouvons bel et bien voir que les biais de prédiction observés précédemment (Figure 4.55 et Figure 4.56) sont grandement atténués.

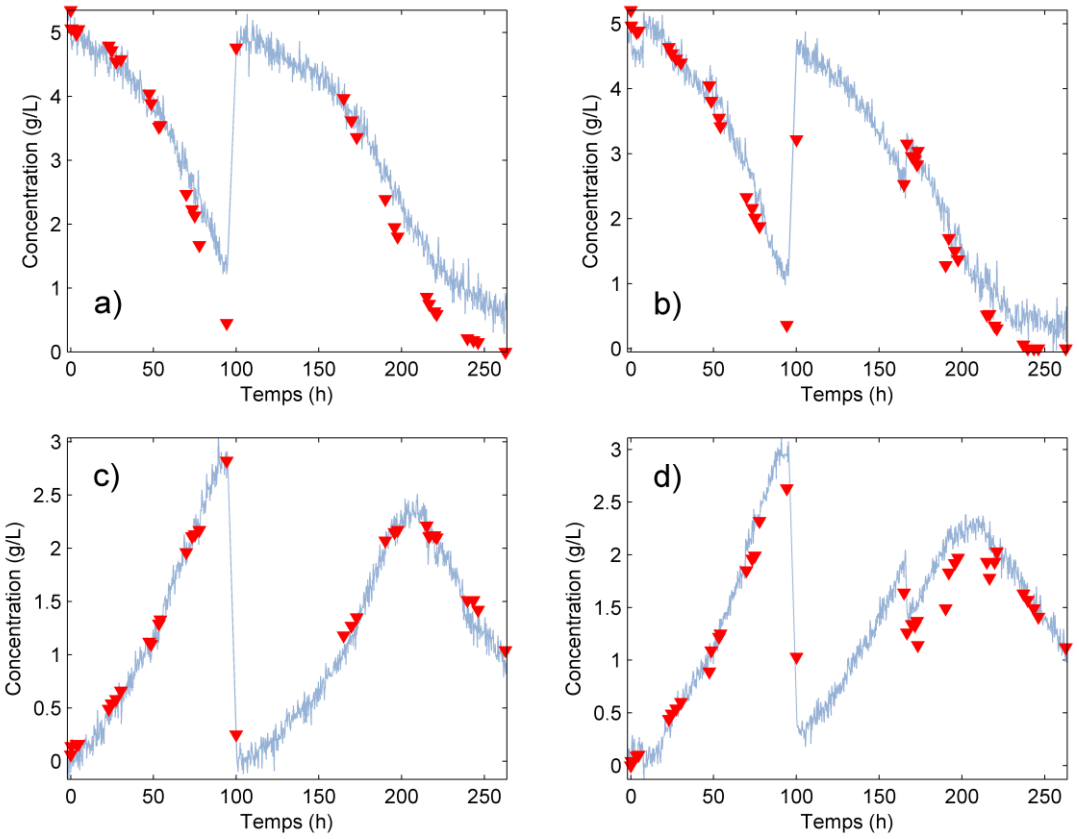

Figure 4.59 Représentations des prédictions réalisées en intégrant les échantillons de la culture de cellules HEK #2 à l'ensemble d'apprentissage pour les concentrations a) en glucose du lot HEK #1, b) en glucose du lot HEK #3, c) en lactate du lot HEK #1 et d) en lactate du lot HEK #3.

Nous pouvons ainsi conclure que les biais observés précédemment étaient en grande partie dus à l'absence d'échantillon des cultures de cellules HEK au sein de l'ensemble de calibration. En ajoutant les données provenant d'une seule culture, nous avons ainsi pu réduire de façon importante les écarts exposés et donc accroitre la robustesse générale des modèles de régression développés pour les prédictions des métabolites glucose et lactate pour les cultures de différents types de cellule.

4 Conclusions sur l'optimisation et la recherche de robustesse pour les modèles de régression

L'amélioration continue des modèles de régression est au centre de ce chapitre. En effet, nous avons tout d'abord commencé par développer des modèles de régression capables de prédire les concentrations des différents métabolites ainsi que les densités cellulaires pour les cultures des cellules CHO, HeLa et Sf9 séparément. Ainsi, en intégrant dans un premier temps les variabilités qui existent entre les différentes cultures d'un même type de cellule, nous avons montré que nous pouvions exporter les modèles à d'autres cultures, absentes de l'ensemble d'apprentissage.

Néanmoins, lors des cultures de cellules, les paramètres physiques peuvent varier selon les conditions du milieu. Ainsi, en étudiant l'influence de plusieurs paramètres critiques, nous avons montré que les variations des conditions de culture influent directement sur les cellules, ce qui perturbe la multiplication cellulaire et donc les gammes de concentration des métabolites mis en jeu (sans pour autant dénaturer la mesure Raman). De ce fait, en réalisant des cultures cellulaires dans des conditions atypiques de culture, nous obtenons des profils de concentration balayant des gammes plus importantes, notamment pour le glucose et le lactate. L'intégration de ces cultures atypiques permet en fin de compte d'accroitre la robustesse des modèles de régression et d'étendre les gammes de calibration.

Nous avons ensuite élaboré des modèles de régression qui s'appuient sur les échantillons de trois types de cellule : CHO, HeLa et Sf9. Ainsi, nous avons eu pour but de créer des modèles de régression étant capables de surmonter les caractéristiques biologiques qui existent entre les différentes lignées cellulaires pour prédire les niveaux des biomarqueurs d'intérêt. Cependant, seuls les modèles développés pour les concentrations en glucose et en lactate présentent de réels potentiels prédictifs. En effet, d'un côté, les gammes de concentration en glutamine et glutamate présentent trop d'écarts d'une cellule à l'autre ce qui empêche de construire des modèles PLS fiables. D'un autre côté, les différences biologiques entre les cellules sont telles que l'élaboration de modèles de régression pour la détermination des densités cellulaires est difficile.

Enfin, nous avons pu montrer que les modèles de régression basés sur les concentrations en glucose et en lactate permettaient de déterminer de manière satisfaisante les différentes teneurs des échantillons d'un quatrième type de cellule (HEK) inconnu au modèle, malgré l'existence d'un certains biais de prédiction. Cependant, ces-derniers peuvent être largement atténués en intégrant plusieurs échantillons de culture du nouveau type de cellule au sein de l'ensemble d'apprentissage, ce qui laisse entrevoir la possibilité d'étendre les modèles de régression pour les suivis d'autres types de cellule encore.

Chapitre 5
Vers de nouvelles valorisations des données spectroscopiques

Durant les cultures de cellules, plusieurs paramètres sont contrôlés et régulés en temps réel, tels que les paramètres physiques de culture, tandis que d'autres sont mesurés par des méthodes de référence sur la base de prélèvements, ce qui est notamment le cas des concentrations métaboliques ou encore des densités cellulaires. Néanmoins, les prélèvements réalisés permettent d'accéder à d'autres informations plus difficiles à obtenir telles que les concentrations en anticorps lors des cultures de cellules, ou encore les titres en virus infectieux pour les cultures virales.

De plus, l'acquisition de spectres Raman d'une culture de cellules en continu permet, au-delà de la prédiction des concentrations métaboliques ou des densités cellulaires sur toute la durée du bioprocédé, de bénéficier de représentations spectroscopiques du contenu du bioréacteur de culture sur l'ensemble du procédé. Ainsi, nous pouvons accéder aux changements chimiques qui peuvent survenir durant la culture sans prendre de mesure de référence.

L'objectif de ce chapitre est donc de valoriser les données à disposition afin de pouvoir développer des outils permettant d'accéder à de plus amples informations sur les cultures cellulaires et virales engagées.

1 Prédiction de la concentration en anticorps

Lors des cultures cellulaires, nous avons pu montrer que de nombreux paramètres doivent être pris en considération pour que le développement des cellules soit suffisant et que la production d'anticorps soit optimale. Il convient donc de mesurer la concentration en anticorps pendant le bioprocédé afin non seulement d'évaluer les performances de la culture cellulaire, mais aussi de déterminer, en cas de problème, à quel moment la production a été impactée. Toutefois, les méthodes de référence ne permettent pas d'accéder aux concentrations en anticorps recombinants de manière instantanée. De ce fait, les mesures réalisées rétrospectivement excluent le développement d'un suivi en temps réel ou n'importe quelle action corrective sur la culture en cas de détection de problème depuis la concentration en anticorps. C'est pourquoi aujourd'hui, les méthodes de mesure permettant

d'accéder rapidement à ces teneurs peuvent parfois être onéreuses, reposant principalement sur des techniques chromatographiques, sans pour autant permettre d'accéder au titre en direct et en temps réel et en ligne.

L'intérêt de la spectroscopie vibrationnelle, couplée aux principes de régression chimiométrique entre donc en jeu afin de développer de nouvelles techniques de mesure du taux d'anticorps présent en culture. Ainsi en 2010, suite à certains travaux sur la caractérisation de protéines, la spectrométrie infrarouge à transformée de Fourier a été associée à la méthode de régression PLS afin de proposer une technique rapide de dosage des anticorps pour les cultures de cellules CHO [153]. Plusieurs prélèvements ont été réalisés afin de mesurer la concentration en anticorps dans le milieu à l'aide de la technique ELISA. Ensuite, une partie du prélèvement (entre 0–20 µL) est analysée à l'aide d'un spectromètre infrarouge à transformée de Fourier. Les spectres et mesures de référence sont finalement reliés à l'aide de la méthode PLS afin de développer les modèles de régression. Ainsi, il s'agit là d'une première approche de caractérisation de la teneur en anticorps qui met en avant les bénéfices des techniques spectroscopiques, notamment la non-destructivité des échantillons mesurés et la rapidité d'analyse. En 2013, des travaux similaires sont conduits et mettent en jeu la spectrométrie Raman [154]. Les méthodes HPLC et ELISA étaient toutes deux mises en jeu afin de prédire les concentrations de référence en anticorps recombinants sur des prélèvements réalisés à partir de cultures de cellules CHO. De plus, 40 µL de chaque prélèvement était exposé au laser UV émettant à 244 nm afin d'acquérir les spectres Raman représentatifs des échantillons. Une nouvelle fois, la méthode PLS était appliquée afin de relier les spectres aux concentrations de référence pour développer le modèle de régression permettant de prédire la teneur en anticorps.

Ainsi, ces travaux permirent de mettre en avant les dispositions des techniques spectroscopiques et chimiométriques à prédire la concentration en anticorps pour une culture cellulaire, exposant ainsi les capacités à conduire le suivi de ce paramètre d'intérêt. Toutefois, bien que les techniques spectroscopiques présentées jusque là soient plus rapides que les techniques chromatographiques ou enzymatiques, la nécessité de réaliser un prélèvement prive le suivi du caractère temps réel et en ligne de la méthode analytique. C'est pourquoi nous avons employé les spectres Raman acquis à l'aide des sondes à immersion, utilisées jusque là pour le suivi spectroscopique des différents biomarqueurs, pour mettre en œuvre la modélisation chimiométrique de la concentration en anticorps. Ainsi, nous nous appuyons non pas sur des spectres de prélèvements, mais sur des spectres acquis *in situ*. Ces travaux font l'objet d'un article scientifique publié en 2015 dans la revue *Analytica Chimica Acta* [155]. Nous présentions alors un modèle de régression PLS basé sur

132 échantillons provenant de trois cultures de cellules CHO différentes, dont les spectres étaient acquis à l'aide d'une sonde à immersion plongée dans le milieu de culture et reliée au spectromètre Raman Rxn2 (Kaiser Optical Systems Inc.). Les mesures de référence associées pour la concentration en anticorps étaient déterminées à partir de prélèvements de culture passés en UPLC (ACQUITY® H-Class Bio UPLC, Waters Corporation).

Afin de développer le modèle de régression PLS, nous avons scindé l'ensemble des 132 échantillons provenant de trois cultures différentes en un jeu de calibration de 99 échantillons et un jeu test de 33 échantillons. Le fait de ne disposer que de trois cultures nous privait d'un ensemble test constitué des données d'une seule culture extérieure à l'ensemble de calibration. En effet, si nous avions étalonné le modèle de régression sur la base des données provenant de deux cultures, la troisième étant conservée dans l'ensemble test, le manque de robustesse aurait certainement entrainé d'importants biais de prédiction. Nous avons donc panaché les échantillons des trois cultures mises en jeu au sein des ensembles d'apprentissage et test. Nous avions alors obtenu un modèle de régression satisfaisant présentant une erreur RMSEP évaluée à 0,054 mg/mL pour une gamme de concentration 0,034–1,290 mg/mL et une erreur RMSECV calculée de 0,037 mg/mL. De plus, les prédictions réalisées sur les spectres acquis en continu pour les trois cultures de cellules CHO mises en jeu, Figure 5.1, permettaient de montrer que nous étions capables de suivre l'évolution de la concentration en anticorps pendant les cultures de cellules.

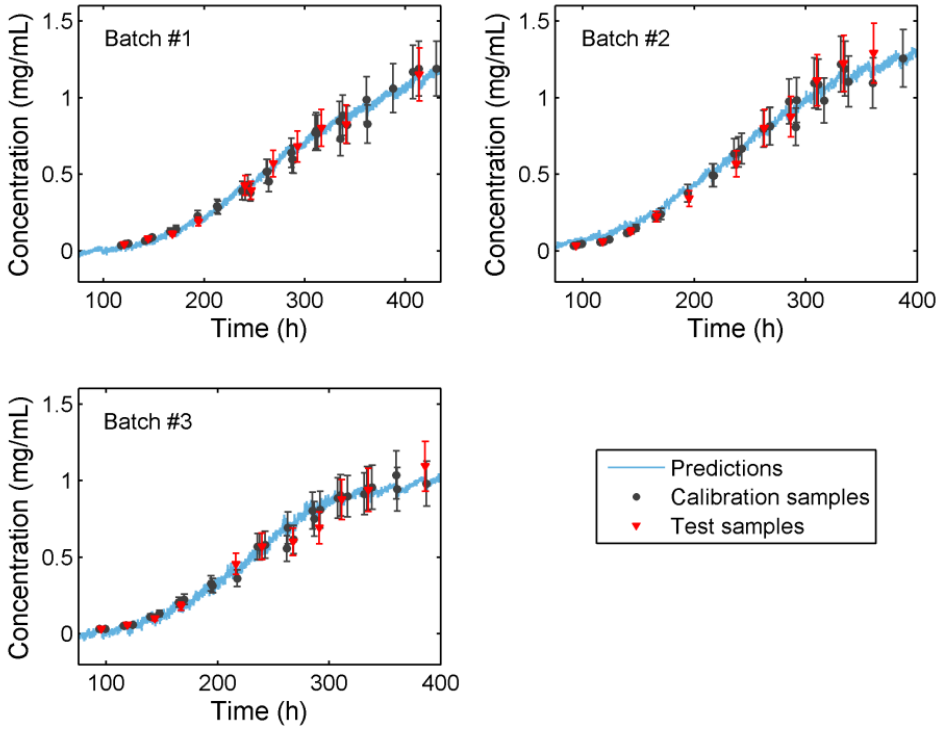

Figure 5.1 Représentations des prédictions de la concentration en anticorps réalisées sur les spectres acquis en continu pour les trois cultures de cellules CHO utilisées pour les premiers travaux sur le développement d'un modèle de régression PLS pour le suivi en ligne et en temps réel du titre en anticorps recombinants. Les mesures de référence (points noirs et triangles rouges) sont accompagnées des 15 % de précision de la méthode de référence. Figure issue de la référence [155].

Nous avons poursuivi ces travaux en prenant en compte les quinze cultures de cellules CHO dont nous disposons. Le nombre de mesures de référence par culture pour la concentration en anticorps est présenté Tableau 5.1. Afin de bénéficier d'un ensemble d'apprentissage suffisamment représentatif de l'ensemble de la population et permettant d'éprouver le modèle de régression, trois cultures sont conservées dans l'ensemble test (et non plus une seule). Ainsi, l'ensemble de calibration est finalement constitué de 439 échantillons provenant de douze cultures différentes contre 118 échantillons dans l'ensemble test appartenant à trois différentes cultures.

Les spectres Raman sont prétraités de la même façon que lors des travaux sur les modélisations des différents biomarqueurs, soit une dérivée première de Savitzky-Golay (fenêtre d'ajustement de 15 points et polynôme de degré 2), suivi d'une normalisation SNV. Puis nous sélectionnons les régions spectrales comprises entre 3000–2800 cm^{-1} et 1775–350 cm^{-1}. Ensuite, tout comme le développement des modèles présentés dans les chapitres précédents, une validation croisée (ici *10-fold cross-validation*) est réalisée initialement.

Tableau 5.1 Récapitulatif du nombre de mesures de référence réalisées pour chacune des quinze cultures de cellules CHO. Les lignes grisées représentent les cultures dont les échantillons appartiennent à l'ensemble test.

Numéro de lot	Nombre d'échantillons
#1	45
#2	15
#3	44
#4	43
#5	50
#6	52
#7	36
#8	36
#9	39
#10	37
#11	37
#12	21
#13	27
#14	35
#15	40

Le modèle PLS développé sur la base des 439 échantillons de calibration est basé sur huit variables latentes et présenté Figure 5.2. L'erreur RMSECV est estimée à 0,068 g/L pour une gamme d'étalonnage comprise entre 0,024 mg/mL et 1,432 mg/mL. En appliquant

ce modèle de régression sur l'ensemble test constitué de 118 échantillons, nous obtenons finalement une erreur RMSEP calculée à 0,072 mg/mL.

Cette erreur est plus importante que ce que nous avons rencontré en ne travaillant que sur trois cultures. Ceci est tout-à-fait normal puisqu'ici, les échantillons provenant de l'ensemble test n'appartiennent pas aux mêmes cultures que ceux employés pour construire le modèle PLS. Ainsi, la variance induite par la diversité des cultures accroit les différentes erreurs, non seulement l'erreur RMSEP, mais aussi l'erreur RMSECV. En considérant les coefficients de régression, Figure 5.2, nous pouvons remarquer que, bien que le profil soit relativement bruité, ces derniers présentent plusieurs signaux importants. Nous retrouvons d'abord la bande située autour de 2900 cm^{-1} qui peut une nouvelle fois être attribuée aux élongations v(C–H) des différentes fonctions chimiques [146]. Ensuite, nous pouvons noter la présence de signaux autour de 1630 cm^{-1}, caractérisant principalement les molécules de phénylalanine, tryptophane et tyrosine [156-159], ainsi que les fonctions amide [160]. Le signal présent à 1350 cm^{-1} traduit également la présence de tyrosine [154,159], mais aussi d'adénosine et de thymidine [154,157-159]. De plus, nous pouvons assigner les signaux autour de 1120 cm^{-1} et 1000 cm^{-1} aux molécules de phénylalanine et tyrosine [154,158-161], ainsi qu'au squelette de la chaine ADN pour la bande à 1120 cm^{-1} [161]. Enfin, la bande autour de 720 cm^{-1} traduit les bandes caractéristiques de la molécule de tryptophane [157] tandis que les bandes comprises entre 520–380 cm^{-1} sont principalement dues aux contributions spectroscopiques des élongations v(S–S) des ponts disulfure permettant de stabiliser la structure des anticorps [159,162].

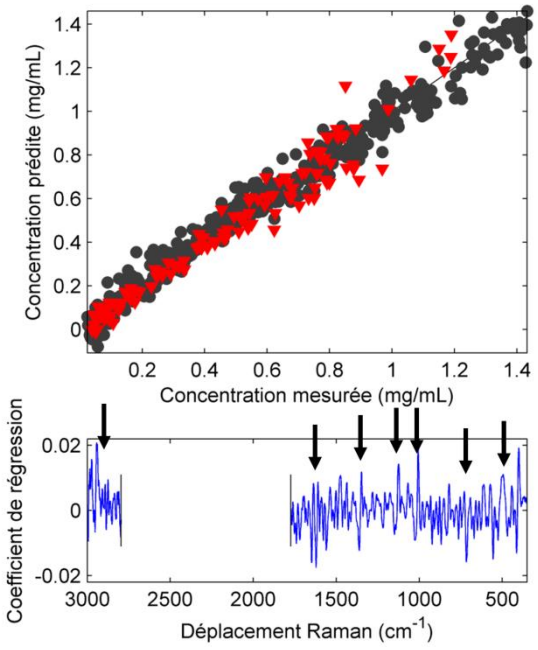

Figure 5.2 Représentation du modèle de régression développé pour la détermination de la concentration en anticorps recombinants pour les cultures de cellules CHO. Les points noirs représentent les échantillons de l'ensemble de calibration, les triangles rouges les échantillons de l'ensemble test.

Ces différentes attributions permettent de valider le modèle de régression en plus des faibles erreurs de prédiction. De ce fait, nous reprenons les spectres acquis en continu pour les trois cultures test, auxquels nous appliquons le modèle de régression PLS. Les prédictions obtenues sont confrontées aux mesures de référence, Figure 5.3. Nous pouvons alors voir que le modèle est tout à fait capable de suivre le profil d'évolution de la concentration en anticorps au sein des cultures test, bien que nous observions quelques écarts en début de procédé pour le lot #7 (Figure 5.3b). Toutefois, les profils des concentrations prédites ne suivent pas d'évolution linéaire, mais bien une concentration nulle en début de culture suivie d'une augmentation croissante au cours de la culture. Nous pouvons également remarquer que le suivi réalisé à partir des spectres Raman semble mieux suivre l'évolution du titre en anticorps par rapport à la méthode de référence employée, notamment à travers certaines mesures dont nous ne pouvons expliquer les baisses de titre (Figure 5.3).

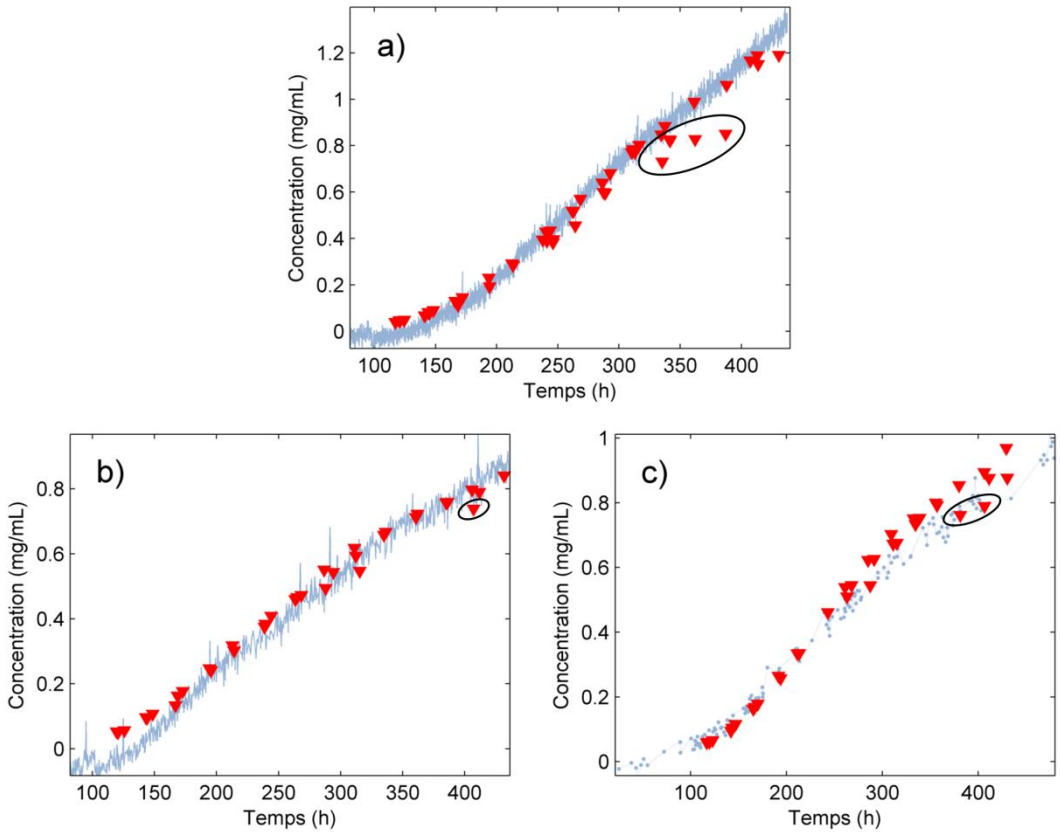

Figure 5.3 Représentations des prédictions réalisées à l'aide du modèle de régression pour les concentrations en anticorps sur les spectres acquis en continu de a) la culture #1, b) la culture #7 et c) la culture #11.

Il apparait finalement que le modèle de régression développé pour la détermination du titre en anticorps recombinants dans le milieu durant les cultures de cellules CHO permet de réaliser le suivi de la teneur au cours du temps. Ainsi, nous montrons qu'il est possible de développer une technique de mesure permettant d'accéder à la concentration en anticorps

en ligne sans solliciter de mesures à l'aide des méthodes de référence (chromatographiques et enzymatiques) ni être contraint de réaliser des prélèvements du bioréacteur durant la culture. Cette méthode prenant en compte des spectres Raman acquis *in situ*, couplés à un modèle PLS robuste peut donc permettre, à terme, de développer des outils de contrôle et d'optimisation en temps réel des productions d'anticorps durant les cultures cellulaires.

2 Prédiction du titre en virus infectieux

Après avoir appliqué les principes de régression multivariée pour la détermination de la concentration en anticorps lors de cultures de cellules CHO, prenant en compte des spectres Raman acquis *in situ*, nous étendons cette idée directement au titre en virus infectieux. En effet, le paramètre d'intérêt durant les cultures virales est la concentration en virus dans le milieu. Que ce soit pour les procédés mettant en jeu les cellules HeLa, Sf9 ou HEK, la phase de pré-inoculation virale permet essentiellement de produire un grand nombre de cellules potentiellement hôtes pour le virus injecté et ne présente donc pas de référence pour la concentration en virus.

2.1 Modélisation classique du titre en virus infectieux

Nous avons pris en compte les cultures de cellules Sf9 pour réaliser les modèles de régression pour le suivi de la teneur en virus au cours du temps. Cinq cultures virales ont été mises en jeu. Le nombre d'échantillons disponible pour chaque culture est présenté Tableau 5.2. Nous pouvons noter que les cultures virales #7 et #8 n'apparaissent pas lors des travaux précédents puisqu'elles ont été réalisées spécialement pour la modélisation de la concentration en virus.

Tableau 5.2 Récapitulatif du nombre de mesures de référence du titre en virus infectieux disponibles pour les cultures impliquant les cellules Sf9.

Numéro de lot	Nombre de références	Nombre de réf. (virus)
Amplification cellulaire		
#1-1	42	-
#2	50	-
Culture virale		
#3-1	42	42
#3-2	48	44
#4	63	58
#5	61	54
#7	37	29
#8	37	35

Les spectres de tous les lots de cultures virales ont ensuite été prétraités (dérivée première, normalisation SNV, sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}) et séparés en deux ensembles. Étant donné que nous disposons uniquement de cinq cultures virales, seuls les échantillons de la culture virale #5 composent l'ensemble test, le reste de la population représentant donc l'ensemble d'apprentissage du modèle de régression. Nous effectuons également une validation croisée préliminaire permettant de déterminer le nombre de variables latentes à intégrer au modèle, ainsi que l'erreur RMSECV. Le modèle de régression obtenu est présenté Figure 5.4.

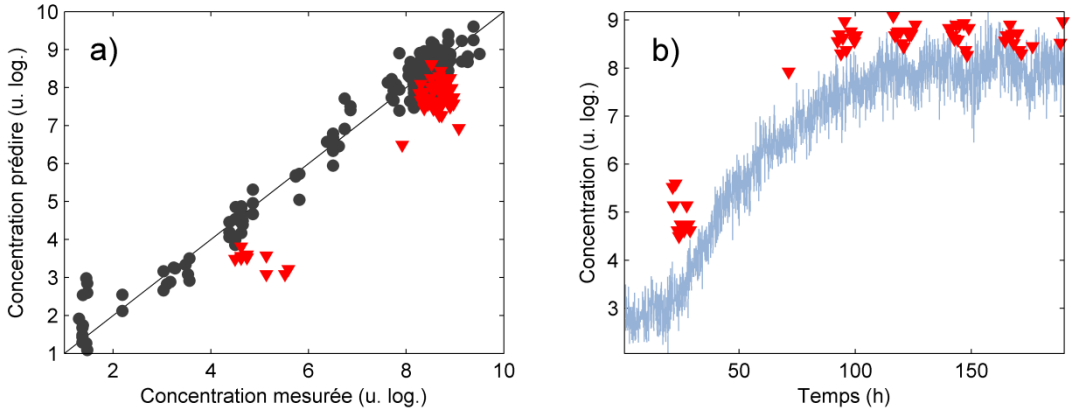

Figure 5.4 Représentation des performances prédictives du modèle de régression développé pour la détermination des concentrations en virus. En a), le graphique des concentrations prédites en fonction des concentrations mesurées pour les échantillons de calibration durant la validation croisée (points noirs) et pour les échantillons de l'ensemble test (triangles rouges). En b), les prédictions de la concentration en virus (en bleu) réalisées sur les spectres acquis en continu de la culture virale #5.

L'erreur de prédiction RMSEP obtenue à partir des prédictions du jeu test est évaluée à 1,09 unité logarithmique (u. log.) pour une erreur RMSECV estimée à 0,62 u. log. et une gamme d'étalonnage comprise entre 1,30–9,50 u. log. (le modèle repose sur huit variables latentes). Les performances prédictives dégagées par le modèle de régression laissent donc supposer que nous sommes en mesure de prédire de façon acceptable la concentration en virus malgré un biais (évalué à -0,96 u. log.). En effet, nous pouvons remarquer, Figure 5.4b, que le modèle de régression capte le profil exponentiel de la concentration en virus jusqu'à atteindre une valeur plateau, traduisant le déclin de la viabilité cellulaire dans le milieu et donc l'absence de cellules potentiellement hôte pour cultiver le virus. Toutefois, afin de vérifier si le modèle de régression se base bien sur les contributions spectrales du virus, nous appliquons celui-ci sur les spectres acquis en continu d'une culture de cellules, présentant donc, *a priori*, des teneurs nulles en virus. Les prédictions réalisées sur les spectres acquis en continu de la culture de cellules #1-1 sont présentées Figure 5.5.

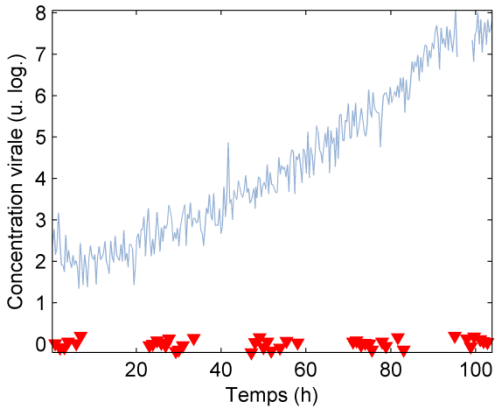

Figure 5.5 Prédictions de la concentration en virus (en bleu) réalisées sur les spectres acquis en continu de la culture de cellules #1-1. Les mesures références (triangles rouges) sont créées artificiellement autour de 0,00 u. log., en prenant en compte l'erreur de la méthode de référence de manière aléatoire.

Nous pouvons clairement voir que les concentrations obtenues ne sont pas nulles et présentent bien une évolution croissante, bien qu'il n'y ait pas de virus présent dans le milieu de culture. En réalité, ce profil d'évolution correspond à la croissance de la concentration en cellules dans le milieu de culture, ce qui signifie que le modèle de régression s'appuie, non pas sur les caractéristiques spectrales du virus mais sur celles des cellules Sf9. De ce fait, nous pouvons finalement conclure que le modèle de régression n'est pas satisfaisant pour réaliser les prédictions de la concentration en virus dans le milieu. Nous avons donc décidé d'apporter plus d'informations à l'ensemble de calibration afin de développer un modèle de régression performant s'appuyant uniquement sur les contributions spectrales du virus.

2.2 Modélisation du titre en virus infectieux tenant compte des cultures cellulaires

Ainsi, nous prenons en considération les cultures de cellules (phase d'amplification cellulaire) dans cette section. Puisque la concentration en virus est sensée être nulle au sein de ces lots, nous créons autant de références en concentration virale qu'il existait de mesures de référence des biomarqueurs. Nous obtenons finalement 42 échantillons pour la culture de cellules #1-1 et 50 échantillons pour la culture #2. De plus, en créant ces échantillons, nous prenons en compte l'erreur de la méthode de référence enzymatique afin de générer des valeurs aléatoires, non pas nulles, mais comprises entre -0,20 u. log. et 0,20 u. log. pour tenir compte de l'incertitude de mesure de la méthode de référence.

Nous incluons les échantillons de la culture cellulaire #2 à l'ensemble de calibration (puisqu'elle dispose d'un plus grand nombre de mesures que la culture #1-1), ce qui porte le nombre d'échantillons de calibration de 157 à 207. L'ensemble de la culture de cellules Sf9 #1-1 est ajouté au jeu test. Nous créons ainsi un nouveau modèle de régression sensé

inclure les variations qui existent lorsque le virus est présent ou non. Les performances prédictives du nouveau modèle de régression sont présentées Figure 5.6.

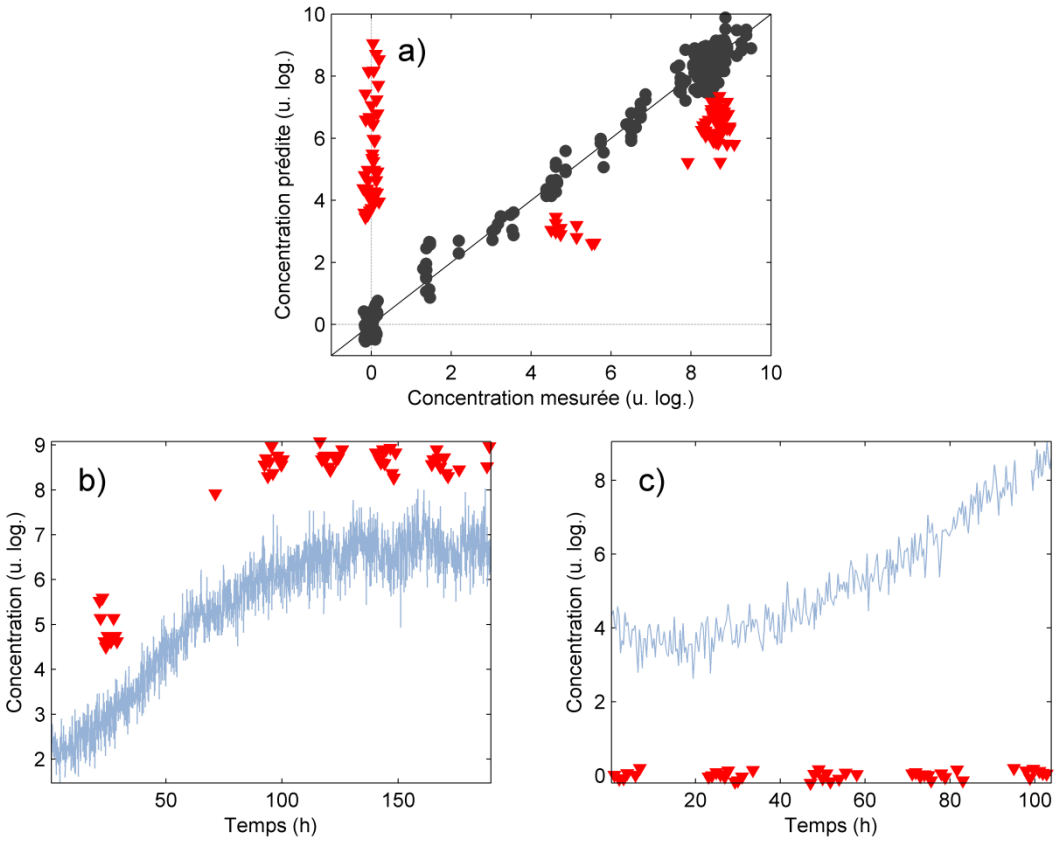

Figure 5.6 Représentation des performances prédictives du modèle de régression développé pour la détermination des concentrations en virus sur la base de quatre cultures virales et une culture cellulaire. En a), le graphique des concentrations prédites en fonction des concentrations mesurées pour les échantillons de calibration durant la validation croisée (points noirs) et pour les échantillons de l'ensemble test (triangles rouges). En b), les prédictions de la concentration en virus réalisées sur les spectres acquis en continu de la culture virale #5. En c), les prédictions de la concentration en virus réalisées sur les spectres acquis en continu de la culture cellulaire #1-1.

Le modèle de régression est développé sur neuf variables latentes, soit une de plus que pour le modèle précédent, ce qui signifie que l'ensemble de calibration présente plus de variabilité, notamment à cause de l'ajout des échantillons de la culture #1-1. Sur la Figure 5.6a, nous pouvons remarquer que les prédictions des échantillons de calibration présentent peu d'écart par rapport aux valeurs de référence. Ceci se traduit par une erreur RMSECV relativement faible (0,62 u. log.) compte tenu de la gamme de calibration (-0,18–9,50 u. log.). Toutefois, le biais de prédiction des échantillons de la culture virale #5, évalué à 1,16 u. log., ainsi que les prédictions des échantillons de la culture cellulaire #1-1, conduisent à une importante erreur de prédiction RMSEP de 4,11 u. log. Les prédictions des spectres acquis en continu pour les deux cultures test permettent d'observer, d'une part le biais de prédiction de la culture #5 (Figure 5.6b), et d'autre part les erreurs de prédiction de la culture #1-1 (Figure 5.6c). Sur ces-dernières, nous pouvons clairement voir que le modèle de régression prédit une croissance du titre en virus infectieux pour la culture de cellules Sf9, alors que le

virus n'est pas présent dans le milieu. Le modèle présente donc les mêmes aberrations que précédemment, ce qui montre une nouvelle fois qu'il ne repose pas sur les variables spectrales caractéristiques du virus.

2.3 Modélisation PLS2 pour la concentration en virus infectieux et les densités en cellules VCD et TCD

Il est acquis que le virus évolue de façon proportionnelle aux densités cellulaires puisque ce sont les cellules vivantes qui servent d'hôte au virus pour sa multiplication. Si nous représentons les évolutions de ces trois paramètres issues de mesures réalisées par les méthodes de référence, nous obtenons finalement les tendances présentées Figure 5.7.

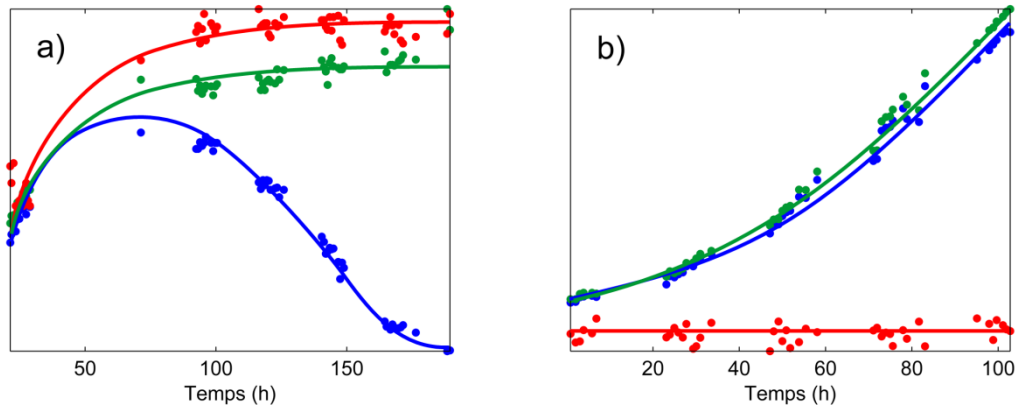

Figure 5.7 Représentations des tendances de la concentration en virus (rouge) et des densités VCD (bleu) et TCD (vert) durant a) la phase de culture virale (cas de la culture #5) et b) la phase d'amplification cellulaire (cas de la culture #1-1).

Nous avons donc une corrélation entre les densités cellulaires et la production de virus durant les cultures impliquant les cellules Sf9. Nous pouvons remarquer que pour les phases de culture virale (Figure 5.7a), les évolutions de la teneur en virus et de la densité totale en cellules sont très similaires. De ce fait, les prédictions de la concentration en virus réalisées sur les spectres des cultures cellulaires (Figure 5.5 et Figure 5.6c) présentent aussi la même évolution que la densité cellulaire totale au cours du bioprocédé de multiplication, présenté Figure 5.7b. Ainsi, nous décidons de construire, non pas un modèle de régression PLS prenant en compte une seule réponse, mais un modèle de régression pour plusieurs réponses simultanément, aussi appelé modèle PLS2. Nous reprenons les mêmes jeux de données que précédemment, soit un ensemble d'apprentissage composé des échantillons des cultures virales #3-1, #3-2, #4, #7 et #8 et la culture cellulaire #2. L'ensemble test reste également inchangé, composé des échantillons des cultures #1-1 (cellulaire) et #5 (virale). Le nombre de variables latentes nécessaires est déterminé à l'aide de la validation croisée initiale en consultant les erreurs RMSECV obtenues pour chaque réponse (titre en virus infectieux, densités VCD et TCD). Nous sélectionnons finalement onze variables latentes

pour le modèle, donc les erreurs RMSECV sont évaluées à 0,62 u. log. pour la concentration en virus, $0{,}45 \cdot 10^6$ cell/mL pour la densité VCD et $0{,}33 \cdot 10^6$ cell/mL pour la densité TCD. La Figure 5.8 permet de représenter les performances du modèle de régression obtenu.

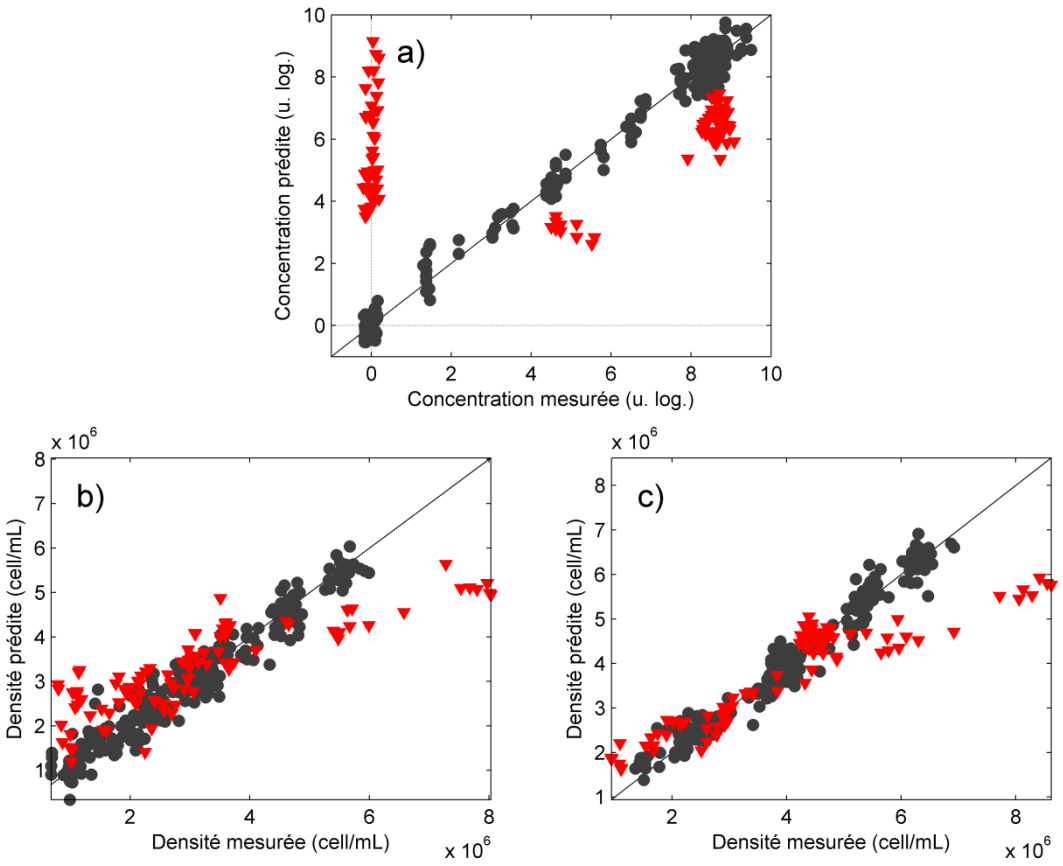

Figure 5.8 Représentations des graphiques des concentrations (ou densités) prédites en fonction des concentrations (ou densités) mesurées pour le modèle de régression PLS2 développé pour a) la concentration en virus infectieux, b) la densité en cellules vivantes et c) la densité en cellules totale. Les triangles rouges représentent les échantillons test, les points noirs les échantillons de calibration.

Les erreurs RMSEP sont déterminées à hauteur de 4,15 u. log. pour la concentration en virus, $1{,}16 \cdot 10^6$ cell/mL pour la densité VCD et $0{,}91 \cdot 10^6$ cell/mL pour la densité TCD. Au-delà du fait que l'erreur de prédiction est plus importante ici que lors de la modélisation PLS1 (modèle PLS à une seule réponse), nous pouvons constater que le graphique des concentrations en virus prédites en fonction des concentrations en virus mesurées, Figure 5.8a, que les résultats sont relativement similaires. Nous retrouvons le biais de prédiction observé sur les échantillons de la culture virale #5 et les erreurs obtenues pour la culture cellulaire #1-1. Ainsi, nous pouvons donc conclure que la prise en compte de la corrélation entre les densités cellulaires et les concentrations en virus, à travers la modélisation PLS2, ne permet pas d'accéder à de réelles prédictions du titre en virus infectieux dans le milieu.

2.4 Création de références supplémentaires pour la modélisation

Les derniers travaux présentés pour la modélisation de la concentration en virus portent sur la création de références pour tous les spectres des cultures disponibles. Lors des cultures virales, ce n'est pas l'introduction du virus qui établit le temps zéro, mais l'introduction des cellules dans le bioréacteur de culture. L'inoculation virale est réalisée plus tard. À titre d'exemple, le virus est introduit à partir de 21 h pour la culture virale test (#5), ou encore à prêt de 23 h pour la culture virale #8. Ainsi, nous pouvons donc considérer que la concentration en virus infectieux est nulle avant ce temps de culture. De plus, nous disposons d'une grande quantité de spectres Raman acquis en continu pour les différentes cultures. L'idée, au cours de cette section, est donc d'attribuer à chaque spectre une valeur de concentration en virus.

Pour cela, nous reprenons pour chaque culture, cellulaire ou virale, les mesures de la concentration en virus réalisées. Nous ajustons ensuite une courbe d'évolution sur les mesures de référence. Dans notre cas, nous appliquons une fonction sigmoïde sur les données de concentration en virus infectieux à l'aide de la boite à outils Curve Fitting Toolbox™ (The MathWorks) disponible sous environnement MATLAB® (The MathWorks). Nous considérons l'équation sigmoïdale logistique, traduite par la fonction décrite équation (5.1), où x_0 est la valeur centre, où la valeur de y est le milieu des deux valeurs limites A_1 et A_2, p étant un paramètre de puissance.

$$y = \frac{A_1 - A_2}{1 + \left({}^{x}/_{x_0}\right)^{p}} + A_2 \tag{5.1}$$

En considérant les écarts entre les mesures de référence, les concentrations calculées sont aléatoirement comprises entre -0,4 u. log. et 0,4 u. log. autour de la valeur y déterminée artificiellement. Un exemple de représentation des références ainsi réalisées est proposé Figure 5.9 où les échantillons utilisés avant et après l'ajout des mesures artificielles pour la culture virale #8 sont exposés. Nous pouvons souligner le fait que les spectres acquis avant l'introduction du virus (ne présentant donc pas de réelle mesure de référence pour la concentration en virus infectieux) présentent également des mesures de référence artificielles comprises entre -0,40 u. log. et 0,40 u. log. autour de la valeur nulle. Ainsi, en plus des échantillons des cultures cellulaires impliquées, nous bénéficions à la fois des spectres présentant des mesures supposément nulles et des spectres durant la culture virale pour un seul et même lot, ce qui nous permettrait d'accroître la distinction spectrale entre la présence ou non de virus au sein des cultures.

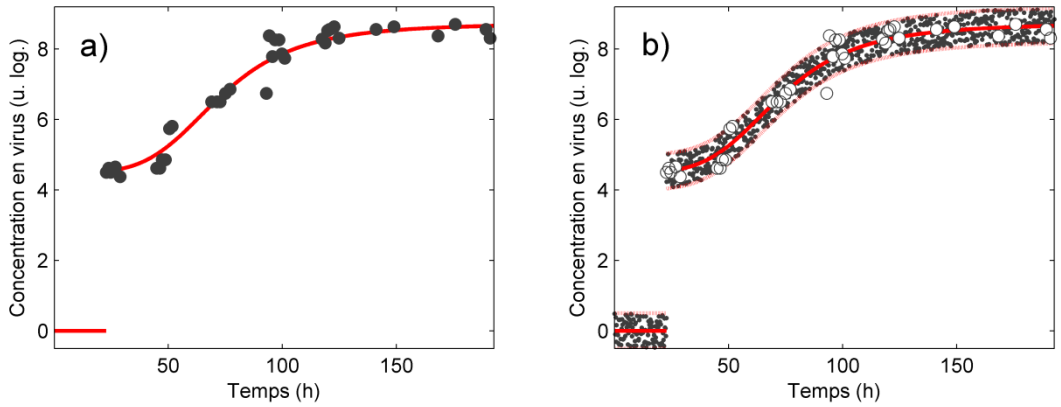

Figure 5.9 Représentation de la création de mesures de référence artificielles afin d'obtenir un jeu de données représentatif de l'évolution de la concentration en virus durant les cultures virales. Sont représentées en a) les mesures des référence de la concentration en virus pour la culture #8 (points noirs) ainsi que la tendance des concentrations (ligne rouge), en b) les mesures de référence nouvellement créées (points noirs) répartis aléatoirement autour de la tendance de la concentration (ligne rouge) et ajoutés aux vraies mesures de référence (cercles noirs).

Le modèle de régression déterminé sur la base de tous les échantillons (artificiels ou non) est présenté Figure 5.10a. Bien que l'erreur RMSECV (0,91 u. log.) obtenue soit relativement acceptable compte tenu du volume de données pris en compte, nous pouvons clairement noter la présence d'un biais très important pour les prédictions des échantillons test, l'erreur RMSEP étant évaluée à 11,41 u. log. Toutefois, ce biais peut être engendré par la sur-modélisation du titre en virus infectieux pour les cultures utilisées en tant que lot de calibration. Il convient donc de réduire la taille de la population de l'ensemble d'apprentissage. Nous sélectionnons alors un échantillon sur dix au sein de l'ensemble de calibration et calculons le modèle de régression, présenté Figure 5.10b.

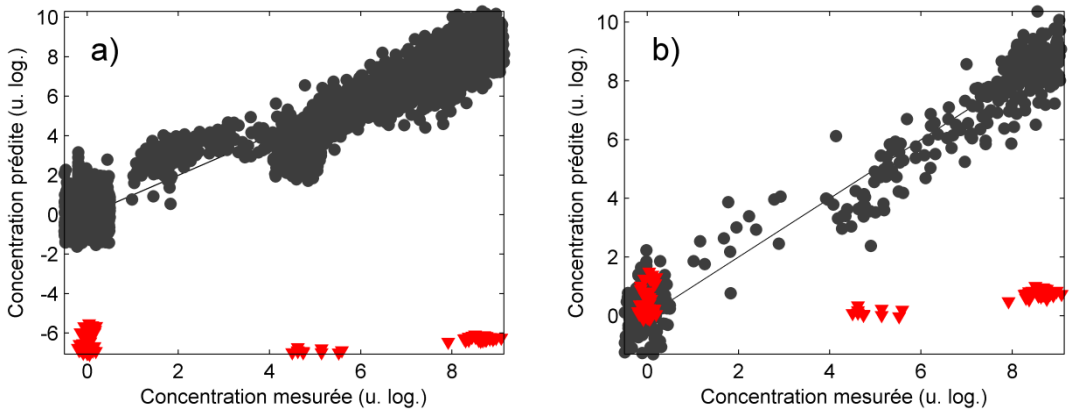

Figure 5.10 Représentations des modèles de régression calculés sur la base des échantillons du jeu de données comprenant a) toutes les références artificiellement calculées et b) un échantillon sur dix de l'ensemble de calibration. Les échantillons de calibration (cultures #2, #3-1, #3-2, #4, #7 et #8) sont les points noirs, les triangles rouges étant les échantillons test (cultures #1-1 et #5).

L'erreur RMSEP est alors grandement réduite (5,57 u. log.) pour une erreur RMSECV équivalente (0,99 u. log). Nous pouvons notamment souligner les prédictions acceptables de la culture cellulaire, présentant beaucoup moins de biais. Néanmoins, les échantillons de la

culture virale (mesures de référence comprises entre 4,00–9,00 u. log.) présentent de très importants écarts de prédiction, d'où la forte erreur RMSEP calculée. Le modèle de régression n'est donc pas adapté à la prédiction de la teneur en virus infectieux des cultures virales, ce qui signifie qu'à partir des mesures de référence créées artificiellement, nous ne pouvons pas générer de modèle de régression robuste à la prédiction de la teneur en virus pour les cultures virales et cellulaires.

Par la suite, d'autres travaux ont été entrepris pour tenter d'obtenir un modèle de régression satisfaisant. D'abord, les algorithmes génétiques ont été employés, suivant la méthode GA-PLS (pour *Genetic Algorithms - Partial Least Squares*) [81], afin de retrouver un nouveau prétraitement spectral qui permettrait de corriger les données de façon plus efficace pour générer un modèle de régression robuste. Toutefois, le critère d'optimisation mis en jeu étant l'erreur RMSECV, nous avons une nouvelle fois été confronté à un surentrainement du modèle de régression pour les données de calibration, et donc un important biais sur les prédictions des échantillons test. Nous avons ensuite appliqué la méthode GA-PLS en co-optimisant d'une part le prétraitement spectral et d'autre part la sélection de variables, sans succès. Puis nous avons appliqué différentes techniques de sélection de variables, dont la sélection par intervalles (i-PLS pour *interval – Partial Least Sqaures*) [163,164], l'élimination de variables non-informatives (UVE-PLS pour *Uninformative Variable Elimination – Partial Least Squares*) et ses variantes [165,166], ainsi que l'optimisation à l'aide de l'algorithme des colonies de fourmis (*Ant colony optmisation*) [167]. Cependant, aucune des techniques citées n'a permis de résoudre la problématique liée à la modélisation de la teneur en virus dans le milieu durant les cultures virales.

Finalement, les importantes erreurs de prédiction obtenues traduisent indéniablement le manque de fiabilité du modèle PLS pour la détermination du titre viral. La détermination directe de la teneur en virus à partir de spectres Raman reste donc un réel challenge, contrairement à la prédiction de la concentration en anticorps dans le milieu pour les cultures de cellules. Nous pouvons alors nous interroger quand à la limite de détection de la spectroscopie Raman pour savoir si cette technique pourrait permettre de prédire les concentrations virales. En effet, pour la suite des travaux sur la prédiction du titre viral, il faut donc déterminer d'une part si le virus est suffisamment présent dans le milieu pour être détecté par la spectroscopie Raman, et d'autre part si cette dernière est en mesure de détecter les contributions chimiques du virus dans le milieu de culture.

Puisque la prédiction directe du titre en virus est difficile, une solution alternative pourrait être mise en place en développant une méthode prenant en compte les densités cellulaires TCD et VCD. En effet, nous avons pu voir que le titre en virus infectieux est dépendant de la viabilité cellulaire dans le milieu de culture virale, Figure 5.7. De ce fait, en considérant les

prédictions réalisées en ligne et en temps réel pour les densités cellulaires grâce aux modèles de régression et au dispositif d'acquisition Raman (sonde à immersion), il pourrait être possible de développer une équation pour déterminer dans un second temps la concentration virale. Les calculs chimiométriques réalisés en temps réel pour les prédictions des densités cellulaires étant quasi-instantanés, l'application de cette équation permettrait d'accéder au titre en virus infectieux en temps réel.

3 Maitrise statistique des bioprocédés à l'aide de données multivariées

Afin d'assurer la qualité d'un produit manufacturé lors de sa fabrication, une technique de contrôle très fréquente consiste à suivre statiquement un paramètre au cours du temps, nous parlons aussi de maitrise statistique des procédés (SPC pour *Statistical Process Control*). Il s'agit des cartes de contrôle qui, une fois mises en place, permettent de détecter les déviations d'un procédé de fabrication par rapport à un état référent pris en compte lors du développement de la carte. En effet, celles-ci sont réalisées sur la base de mesures d'un paramètre pour les échantillons d'une population de référence. À partir de ces mesures, il est ensuite possible de déterminer une valeur moyenne cible pour le paramètre pris en compte, ainsi que des bornes limites au-delà desquelles l'individu dont provient l'échantillon analysé est considéré comme hors de contrôle.

Aujourd'hui, dans le milieu biopharmaceutique et plus précisément pour les cultures de cellules, les cartes de contrôle sont principalement basées sur le suivi de paramètres conditionnant l'état physique du milieu de culture. En effet, ces cartes univariées reposent sur le suivi en temps réel de paramètres tels que le pH, la température ou encore la pO_2 du milieu, qui, combinés, permettent d'avoir une vision générale de la culture cellulaire en cours [168,169]. Toutefois, il existe aussi des cartes de contrôle multivariées qui prennent en compte plusieurs paramètres univariés [170]. Le bioprocédé est modélisé à l'aide d'une analyse ACP réalisée sur le cumul des paramètres conditionnant la culture. Les cartes de contrôles ainsi élaborées permettent de suivre l'évolution globale du procédé à travers les composantes principales de l'analyse ACP. Nous pouvons parler de contrôle SPC multivarié, ou MSPC (*Multivariate Statistical Process Control*).

Cependant, ces cartes de contrôle ne s'appuient que sur des données physiques et ne représentent donc pas le vrai caractère biochimique de la culture de cellules qui est variable par nature. C'est pourquoi nous mettons en œuvre les spectres Raman acquis en continu lors des travaux précédents, qui contiennent toute l'information complexe contenue au sein

du milieu de culture, pour développer des cartes permettant de suivre les bioprocédés et détecter les cultures anormales.

3.1 L'analyse en composantes principales multivoie

Ainsi, des moyens statistiques ont été développés pour prendre en compte les échantillons présentant un très grand nombre de variables telles que les données spectroscopiques. C'est notamment le cas de l'analyse en composantes principales à plusieurs dimensions (MPCA pour *Multiway Principal Component Analysis*) [171], qui reste la méthode la plus employée pour implémenter le MSPC sur les procédés. Cette méthode est relativement proche de l'analyse PCA puisqu'elle emploie les mêmes algorithmes de calcul. Si nous prenons l'équation (2.3), page 62, nous pouvons la réécrire considérant les dimensions des matrices, équation (5.2), où i traduit le nombre de cultures à disposition, j le nombre de variables spectrales, k le temps de culture et n le nombre de composantes principales pris en compte.

$$\mathbf{X}_{i\times j\times k} = \mathbf{T}_{i\times n}\,\mathbf{P}^{\mathrm{T}}{}_{j\times k\times n} + \mathbf{E}_{i\times j\times k} \tag{5.2}$$

Elle nécessite néanmoins l'application d'une étape de dépliement des matrices afin de passer d'un jeu de données tridimensionnel à un jeu bidimensionnel. Cette opération est représentée Figure 5.11, où nous passons d'un jeu en trois dimensions à un ensemble bidimensionnel. L'orientation suivant le nombre de cultures reste inchangée tandis que la taille de la seconde dimension de la matrice nouvellement formée est le produit des dimensions suivant les déplacements Raman (variables spectrales) par les temps de culture (ou nombre d'échantillons). Si nous considérons I la taille de la dimension suivant le nombre de cultures, J le nombre de variables spectrales et K le nombre de spectres disponibles pour chaque culture (soit le nombre de temps disponibles), nous passons alors d'une matrice tridimensionnelle $I \times J \times K$ à une matrice plane de dimensions $I \times (J \times K)$.

Si nous considérons une culture cellulaire test, extérieure à l'ensemble de calibration du modèle MPCA, nous pouvons alors déterminer si ce lot test peut être affilié ou non à l'ensemble d'apprentissage. Toutefois, l'application du modèle MPCA n'est réalisée qu'en fin de culture, contraignant donc la détection qui interviendrait *a posteriori*. C'est pourquoi il est nécessaire d'adopter une approche fractionnée du jeu de données de calibration afin d'être capable de déceler directement en temps réel les déviations des procédés de culture cellulaire et déclencher instantanément une alerte.

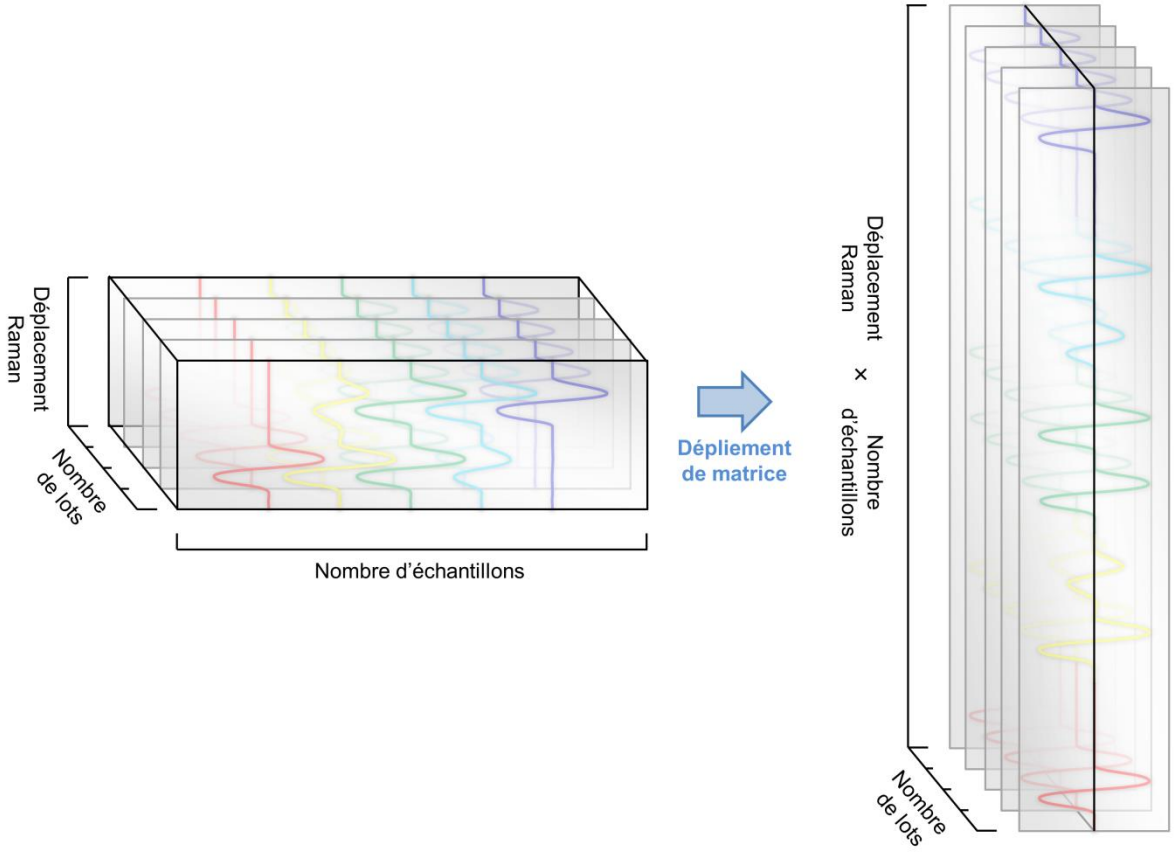

Figure 5.11 Représentation du dépliement effectué sur la matrice tridimensionnelle contenant les spectres Raman pour réaliser l'analyse MPCA.

3.2 Synchronisation et fractionnement de l'ensemble de calibration

Le principal inconvénient des jeux de données provenant de cultures cellulaires vient de la nature variable du support biologique. En effet, au-delà des caractéristiques propres de chaque culture, il est important de noter que le calcul du temps de culture écoulé au moment de l'acquisition d'un spectre dépend irrémédiablement de la date (date et heure exactes) de l'introduction des cellules au sein du milieu de culture, conditionnant alors le temps t = 0 h. Cependant, le protocole d'acquisition des spectres Raman est amorcé avant l'introduction des cellules afin d'être sûr d'acquérir les données spectroscopiques du début à la fin de la culture. De ce fait, il peut apparaître certains écarts dans les temps des spectres, ce qui est présenté à travers les cas A et B de la Figure 5.12.

De plus, le dispositif d'acquisition des spectres Raman utilisé lors de ces travaux de recherche permet de réaliser le suivi spectroscopique de plusieurs cultures simultanément. Cependant, il est impossible d'acquérir deux spectres provenant de deux sondes différentes au même moment (les sondes étant reliées au même spectromètre et disposant donc d'un détecteur commun). De ce fait, le protocole d'acquisition prévoit d'alterner les acquisitions,

cas C et D présentés Figure 5.12. Ainsi, les temps de culture attribués aux spectres et les volumes de données présentent des différences d'une culture à l'autre.

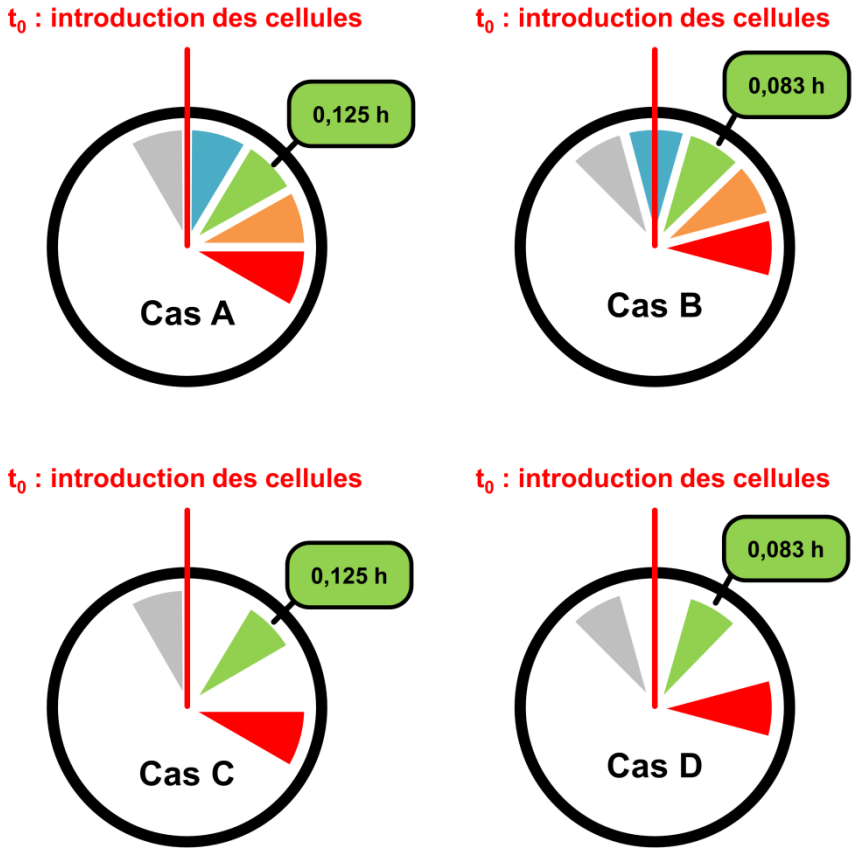

Figure 5.12 Schéma des différentes configurations possibles du procédé d'acquisition des spectres Raman pour le suivi spectroscopique des cultures cellulaires. Le cas A met en jeu une seule sonde, l'acquisition du premier spectre débutant lors de l'introduction de cellules. Le cas B met également en jeu une seule sonde, mais l'introduction des cellules est réalisée durant l'acquisition d'un spectre. Les cas C et D présentent les mêmes caractéristiques d'acquisition que les cas A et B respectivement, mettant cette fois deux sondes en jeu.

Afin de corriger ces différences, une technique de synchronisation est employée et permet d'aligner les temps des différents spectres en allongeant ou compressant les trajectoires des procédés afin qu'elles soient le plus corrélées possible à un profil de référence assigné. Cette techniques, intitulée *Correlation Optimized Warping* (COW) [172], permet finalement de conditionner les données sous forme de cube afin de pouvoir appliquer les principes des techniques telle que l'analyse MPCA présentée précédemment. Une représentation visuelle des résultats de l'application de la technique COW est proposée Figure 5.13.

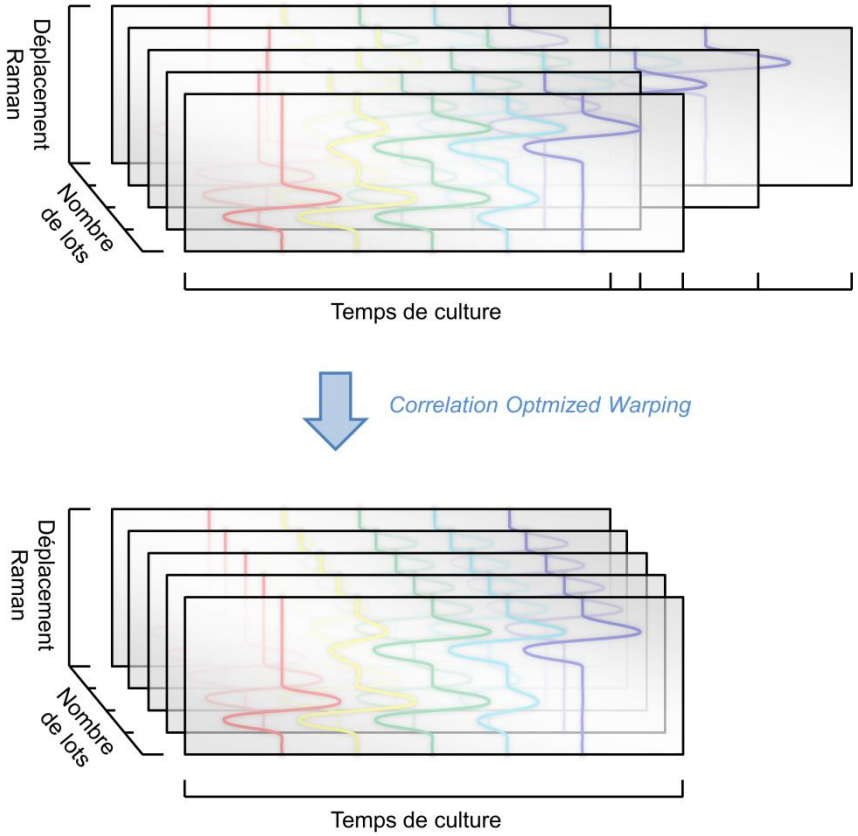

Figure 5.13 Représentation de l'application de la méthode COW pour la synchronisation des données acquises pour chaque culture (lot).

Bien que les données soient correctement conditionnées, l'application de MPCA ne permettrait tout de même pas de réaliser un suivi spectroscopique en temps réel de l'évolution de la culture. En effet, l'application du modèle MPCA serait réalisée sur l'ensemble du lot test, soit nécessairement en fin de culture. C'est pourquoi nous fractionnons le jeu de données en plusieurs sous-ensembles permettant de représenter l'état d'une culture de référence au temps t. Néanmoins, le fractionnement doit être homogène sur l'ensemble du bioprocédé et doit être optimisé afin, d'une part, de ne pas prendre d'intervalles trop courts. En effet, ceci pourrait conduite à développer des modèles MPCA non-représentatifs présentant un nombre d'échantillons trop insuffisant pour réaliser les analyses MPCA. D'autre part, en prenant des intervalles trop longs, nous empêcherions toute rétroaction sur la culture puisque la détection serait réalisée trop tardivement. La Figure 5.14 présente le fractionnement du jeu de données initial en plusieurs sous-ensembles qui permettront de construire des modèles MPCA représentatifs de régions chronologiques traduisant l'avancement du bioprocédé.

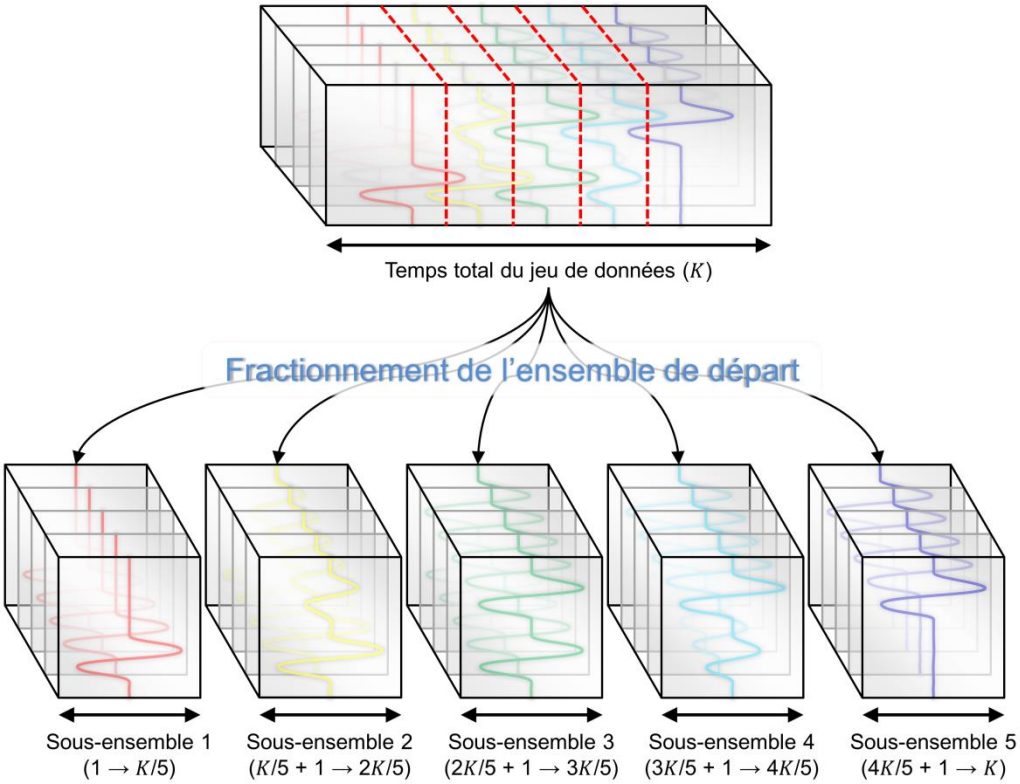

Figure 5.14 Représentation du fractionnement de la matrice de données initiale par rapport au temps de culture en plusieurs sous-ensembles de taille homogène.

C'est sur ces différents sous-ensembles que l'analyse MPCA sera réalisée, ce qui permet *in fine* de modéliser l'état du bioprocédé pour la période balayée. En prenant les échantillons d'un ensemble test absent de l'ensemble d'apprentissage initial, nous pouvons alors appliquer les modèles MPCA et calculer certains critères afin de vérifier si la culture test suit le profil de référence ou si elle présente une trajectoire différente.

3.3 Évaluation des modèles MPCA et développement des cartes de contrôle

Afin de vérifier la faisabilité de l'application du contrôle MSPC à l'aide de spectres Raman, nous n'avons travaillé que sur les échantillons provenant de cultures du même type de cellule, en l'occurrence, les cellules CHO. Au total, dix cultures cellulaires sont mises en jeu dont huit réalisées dans les conditions normales de cultures qui tendent donc à suivre le profil normal d'évolution du bioprocédé. De plus, un lot a été réalisé en stoppant les *feeds* après 10 j de culture (dernier *feed* à 245 h de culture), ce qui provoque une pénurie en glucose dans le milieu et donc une forte mortalité cellulaire. Enfin, une dernière culture, également réalisée dans les conditions normales (*feeds* compris), présente une contamination depuis le départ, détectée uniquement en fin de culture.

D'une part, une des huit cultures normales a été exclue suite à certains problèmes d'acquisition rencontrés. D'autre part, une autre culture normale, la culture stoppant le *feeding* et la culture contaminée constituent l'ensemble test pour évaluer les modèles MPCA, développés à partir d'échantillons provenant donc de sept cultures cellulaires. Il convient de rappeler ici que les échantillons pris en compte ne sont pas les mesures de référence et spectres Raman associés tels que les travaux réalisés lors des sections précédentes, mais bien tous les spectres Raman acquis en continu pour chaque culture de cellules CHO. Pour une culture normale, nous considérons que le bioprocédé dure 300 h, ce qui présente donc un total de près de 900 spectres (minimum) par culture, les spectres étant acquis sur 10×30 s (5 min). Rappelons que dans le cas où quatre sondes sont employées simultanément, un spectre Raman de 5 min est acquis toutes les 20 min, les 15 min restantes étant allouées aux autres sondes. C'est pourquoi nous disposons d'un minimum de 900 spectres par culture. Enfin, nous segmentons l'ensemble de calibration en sous-ensembles recouvrant des périodes d'une heure, ce qui est assez important pour bénéficier de suffisamment de spectres pour les modélisations MPCA, mais qui reste relativement court compte tenu des 300 h de culture. L'information sur une éventuelle contamination n'est donc pas instantanée, mais parvient au maximum une heure après l'acquisition des spectres Raman.

Deux graphiques des *scores* provenant des analyses MPCA réalisées à 100 h et 300 h sont proposés Figure 5.15. Les graphiques comprennent les *scores* de chaque culture dans un espace à trois dimensions composé des premières composantes principales, pour un total de près de 94 % de variance expliquée dans les deux cas. Un critère traduisant les variations existantes durant le développement du modèle MPCA est le T^2 de Hotelling qui, en partant de l'hypothèse que les *scores* sont distribués normalement, permet de représenter une limite de contrôle autour des *scores* des cultures de l'ensemble de calibration suivant l'équation (5.3) [173], où I est le nombre de lots de l'ensemble de calibration, N le nombre de composantes principales du modèle MPCA, et $\mathrm{F}_\alpha(p, m-p)$ la valeur de la distribution de Fisher au risque α pour p et $m-p$ degrés de liberté.

$$T^2 \sim \frac{(I-1)(I+1)}{I(I-N)} \mathrm{F}_\alpha(N, I-N) \tag{5.3}$$

La représentation de cette limite prend la forme d'un ellipsoïde sur les graphiques des *scores* présentés Figure 5.15. Les *scores* des cultures de l'ensemble d'apprentissage sont tous compris dans cette limite. En appliquant les modèles MPCA aux cultures test, il est alors possible de représenter les *scores* de ces cultures sur les graphiques. En considérant le modèle MPCA à 100 h, Figure 5.15a, nous pouvons voir que les cultures test #8 (culture normale) et #9 (culture stoppant le *feeding* vers 245 h) sont comprises au sein de l'ellipsoïde tracé par le coefficient T^2 calculé. Nous pouvons également remarquer que la culture #10,

contaminée depuis le départ, présente un *score* en dehors de la limite de Hotelling calculée. En prenant maintenant les résultats obtenus par le modèle MPCA à 300 h de culture, Figure 5.15b, nous devons d'abord signaler l'absence de résultat pour la culture #10. En effet, celle-ci est stoppée à 245 h. Ensuite, nous pouvons remarquer que les *scores* calculés pour la culture test #9 sont en dehors des limites définies par l'ellipsoïde de Hotelling, ce qui pourrait bien s'expliquer par l'arrêt des *feeds* entrainant un forte mortalité des cellules et donc une déviation du procédé.

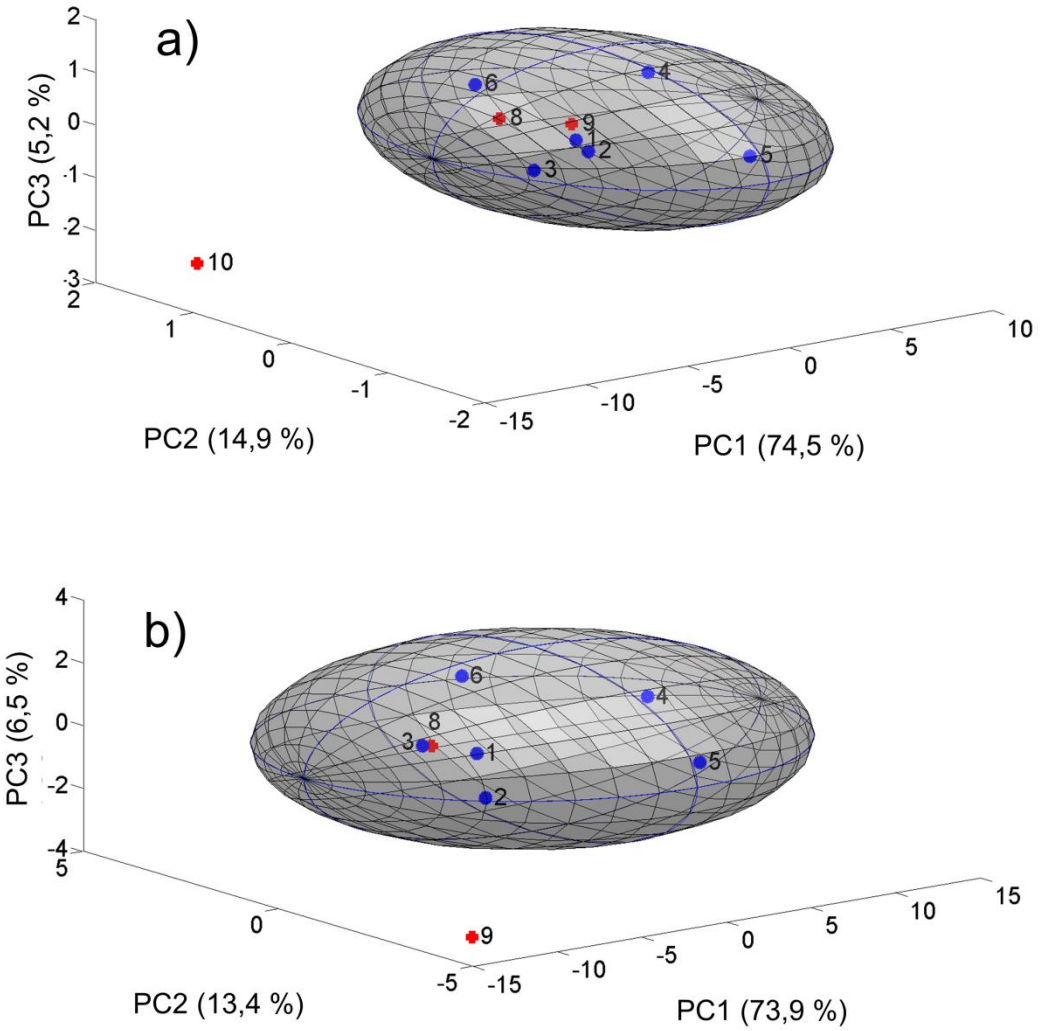

Figure 5.15 Représentations des graphiques des scores pour les trois premières composantes principales de l'analyse MPCA à a) 100 h de culture et b) 300 h de culture. Les cultures de l'ensemble de calibration sont représentées par les points bleus, les croix rouges traduisant les cultures test. La culture #10 est absente de la partie b) car le procédé a été stoppé autour de 245 h.

Toutefois, au-delà de la simple interprétation des projections des cultures test sur les graphiques des *scores* obtenus pour chaque modèle MPCA calculé, il est nécessaire d'instaurer des facteurs statistiques qui peuvent être calculés au cours du temps pour

l'élaboration de cartes de contrôle du bioprocédé. Le premier critère possible est le coefficient T^2 de Hotelling traduisant les variations au sein du modèle MPCA. Pour un lot i donné, le critère est obtenu suivant l'équation (5.4) où $\mathbf{t}_i$ est le $i^{\text{ème}}$ vecteur de la matrice des *scores* $\mathbf{T}_n$ de la composante principale n [174].

$$T_i^2 = \mathbf{t}_i\left(\mathbf{T}_n^{\mathrm{T}}\mathbf{T}_n\right)^{-1}\mathbf{t}_i^{\mathrm{T}} \tag{5.4}$$

La limite de contrôle pour le coefficient T^2 de Hotelling a été présentée équation (5.3). Un second critère possible est la statistique Q qui présente le manque d'adéquation du modèle MPCA pour les cultures test. Ce critère se calcule sur la base des résidus obtenus suite à l'application du modèle suivant l'équation (5.5) [174], où $\mathbf{e}_i$ est le vecteur des résidus pour la culture i, $\mathbf{x}_i$ est le vecteur représentant la culture i (ici le vecteur déplié provenant de la matrice contenant les spectres Raman), $\mathbf{P}_n$ la matrice des *loadings* issue de l'analyse MPCA pour la composante n, et $\mathbf{I}$ la matrice identité.

$$Q_i = \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i\left(\mathbf{I} - \mathbf{P}_k\mathbf{P}_k^{\mathrm{T}}\right)\mathbf{x}_i^{\mathrm{T}} \tag{5.5}$$

La limite de contrôle calculée pour la statistique Q est présentée équation (5.6) et se base sur les distributions χ^2, prenant $(I-1)$ pour degré de liberté, avec I le nombre de cultures. La valeur de la distribution est pondérée par la variance des résidus σ_r^2, ce qui génère donc une limite de contrôle variable au cours du temps.

$$Q \sim \sigma_r^2 \chi_{I-1}^2 \tag{5.6}$$

La Figure 5.16 présente les résultats pour le suivi des paramètres T^2 et Q pendant toute la durée considérée des bioprocédés, soit 300 h de culture. Tout d'abord, en prenant la culture cellulaire contaminée depuis le début (culture #10), nous pouvons remarquer que quelque soit la carte envisagée, il apparait clairement que la culture est en dehors des limites de contrôle. Ensuite, en ce qui concerne la culture #9, nous pouvons remarquer que les coefficients T^2 et Q sortent des limites au-delà de 260 h. Ceci est cohérent avec la réalité de la nature du milieu contenu dans le bioréacteur : le *feeding* a été stoppé à 245 h de culture. Mais le temps que tout le glucose soit entièrement consommé par les cellules, la chute drastique de la viabilité intervient plusieurs heures (quinze en l'occurrence) après le dernier *feed*. Enfin, nous pouvons voir que les niveaux des critères d'évaluation de la culture #8 (culture de test normale) ne dépassent jamais les limites de contrôle fixées.

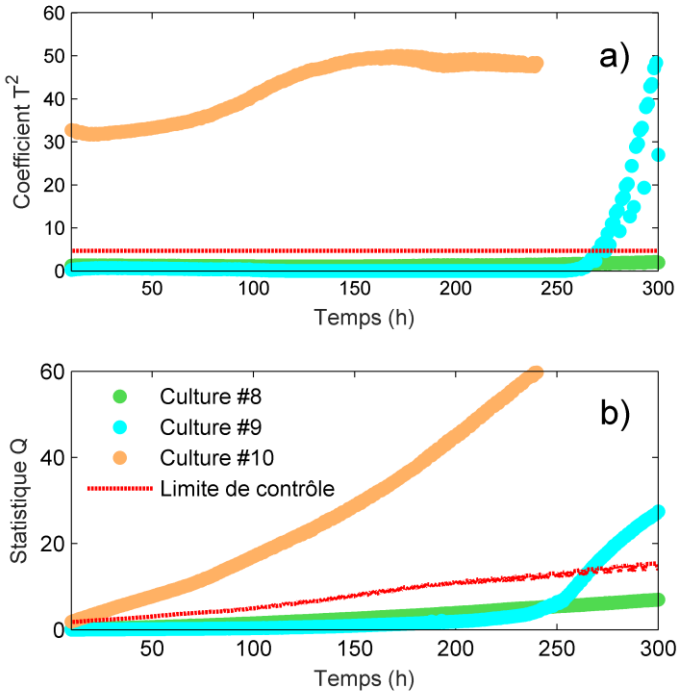

Figure 5.16 Représentation des cartes de contrôle réalisées pour les cultures test #8 (culture normale), #9 (culture dont le dernier *feed* est à 245 h) et #10 (culture contaminée depuis le début) sur la base a) du T^2 de Hotelling et b) de la statistique Q.

Il apparaît donc finalement que les cartes de contrôle établies sur la base des variations entre les cultures (coefficient T^2) et du manque d'adéquation lors de l'application du modèle MPCA sur une culture test (statistique Q) permettent de détecter les cultures anormales. Que ce soit depuis le début de la culture, cas de la culture #10, ou lors d'un évènement particulier, cas de la culture #9, nous pouvons remarquer que ces cartes permettent de détecter rapidement d'éventuels disfonctionnements. Ainsi, l'application de ces modèles MPCA pour instaurer un contrôle MSPC sur les cultures cellulaires permet d'alerter quant à d'éventuelles contaminations des cultures. Toutefois, il est important de rappeler que ces techniques ne permettent pas d'accéder à la cause du disfonctionnement du bioprocédé. Il faut donc recueillir cette information suivant les méthodes traditionnelles hors ligne.

4 Conclusions sur la valorisation des données biologiques et spectroscopiques

Au cours de ce chapitre, nous avons présenté divers travaux mettant en jeu les données biologiques et spectroscopiques. Les premières recherches concernent la modélisation de la concentration en anticorps dans les cultures de cellules CHO. En effet, les anticorps sont des produits d'intérêt des cultures cellulaires. C'est pourquoi la capacité à déterminer en ligne et en temps réel leur teneur présente un intérêt certains pour la production

biopharmaceutique. La mise en place d'un système d'acquisition de spectres Raman *in situ*, couplé à un modèle chimiométrique performant permet de réaliser ce suivi. Le modèle PLS développé atteint des niveaux d'erreurs acceptables compte tenu de la gamme de variation rencontrée, ce qui permet de valider la faisabilité d'une méthode de détermination de la concentration en anticorps en ligne et en temps réel à l'aide des outils spectroscopiques et chimiométriques.

Si pour les cultures de cellules, le principal intérêt concerne la production d'anticorps, pour les cultures virales, le produit d'intérêt est indéniablement le virus. Nous avons alors déployé les mêmes outils pour la modélisation de la concentration en virus. Cependant, l'application de la régression PLS ne permet pas de générer de modèle satisfaisant. C'est pourquoi nous avons d'abord appliqué des techniques plus avancées, prenant notamment en compte la corrélation qui existe entre la concentration virale et les densités en cellules totales et vivantes au sein d'une PLS2 (PLS ayant une matrice comportant plus d'une réponse). De plus, nous avons également tenté d'intégrer aux modèles de régression des informations supplémentaires provenant de cultures saines, ne présentant donc pas de virus, afin de procurer au modèle de régression des informations concernant les différences spectrales existantes entre la présence et l'absence de virus. Toutefois, les résultats montrèrent que les prédictions de la concentration virale ne permettaient toujours pas de réaliser le suivi du titre en virus infectieux. Enfin, nous avons généré de nouvelles références développées à partir des tendances (sigmoïdes) des concentrations en virus mesurées depuis les prélèvements du milieu. Néanmoins, nous ne sommes toujours pas en mesure de développer de modèle acceptable. D'autres outils, telles que diverses méthodes de sélection de variables, ont également été employés mais ne permettent pas non plus de fournir de modèle de régression acceptable pour la détermination de la teneur en virus. Une solution envisageable serait alors d'investir des techniques de calculs non-linéaires, type réseaux de neurones afin de pouvoir corréler les spectres Raman aux mesures de la concentration en virus dans le milieu de culture.

Le dernier point abordé dans le chapitre a pour but de valoriser les données spectrales en développant des cartes de contrôle multivariées. En effet, les cartes pour le suivi, le contrôle et la détection de possibles contaminations ou de déviations potentielles du bioprocédé sont généralement univariées dans le sens où elles ne reposent que sur un seul paramètre, généralement physique (type pH, température, pO_2). Il existe déjà des cartes de contrôle multivariées qui s'appuient sur des analyses en composantes principales développées à partir des différents paramètres physiques mesurés. Néanmoins, ces cartes ne retracent que les caractéristiques physiques de la culture, soit les conditions de culture, et ne traduisent donc pas l'état biologique du bioprocédé. Il n'est donc pas possible, à l'aide de

ces cartes multivariées, de déceler une quelconque contamination du milieu. C'est pourquoi nous proposons des cartes de contrôles multivariées tenant compte, non pas des paramètres physiques de culture, mais des spectres Raman acquis en continu tout au long des cultures cellulaires. Ces travaux, développés à partir des données recueillies pour les cellules CHO, proposent d'utiliser des modèles MPCA générés à partir de cultures de référence, réalisées dans les conditions normales, et de les appliquer sur trois cultures test. Un modèle est calculé pour chaque tranche horaire d'une heure, proposant ainsi près de 300 modèles pour le contrôle MSPC des cultures à appliquer aux échantillons correspondant des cultures test. Parmi ces trois cultures, nous retrouvons une culture réalisée dans les conditions normales ne présentant aucun problème particulier, une culture dont le dernier *feed* est réalisé à près de 245 h de culture, provoquant ainsi une pénurie en glucose dans le milieu au bout de 260 h et donc une forte mortalité cellulaire, et enfin le troisième lot test présente une contamination extérieure depuis l'introduction des cellules en début de culture. Si nous appliquons uniquement les modèles MPCA aux cultures test, nous sommes en mesure d'avoir un aperçu du comportement général des cultures vis-à-vis de l'ensemble de calibration constitué des cultures acquises dans les conditions normales. Cependant, l'évaluation du comportement nécessite la mise en place de critères mathématiques permettant de fixer les limites de contrôle des différentes cartes. Nous utilisons ainsi le coefficient T^2 de Hotelling qui permet de comparer la position d'une culture test par rapport aux variations rencontrées lors de l'élaboration du modèle MPCA, et le critère statistique Q qui permet de juger le manque d'adéquation d'un modèle MPCA sur l'ensemble test. Nous avons alors montré que les deux cartes proposées permettaient de correctement interpréter les comportements observés pour les cultures test. Nous pouvons donc dire qu'il est possible de mettre en place de cartes multivariées pour le suivi et le contrôle de l'état biologique d'une culture à l'aide de spectres Raman acquis *in situ* et en continu. Dans notre cas, ceci pourrait permettre de bénéficier d'informations dans l'heure sur l'apparition d'un comportement anormal ou une éventuelle contamination de la culture. Ces informations pourraient finalement conduire à stopper le bioprocédé de culture avant son terme pour un gain en temps et en ressources.

Conclusions

Les recherches présentées au cours de ces travaux s'inscrivent directement au sein du projet global CellPAT avec pour objectif de développer des capteurs de mesure de concentrations métaboliques de plusieurs paramètres biochimiques d'intérêt durant les cultures cellulaires et virales pour la production de vaccins. Les cultures tendent à produire des anticorps ou à cultiver du virus selon le type de cellule mis en jeu, mais toujours en conservant des traceurs métaboliques similaires tels que la consommation du glucose ou la production de lactate. Toutefois, les techniques de mesure actuelles de ces paramètres biologiques nécessitent le prélèvement d'échantillons du milieu de culture pour appliquer des méthodes externes, ce qui prive donc l'analyse de toute rétroaction instantanée sur la culture en elle-même. En effet, bien que les conditions physiques de culture, tels que le pH, la température ou l'oxygénation du milieu soient mesurées et régulées en temps réel, le suivi en direct des concentrations métaboliques reste un challenge posé par les grandes instances sanitaires aux acteurs du milieu pharmaceutique.

Les travaux de recherche présentés au sein de cette thèse ont plusieurs objectifs majeurs afin d'améliorer les techniques de suivi existantes dans la littérature. Le premier consiste à paramétrer et optimiser le protocole d'acquisition des spectres Raman pour le suivi des cultures de cellules. Nous soulignons que les derniers travaux rapportés dans la littérature scientifique concernent essentiellement les cultures de cellules de type CHO. Dans notre cas, les différents travaux portent non seulement sur cette lignée cellulaire, mais aussi sur les types HeLa et Sf9, présentant certaines différences significatives dans le sens où ces cellules sont utilisées pour produire non pas des anticorps mais du virus. Trois protocoles de culture différents sont donc engagés dans ces travaux. Le système d'acquisition, commun aux trois protocoles, est composé d'une sonde à immersion plongée dans le bioréacteur de culture et reliée à un spectromètre. Les premiers travaux présentés définissent donc les conditions d'acquisition des spectres Raman pour chacune des lignées cellulaires, ainsi que les techniques de correction appliquées. Dans un premier temps, le matériel d'acquisition étant commun et défini (source, détecteur, etc.), le paramètre le plus critique pour la qualité des signaux spectraux a été optimisé. Il s'agit du temps d'acquisition des spectres Raman, composé d'une part du temps d'exposition au détecteur et d'autre part du nombre d'accumulations réalisées pour composer le spectre final. Nous avons défini le temps d'exposition à 30 s, ce qui permet de se placer entre 40 % et 80 % de la gamme du convertisseur analogique-numérique du détecteur CCD. Ensuite, un protocole original pour la détermination du nombre d'accumulations à réaliser a été développé afin d'obtenir des

temps d'acquisition finaux ni trop courts pour bénéficier d'une qualité spectrale satisfaisante (rapport signal sur bruit suffisamment élevé), ni trop long afin que les spectres Raman soient représentatifs de l'état biologique du milieu de culture à l'instant d'acquisition. Réalisés sur une culture unique pour chaque type de cellule, nous avons moyenné plusieurs spectres Raman de 30 s afin de générer des jeux de données composés de spectres de 1 min, 2 min, 5 min, 10 min, 16 min et 25 min. Ces différents jeux de données ont été évalués en réalisant une modélisation PLS pour chaque biomarqueur, soit les concentrations en glucose, glutamine, glutamate, lactate, ammonium, ainsi que sur les densités cellulaires VCD et TCD. En comparant les erreurs RMSEP obtenues pour chaque modèle, nous avons obtenu un temps optimal d'acquisition pour les spectres Raman, évalué à 5 min, soit 10×30 s, pour tous les bioprocédés engagés.

Suite à l'optimisation du temps d'acquisition total des spectres Raman, nous avons évalué l'influence du montage de la sonde Raman sur le bioréacteur de culture, paramètre critique tant la variabilité spectrale existante entre deux cultures cellulaires peut être accrue par un montage différent. Plusieurs points ont alors été étudiés, à commencer par le montage de la tête de sonde, reliant les deux parties principales de la sonde, soit la tête reliée au spectromètre, et le corps, partie immergée dans le bioréacteur et stérilisée pour chaque nouvelle culture. Nous avons montré que certaines variations spectrales apparaissaient lors du montage de la sonde, quelque soit l'opérateur, traduisant donc une influence du montage de la sonde sur les spectres. Néanmoins, en modifiant le prétraitement des données, soit une dérivée première de Savitzky-Golay suivi d'une normalisation SNV, puis une sélection des régions spectrales 3000–2800 cm^{-1} et 1775–350 cm^{-1}, nous sommes en mesure de réduire l'influence du montage de la tête de sonde. Nous avons également étudié l'influence de la profondeur d'immersion de la sonde Raman, démontrant l'existence d'effets certains sur les mesures Raman, notamment à cause de contributions dues aux parois du bioréacteur de culture. Toutefois, nous avons montré que le fait de travailler à une profondeur de travail variable, ne se rapprochant ni du fond du biogénérateur, ni de la surface du milieu de culture permettait de s'affranchir de la variabilité liée à l'immersion de la sonde. Le dernier point évalué par rapport au montage de la sonde sur le bioréacteur est la position de la fibre optique reliant la sonde au spectromètre Raman. Cette fois, nous avons montré que ce paramètre n'avait aucune influence sur les signaux après avoir appliqué le prétraitement spectral évoqué ci-dessus. Finalement, nous avons appliqué ce nouveau prétraitement à un ensemble d'échantillons provenant de deux cultures de cellules CHO différentes. Nous avons alors montré qu'en développant un modèle PLS uniquement basé sur la population d'une seule culture, il existait un certain biais de prédiction lorsque les échantillons de la seconde culture étaient employés en tant que jeu test. Ceci expose alors

l'existence d'une variabilité inter-culture à prendre en compte au sein des ensembles de calibration pour le développement de modèles de régression robustes.

Il s'agit là du deuxième axe de recherche employé durant ces travaux de recherche, soit la recherche de robustesse pour les modèles de régression. En effet, le développement de modèles de régression robustes est un point crucial pour l'essor de l'application des outils spectroscopiques pour la mesure des paramètres biochimiques. C'est pourquoi les derniers travaux de recherche présentés dans la littérature sur l'application de la spectroscopie Raman pour le suivi de cultures tendent vers la réalisation de modèles de régression robustes à différents changements de paramètre (modifications du bioprocédé). Néanmoins, la plupart de ces travaux portent uniquement sur les cellules CHO, les plus utilisées dans le milieu biopharmaceutique pour la production d'anticorps. Dans notre cas, pour le développement de modèles de régression robustes, nous avons d'une part employé les données acquises depuis les cultures de cellules CHO pour modéliser les paramètres biochimiques lors de la production d'anticorps, mais aussi d'autre part, des données provenant de cultures de cellules HeLa et Sf9 pour développer des modèles de régression efficaces pour les suivi métabolique durant les cultures virales. Dans un premier temps, nous avons accumulé, pour chaque type de cellule, les échantillons provenant de plusieurs cultures pour développer les modèles de régression les plus robustes possibles. Pour les modèles PLS développés sur les biomarqueurs des cultures de cellules CHO, nous avons finalement montré que nous étions en mesure de suivre de façon satisfaisante les tendances de tous les paramètres d'intérêt, hormis la concentration en ammonium. En effet, pour ce paramètre, le modèle de régression ne s'appuie pas sur les bonnes variables spectrales et ne permet donc pas de capter convenablement les tendances du paramètre biochimique. Pour le suivi des concentrations métaboliques des cultures de cellules HeLa, il apparait que la majorité des paramètres sont bien prédits. Toutefois, dans le cas des cultures virales, nous devons souligner la nécessité de distinguer les données provenant des différentes phases du bioprocédé (amplification cellulaire puis culture virale) dans la modélisation des densités cellulaires. Ainsi, nous avons pu montrer que le suivi des densités VCD et TCD était très bien réalisé lors de la phase de croissance cellulaire, mais que l'infection virale empêchait de développer des modèles pleinement satisfaisants. Ceci est directement causé par les méthodes de référence pour le comptage cellulaire, moins fiables durant cette deuxième phase. Enfin, pour les cellules Sf9, également employées pour cultiver du virus, nous avons montré que les suivis métaboliques s'avèrent plus difficiles. En effet, les gammes de concentration présentent très peu de variations, notamment en ce qui concerne le glutamate, ce qui rend donc la modélisation plus périlleuse puisque les signaux Raman présentent peu de variation. De plus, nous ne pouvons pas présenter de modèle de

régression pour la concentration en lactate puisque les niveaux de concentration sont inférieurs à la limite de quantification. Néanmoins, les modèles de régression développés pour les suivis des densités cellulaires sont acceptables tant ils parviennent à démarquer d'une part les modèles pour les cellules vivantes (VCD) et d'autre part les modèles pour la densité totale en cellules (TCD).

Après avoir développé les modèles de régression basés sur les échantillons provenant de plusieurs cultures cellulaires, la robustesse des modèles ne peut plus être améliorée en accumulant encore plus de données similaires. En effet, nous risquerions de surentrainer les modèles de régression qui perdraient donc en capacités prédictives par rapport à certaines cultures légèrement différentes. Nous avons donc étudié les variations de paramètres physiques pouvant influencer les gammes de concentration des cultures cellulaires. Ces travaux, réalisés sur les cultures de cellules CHO, visent à modifier les conditions de culture afin d'entrainer le bioprocédé dans des situations inédites lors des lots précédents et donc agrandir les gammes de concentration des différents biomarqueurs. De plus, cela nous a permis d'évaluer l'impact des variations de paramètres sur la mesure Raman en elle-même. Nous avons ainsi étudié l'influence de variations de pH, de pO_2, de température et de vitesse d'agitation du milieu. Nous avons alors observé des variations des paramètres métaboliques et donc des changements sur les gammes de concentration. Ainsi, nous avons ajouté plusieurs cultures acquises dans des conditions particulières à l'ensemble d'apprentissage, afin d'accroitre les gammes de concentration et donc ajouter au jeu de calibration plus de variabilité. Nous en avons tiré des modèles PLS plus performants que ceux développés sur la seule base d'échantillons provenant de cultures cellulaires réalisées uniquement dans les conditions normales de culture. Nous avons ainsi accru la robustesse des modèles de régression développés pour le suivi des biomarqueurs des cultures de cellules CHO.

Par la suite, nous avons développé des modèles de régression PLS généraux pour la prédiction des concentrations métaboliques et des densités cellulaires de plusieurs types de cellule, en l'occurrence CHO, HeLa et Sf9. Toutefois, les différences métaboliques entre les bioprocédés mis en jeu sont de réels obstacles au développement de modèles prenant en compte les variabilités qui existent entre les différents types de culture. De ce fait, seuls les modèles développés pour les concentrations en glucose et en lactate présentent de réels potentiels prédictifs puisque les gammes de concentration de ces produits sont homogènes d'une lignée à l'autre. Ces deux modèles ont été évalués à l'aide d'un quatrième type de cellule (HEK) dont les échantillons ont été utilisés pour éprouver les différents modèles de régression. Nous avons alors montré qu'ils étaient capables de prédire les tendances métaboliques du nouveau type cellulaire sans que l'information ne soit introduite au sein de l'ensemble d'apprentissage. En incluant des échantillons provenant des cultures de cellules

HEK au jeu de calibration, nous avons obtenu des résultats très satisfaisants permettant de montrer qu'il est possible de développer des modèles de régression uniques pour le suivi métabolique des cultures cellulaires de plusieurs types de bioprocédé. Ceci est un point important, puisque pour trois types de cellule (voire quatre en comptant HEK), nous passons de trois modèles PLS à un seul, ce qui réduit de façon importante les données nécessaires à la prédiction en ligne de toute nouvelle culture, simplifiant ainsi le développement d'outils de suivi spectroscopique en temps réel des paramètres biochimiques des cultures cellulaires de différentes lignées.

Les derniers travaux de recherche présentés au cours de cette thèse sont focalisés sur la valorisation des données biologiques et spectrales mises à disposition pour le développement des modèles de régression précédents. Dans un premier temps, nous avons travaillé sur la modélisation de la concentration en anticorps des cultures de cellules CHO. Nous avons alors pu développer les premiers modèles de prédictions du titre en anticorps basés sur des spectres acquis au sein même du bioréacteur, démontrant donc la faisabilité du suivi en ligne et en temps réel de la concentration en anticorps durant les cultures de cellules. Nous avons ensuite exporté ce concept au suivi en temps réel du titre en virus infectieux des cultures virales, prenant pour support les cellules Sf9. Toutefois, malgré la mobilisation d'outils de calculs plus élaborés que pour le développement des modèles PLS précédents, nous n'avons pas été en mesure de générer de modèle prédictif satisfaisant. La prédiction de la concentration en virus requiert certainement l'emploi d'autres techniques de modélisation. Enfin, nous avons également employé les spectres Raman acquis en continu durant les cultures de cellules pour développer des outils de contrôle statistique des procédés à partir de données multivariées (MSPC). Nous avons alors construit des cartes de contrôle permettant de déterminer toutes les heures si le système biologique contrôlé présente une déviation par rapport à un état nominal de référence. Toutefois, il convient de noter ici que ces travaux ne sont qu'un premier pas vers le suivi statistique de bioprocédés dans le sens où les données recueillies tout au long de ce projet de recherche ne s'inscrivent pas directement dans l'optique de telles applications. En effet, la nécessité de disposer de cultures contaminées ou atypiques pour l'élaboration et la validation des cartes de contrôle multivariées est un frein pour la réalisation des modèles MPCA au cours de ce projet, davantage porté sur le suivi de biomarqueurs des cultures. De ce fait, l'ajout de cultures présentant des variations d'évolution des cultures cellulaires pourrait notablement enrichir l'ensemble d'apprentissage en vue du déploiement du MSPC sur ces bioprocédés.

Les travaux présentés au sein de cette thèse exposent plusieurs avancées dans le domaine du suivi spectroscopique de cultures cellulaires, d'une part proposant des techniques d'optimisation pour l'installation et le paramétrage du système d'acquisition et

d'autre part présentant les modèles de régression robustes pour la détermination des concentrations métaboliques et les densités cellulaires. Mais il faut noter que ces travaux peuvent encore être améliorés de différentes manières. En effet, à ce jour, les techniques de sélection de variables ont démontré leur plein potentiel pour amplifier les capacités prédictives des modèles de régression, notamment en supprimant des régions spectrales inutiles ou présentant peu d'intérêt lors de la modélisation chimiométrique. La richesse de la littérature sur le sujet, notamment grâce aux différents articles et revues réalisés, permet d'avoir accès à de nombreuses notions en la matière. De plus la diversité d'applications sur des données spectroscopiques variées permet de rapidement évaluer les différentes techniques pour retrouver celles qui sont les plus appropriées à nos jeux de données. Ainsi, les modèles PLS présentés jusqu'ici pourraient atteindre des niveaux d'erreur plus faibles encore en appliquant les sélections adéquates. Ceci aurait alors pour but final d'accroitre la robustesse des modèles de régression développés pour le suivi en ligne des concentrations métaboliques et des densités cellulaires.

Toutefois, il faut rappeler que l'emploi de tels algorithmes de calcul ne permettrait que d'améliorer à la marge les modèles de régression déjà établis. En effet, les techniques de sélection de variables ne sauraient palier à un manque de corrélation entre les spectres Raman acquis et les mesures de références associées, principalement dus à des niveaux de concentration extrêmement faibles au sein des cultures. De ce fait, il pourrait être intéressant, notamment pour les lignées cellulaires peu consommatrices ou productrices telles que les cellules Sf9, d'enrichir artificiellement l'ensemble des gammes de concentration à travers l'ajout de produits purs au sein des bioréacteurs de culture. Cette démarche, également appelée *spiking*, pourrait alors permettre d'accéder à des niveaux jamais rencontrés par les modèles de régression. De plus, cela pourrait permettre de débloquer certains verrous rencontrés lors de modélisations chimiométriques n'ayant pas aboutis, notamment en ce qui concerne le titre en virus infectieux.

Enfin, une perspective audacieuse d'un point de vue analytique pourrait être le couplage de différentes techniques spectroscopiques *in situ* pour l'amélioration des prédictions de niveaux métaboliques. Dans la littérature, certaines comparaisons existent et portent principalement sur les avantages et inconvénients de chaque méthode. Il serait donc très intéressant de coupler des techniques telles que les spectroscopies NIR et MIR aux mesures Raman afin de réellement optimiser les modèles de régression générés. Néanmoins, au-delà du couplage des données qui pourraient être recueillies, c'est surtout la mise en place des différentes sondes à immersion sur un seul et même bioréacteur de culture qui pourrait être la réelle problématique de ce type de travail.

Références bibliographiques

[1] Fish, R. C.; Smith, B.-H.; Sze, T. T.; Rice, K. A.; Biotechnology inspection guide reference materials and training aids. *U.S. Food and Drug Administration*. [en ligne] consulté le 30 septembre 2016. http://www.fda.gov/ICECI/Inspections/InspectionGuides/ucm074181.htm#CELL_CULTURE_AND.

[2] Définition statistique de la biotechnologie. *Organisation de coopération et de développement économiques (OCDE)*. [en ligne] consulté le 30 septembre 2016. http://www.oecd.org/fr/sti/biotech/definitionstatistiquedelabiotechnologiemiseajouren2005.htm.

[3] Cristofari, J.-J.; Vaccins, l'Europe mène la bal. *Pharmaceutiques*. Juin/Juillet **2007**, 54-58.

[4] Nolan, R. P.; Lee, K.; Dynamic model of CHO cell metabolism. *Metabolic Engineering* **2011**, *13*, 108-124.

[5] Ghorbaniaghdam, A.; Henry, O.; Jolicoeur, M.; An in-silico study of the regulation of CHO cells glycolysis. *Journal of Theoretical Biology* **2014**, *357*, 112-122.

[6] Kay, J.; Weitzman, P. D. J. *Krebs' citric acid cycle: half a century and still turning*; The Biochemical Society: London, 1987.

[7] Krebs, H. A.; Bellamy, D.; The interconversion of glutamic acid and aspartic acid in respiring tissues. *Biochemical Journal* **1960**, *75*, 523-529.

[8] McKeehan, W. L.; Glycolysis, glutaminolysis and cell-proliferation. *Cell Biology International Reports* **1982**, *6*, 635-650.

[9] Callaway, E.; Deal done over HeLa cell line. *Nature* **2013**, *500*, 132-133.

[10] Stanisz, J.; Wice, B. M.; Kennell, D. E.; Comparative energy metabolism in cultured heart muscle and HeLa cells. *Journal of Cellular Physiology* **1983**, *115*, 320-330.

[11] Reitzer, L. J.; Wice, B. M.; Kennell, D.; Evidence that glutamine, not sugar, is the major energy-source for cultured HeLa cells. *Journal of Biological Chemistry* **1979**, *254*, 2669-2676.

[12] Drews, M.; Doverskog, M.; Ohman, L.; Chapman, B. E.; Jacobsson, U.; Kuchel, P. W.; Haggstrom, L.; Pathways of glutamine metabolism in Spodoptera frugiperda (Sf9) insect cells: evidence for the presence of the nitrogen assimilation system, and a metabolic switch by H-1/N-15 NMR. *Journal of Biotechnology* **2000**, *78*, 23-37.

[13] Bernal, V.; Carinhas, N.; Yokomizo, A. Y.; Carrondo, M. J. T.; Alves, P. M.; Cell density effect in the baculovirus-insect cells system: a quantitative analysis of energetic metabolism. *Biotechnology and Bioengineering* **2009**, *104*, 162-180.

[14] Graham, F. L.; Smiley, J.; Russell, W. C.; Nairn, R.; Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *Journal of General Virology* **1977**, *36*, 59-72.

[15] Petiot, E.; Cuperlovic-Culf, M.; Shen, C. F.; Kamen, A.; Influence of HEK293 metabolism on the production of viral vectors and vaccine. *Vaccine* **2015**, *33*, 5974-5981.

[16] Martínez, V. S.; Dietmair, S.; Quek, L. E.; Hodson, M. P.; Gray, P.; Nielsen, L. K.; Flux balance analysis of CHO cells before and after a metabolic switch from lactate production to consumption. *Biotechnology and Bioengineering* **2013**, *110*, 660-666.

[17] Furukawa, K.; Ohsuye, K.; Effect of culture temperature on a recombinant CHO cell line producing a C-terminal alpha-amidating enzyme. *Cytotechnology* **1998**, *26*, 153-164.

[18] Rao, P. N.; Engelber, J.; HeLa cells: effects of temperature on life cycle. *Science* **1965**, *148*, 1092-1094.

[19] Watanabe, I.; Okada, S.; Effects of temperature on growth rate of cultured mammalian cells (L5178Y). *Journal of Cell Biology* **1967**, *32*, 309-323.

[20] Ceccarin, C.; Eagle, H.; pH as a determinant of cellular growth and contact inhibition. *Proceedings of the National Academy of Sciences of the United States of America* **1971**, *68*, 229-233.

[21] Borys, M. C.; Linzer, D. I. H.; Papoutsakis, E. T.; Culture pH affects expression rates and glycosylation of recombinant mouse placental-lactogen proteins by Chinese-Hamster Ovary (CHO) cells. *Bio-Technology* **1993**, *11*, 720-724.

[22] Ozturk, S. S.; Palsson, B. O.; Growth, metabolic, and antibody-production kinetics of hybridoma cell-culture: 2. effects of serum concentration, dissolved-oxygen concentration, and medium pH in a batch reactor. *Biotechnology Progress* **1991**, *7*, 481-494.

[23] Lin, A. A.; Kimura, R.; Miller, W. M.; Production of tPA in recombinant CHO cells under oxygen-limited conditions. *Biotechnology and Bioengineering* **1993**, *42*, 339-350.

[24] Truesdale, G. A.; Downing, A. L.; Solubility of oxygen in water. *Nature* **1954**, *173*, 1236-1236.

[25] Garcia-Ochoa, F.; Gomez, E.; Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnology Advances* **2009**, *27*, 153-176.

[26] Kramers, H.; Baars, G. M.; Knoll, W. H.; A comparative study on the rate of mixing in stirred tanks. *Chemical Engineering Science* **1953**, *2*, 35-42.

[27] Rushton, J. H.; Costich, E. W.; Everett, H. J.; Power characteristics of mixing impellers. *Chemical Engineering Progress* **1950**, *46*, 395-404.

[28] Accesoires pour bioréacteurs. *Sartorius*. [en ligne] consulté le 30 septembre 2016. https://www.sartorius-france.fr/fr/produits/bioreacteurs-fermenteurs/accessoires/.

[29] Bioprofile FLEX Analyzer. *Nova Biomedical*. [en ligne] consulté le 30 septembre 2016. http://www.novabiomedical.com/responsive/bioprof_flex.php.

[30] RX Daytona^{+}. *Randox Laboratories Ltd.* [en ligne] consulté le 30 septembre 2016. http://www.randox.com/rx-daytona-plus/.

[31] Cedex Bio Analyzer. *CustomBiotech, Roche*. [en ligne] consulté le 30 septembre 2016. http://custombiotech.roche.com/home/Product_Details/3_8_10_2_1_1.html.

[32] Ammoniaque. *R-Biopharm*. [en ligne] consulté le 30 septembre 2016. http://www.r-biopharm.com/fr/produits/diagnostic-alimentaire/constituants/coffrets-enzymatique/gamme-roche-boehringer/item/ammoniaque.

[33] 2300 STAT Plus™ Glucose & Lactate Analyzer. *YSI*. [en ligne] consulté le 30 septembre 2016. https://www.ysi.com/ysi-2300-stat-plus-glucose-lactate-analyzer.

[34] NuceloCounter® SCC-100™. *Chemometec*. [en ligne] consulté le 30 septembre 2016. http://chemometec.com/cell-counters/somatic-cell-counter-scc-100-nucleocounter/.

[35] Vi-CELL® XR. *Beckman Coulter*. [en ligne] consulté le 30 septembre 2016. http://www.beckmancoulter.fr/Sciences+de+la+vie/Cytom%C3%A9trie+en+flux+_+Analyseurs+de+cellules/Analyseurs+de+cellules/Vi_CELL+XR++.html.

[36] Brown, M.; Wittwer, C.; Flow cytometry: Principles and clinical applications in hematology. *Clinical Chemistry* **2000**, *46*, 1221-1229.

[37] Strober, W.; Trypan blue exclusion test of cell viability. *Current protocols in immunology* **2015**, *111*.

[38] ACQUITY UPLC H-Class Bio *Waters*. [en ligne] consulté le 30 septembre 2016. http://www.waters.com/waters/en_GB/ACQUITY-UPLC-H-Class-Bio/nav.htm?cid=10166246&locale=en_GB.

[39] Van Deemter, J. J.; Zuiderweg, F. J.; Klinkenberg, A.; Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. *Chemical Engineering Science* **1956**, *5*, 271-289.

[40] Engvall, E.; Perlmann, P.; Enzyme-linked immunosorbent assay (ELISA) Quantitative assay of immunoglobulin G. *Immunochemistry* **1971**, *8*, 871-874.

[41] Wide, L.; Porath, J.; Radioimmunoassay of proteins with the use of Sephadex-coupled antibodies. *Biochimica Et Biophysica Acta* **1966**, *130*, 257-260.

[42] Rathore, A. S.; Bhambure, R.; Ghare, V.; Process analytical technology (PAT) for biopharmaceutical products. *Analytical and Bioanalytical Chemistry* **2010**, *398*, 137-154.

[43] Hinz, D. C.; Process analytical technologies in the pharmaceutical industry: the FDA's PAT initiative. *Analytical and Bioanalytical Chemistry* **2006**, *384*, 1036-1042.

[44] U.S. Department of Health and Human Services; Food and Drug Administration, **2004**, *Guidance for industry: PAT – A framework for innovative pharmaceutical development, manufacturing and quality asurance*; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidance/ucm070305.pdf, consulté le 30 septembre 2016.

[45] International Council for Harmonisation, **2009**, *Pharmaceutical Development Q8(R2)*; http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q8_R1/Step4/Q8_R2_Guideline.pdf, consulté le 30 septembre 2016.

[46] Râman, C. V.; Krishnan, K. S.; A new type of secondary radiation. *Nature* **1928**, *121*, 501-502.

[47] Landsberg, G.; Mandelstam, L.; Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen. *Naturwissenschaften* **1928**, *16*, 557-558.

[48] Râman, C. V.; Krishnan, K. S.; A new class of spectra due to secondary radiation, Part I. *Indian Journal of Physics* **1928**, *16*, 399-419.

[49] Râman, C. V.; A new radiation. *Indian Journal of Physics* **1928**, *2*, 387-398.

[50] Javan, A.; Herriott, D. R.; Bennett, W. R.; Population inversion and continuous optical maser oscillation in a gas discharge containing a He-Ne mixture. *Physical Review Letters* **1961**, *6*, 106-113.

[51] Geusic, J. E.; Marcos, H. M.; Vanuitert, L. G.; Laser oscillations in Nd-doped yttrium aluminium, yttrium gallium and godalium garnets (continuous operation of $Y_3Al_5O_{12}$; pulsed operation of $Y_3Ga_5O_{12}$ and $Gd_3Ga_5O_{12}$; rm. temp; E). *Applied Physics Letters* **1964**, *4*, 182-184.

[52] Moulton, P. F.; Spectroscopic and laser characteristics of Ti:Al_2O_3. *Journal of the Optical Society of America B-Optical Physics* **1986**, *3*, 125-133.

[53] Boyle, W.; Smith, G.; *Buried channel charge coupled devices*. US 3792322 A, **1974**.

[54] Boyle, W.; Smith, G.; *Three dimensional charge coupled devices*. US 3796927 A, **1974**.

[55] Stone, J.; Walrafen, G.; *Method utilizing an optical fiber raman cell*. US 3770305 A, **1973**.

[56] McLachlan, R. D.; Jewett, G. L.; Evans, J. C.; *Fiber-optic probe for sensitive Raman analysis*. US 4573761 A, **1986**.

[57] Bowen, J. M.; Sullivan, P. J.; Sterling Blanche, M.; Essington, M.; Noe, L. J.; *Optical-fiber raman spectroscopy used for remote in-situ environmental analysis*. US 4802761 A, **1989**.

[58] Carrabba, M. M.; Rauh, R. D.; *Apparatus for measuring Raman spectra over optical fibers*. US 5112127 A, **1992**.

[59] Nave, S. E.; Livingston, R. R.; Prather, W. S.; *Fiber optic probe for light scattering measurements*. US 5404218 A, **1995**.

[60] Yanan, J.; *Raman probe with spatial filter and semi-confocal lens*. US 6310686 B1, **2001**.

[61] Berger, A. J.; Brennan III, J. F.; Dasari, R. R.; Feld, M. S.; Itzkan, I.; Tanaka, K.; Wang, Y.; *Apparatus and methods of raman spectroscopy for analysis of blood gases and analytes*. US 5615673 A, **1997**.

[62] Doyle, W. M.; *Raman probe having a small diameter immersion tip*. US 6876801 B2, **2005**.

[63] Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier Science B.V.: Amsterdam, 1997.

[64] Oldroyd, D. *The Arch of Knowledge*; Methuen: New York, 1986.

[65] Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Computerized learning machines applied to chemical problems - Molecular formula determination from low resolution mass spectrometry. *Analytical Chemistry* **1969**, *41*, 21-27.

[66] Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Reilley, C. N.; Computerized learning machines applied to chemical problems - Investigation of convergence rate and predictive ability of adaptive binary pattern classifiers. *Analytical Chemistry* **1969**, *41*, 690-695.

[67] Kowalski, B. R.; Jurs, P. C.; Isenhour, T. L.; Reilley, C. N.; Computerized learning machines applied to chemical problems - Multicategory pattern classification by least squares. *Analytical Chemistry* **1969**, *41*, 695-700.

[68] Kowalski, B. R.; Jurs, P. C.; Isenhour, T. L.; Reilley, C. N.; Computerized learning machines applied to chemical problems - Interpretation of infrared spectrometry data. *Analytical Chemistry* **1969**, *41*, 1945-1949.

[69] Wold, S.; Esbensen, K.; Geladi, P.; Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37-52.

[70] Pearson, K.; On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **1901**, *2*, 559-572.

[71] Wold, H. In *Multivariate Analysis*, Krishnaiah, P. R., Ed.; Academic Press, 1966.

[72] Hotelling, H.; Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **1933**, *24*, 417-441.

[73] Ramos, L. S.; Beebe, K. R.; Carey, W. P.; Sanchez, E.; Erickson, B. C.; Wilson, B. E.; Wangen, L. E.; Kowalski, B. R.; Chemometrics. *Analytical Chemistry* **1986**, *58*, R294-R315.

[74] Wold, H.; Causal flows with latent variables: Partings of ways in light of NIPALS modelling. *European Economic Review* **1974**, *5*, 67-86.

[75] Wold, H. In *Evaluation of econometric models*, Kmenta, J.; Ramsey, J. B., Eds.; Academic Press, 1980; pp 47-74.

[76] Wold, S.; Martens, H.; Wold, H.; The multivariate calibration-problem in chemistry solved by the PLS method. *Lecture Notes in Mathematics* **1983**, *973*, 286-293.

[77] de Jong, S.; SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **1993**, *18*, 251-263.

[78] Cozzolino, D. In *Mathematical and Statistical Methods in Food Science and Technology*, Granato, D.; Ares, G., Eds.; John Wiley & Sons, Ltd: New York, 2014; pp 19-30.

[79] de Noord, O. E.; The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems* **1994**, *23*, 65-70.

[80] Rinnan, A.; van den Berg, F.; Engelsen, S. B.; Review of the most common pre-processing techniques for near-infrared spectra. *Trac-Trends in Analytical Chemistry* **2009**, *28*, 1201-1222.

[81] Devos, O.; Duponchel, L.; Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. *Chemometrics and Intelligent Laboratory Systems* **2011**, *107*, 50-58.

[82] Norris, K. H.; Williams, P. C.; Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particule size. *Cereal Chemistry* **1984**, *61*, 158-165.

[83] Savitzky, A.; Golay, M. J. E.; Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **1964**, *36*, 1627-1639.

[84] Alberti, J. C.; Phillips, J. A.; Jink, D. J.; Wacasz, F. M.; Off-line monitoring of fermentation samples by FTIR/ATR: a feasibility study for real-time process control. In *Biotechnology and Bioengineering Symposium*; John Wiley and Sons, Inc., **1985**, *15*, 689-722.

[85] Cavinato, A. G.; Mayes, D. M.; Ge, Z. H.; Callis, J. B.; Noninvasive method for monitoring ethanol in fermentation processes using fiber-optic near-infrared spectroscopy. *Analytical Chemistry* **1990**, *62*, 1977-1982.

[86] Ge, Z. H.; Cavinato, A. G.; Callis, J. B.; Noninvasive spectroscopy for monitoring cell density in a fermentation process. *Analytical Chemistry* **1994**, *66*, 1354-1362.

[87] Vaccari, G.; Dosi, E.; Campi, A. L.; Gonzalezvara, A.; Matteuzzi, D.; Mantovani, G.; A near-infrared spectroscopy technique for the control of fermentation processes: an application to lactic acid fermentation. *Biotechnology and Bioengineering* **1994**, *43*, 913-917.

[88] Yamane, Y.; Mikami, T.; Higashida, K.; Kakizono, T.; Nishio, N.; Estimation of the concentrations of cells, astaxanthin and glucose in a culture of Phaffia rhodozyma by near infrared reflectance spectroscopy. *Biotechnology Techniques* **1996**, *10*, 529-534.

[89] Riley, M. R.; Rhiel, M.; Zhou, X. J.; Arnold, M. A.; Murhammer, D. W.; Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy. *Biotechnology and Bioengineering* **1997**, *55*, 11-15.

[90] Yeung, K. S. Y.; Hoare, M.; Thornhill, N. F.; Williams, T.; Vaghjiani, J. D.; Near-infrared spectroscopy for bioprocess monitoring and control. *Biotechnology and Bioengineering* **1999**, *63*, 684-693.

[91] Arnold, S. A.; Crowley, J.; Woods, N.; Harvey, L. M.; McNeill, B.; In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation. *Biotechnology and Bioengineering* **2003**, *84*, 13-19.

[92] Tosi, S.; Rossi, M.; Tamburini, E.; Vaccari, G.; Amaretti, A.; Matteuzzi, D.; Assessment of in-line near-infrared spectroscopy for continuous monitoring of fermentation processes. *Biotechnology Progress* **2003**, *19*, 1816-1821.

[93] Tamburini, E.; Vaccari, G.; Tosi, S.; Trilli, A.; Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe. *Applied Spectroscopy* **2003**, *57*, 132-138.

[94] Navratil, M.; Norberg, A.; Lembren, L.; Mandenius, C. F.; On-line multi-analyzer monitoring of biomass, glucose and acetate for growth rate control of a Vibrio cholerae fed-batch cultivation. *Journal of Biotechnology* **2005**, *115*, 67-79.

[95] Roychoudhury, P.; O'Kennedy, R.; McNeil, B.; Harvey, L. M.; Multiplexing fibre optic near infrared (NIR) spectroscopy as an emerging technology to monitor industrial bioprocesses. *Analytica Chimica Acta* **2007**, *590*, 110-117.

[96] Nordon, A.; Littlejohn, D.; Dann, A. S.; Jeffkins, P. A.; Richardson, M. D.; Stimpson, S. L.; In situ monitoring of the seed stage of a fermentation process using non-invasive NIR spectrometry. *Analyst* **2008**, *133*, 660-666.

[97] Li, H. Q.; Chen, H. Z.; Near-infrared spectroscopy with a fiber-optic probe for state variables determination in solid-state fermentation. *Process Biochemistry* **2008**, *43*, 511-516.

[98] Petersen, N.; Odman, P.; Padrell, A. E. C.; Stocks, S.; Lantz, A. E.; Gernaey, K. V.; In situ near infrared spectroscopy for analyte-specific monitoring of glucose and ammonium in Streptomyces coelicolor fermentations. *Biotechnology Progress* **2010**, *26*, 263-271.

[99] Kim, J. Y.; Kim, Y. G.; Lee, G. M.; CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Applied Microbiology and Biotechnology* **2012**, *93*, 917-930.

[100] Clavaud, M.; Roggo, Y.; Von Daeniken, R.; Liebler, A.; Schwabe, J. O.; Chemometrics and in-line near infrared spectroscopic monitoring of a biopharmaceutical Chinese hamster ovary cell culture: Prediction of multiple cultivation variables. *Talanta* **2013**, *111*, 28-38.

[101] Kozma, B.; Parta, L.; Zalai, D.; Gergely, S.; Salgo, A.; A model system and chemometrics to develop near infrared spectroscopic monitoring for Chinese hamster ovary cell cultivations. *Journal of near Infrared Spectroscopy* **2014**, *22*, 401-410.

[102] Milligan, M.; Lewin-Koh, N.; Coleman, D.; Arroyo, A.; Saucedo, V.; Semisynthetic model calibration for monitoring glucose in mammalian cell culture with in situ near infrared spectroscopy. *Biotechnology and Bioengineering* **2014**, *111*, 896-903.

[103] Sivakesava, S.; Irudayaraj, J.; Ali, D.; Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques. *Process Biochemistry* **2001**, *37*, 371-378.

[104] Fayolle, P.; Picque, D.; Corrieu, G.; On-line monitoring of fermentation processes by a new remote dispersive middle-infrared spectrometer. *Food Control* **2000**, *11*, 291-296.

[105] Acha, V.; Meurens, M.; Naveau, H.; Agathos, S. N.; ATR-FTIR sensor development for continuous on-line monitoring of chlorinated aliphatic hydrocarbons in a fixed-bed bioreactor. *Biotechnology and Bioengineering* **2000**, *68*, 473-487.

[106] Rhiel, M.; Ducommun, P.; Bolzonella, I.; Marison, I.; von Stockar, U.; Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements. *Biotechnology and Bioengineering* **2002**, *77*, 174-185.

[107] Rhiel, M. H.; Amrhein, M. I.; Marison, I. W.; von Stockar, U.; The influence of correlated calibration samples on the prediction performance of multivariate models based on mid-infrared spectra of animal cell cultures. *Analytical Chemistry* **2002**, *74*, 5227-5236.

[108] Kornmann, H.; Rhiel, M.; Cannizzaro, C.; Marison, I.; von Stockar, U.; Methodology for real-time, multianalyte monitoring of fermentations using an in-situ mid-infrared sensor. *Biotechnology and Bioengineering* **2003**, *82*, 702-709.

[109] Kornmann, H.; Valentinotti, S.; Duboc, P.; Marison, I.; von Stockar, U.; Monitoring and control of Gluconacetobacter xylinus fed-batch cultures using in situ mid-IR spectroscopy. *Journal of Biotechnology* **2004**, *113*, 231-245.

[110] Mazarevica, G.; Diewok, J.; Baena, J. R.; Rosenberg, E.; Lendl, B.; On-line fermentation monitoring by mid-infrared spectroscopy. *Applied Spectroscopy* **2004**, *58*, 804-810.

[111] Schenk, J.; Marison, I. W.; von Stockar, U.; A simple method to monitor and control methanol feeding of Pichia pastoris fermentations using mid-IR spectroscopy. *Journal of Biotechnology* **2007**, *128*, 344-353.

[112] Veale, E. L.; Irudayaraj, J.; Demirci, A.; An on-line approach to monitor ethanol fermentation using FTIR spectroscopy. *Biotechnology Progress* **2007**, *23*, 494-500.

[113] Trevisan, M. G.; Poppi, R. J.; Direct determination of ephedrine intermediate in a biotransformation reaction using infrared spectroscopy and PLS. *Talanta* **2008**, *75*, 1021-1027.

[114] Sandor, M.; Rudinger, F.; Bienert, R.; Grimm, C.; Solle, D.; Scheper, T.; Comparative study of non-invasive monitoring via infrared spectroscopy for mammalian cell cultivations. *Journal of Biotechnology* **2013**, *168*, 636-645.

[115] Lindemann, C.; Marose, S.; Nielsen, H. O.; Scheper, T.; 2-dimensional fluorescence spectroscopy for on-line bioprocess monitoring. *Sensors and Actuators B-Chemical* **1998**, *51*, 273-277.

[116] Marose, S.; Lindemann, C.; Scheper, T.; Two-dimensional fluorescence spectroscopy: A new tool for on-line bioprocess monitoring. *Biotechnology Progress* **1998**, *14*, 63-74.

[117] Boehl, D.; Solle, D.; Hitzmann, B.; Scheper, T.; Chemometric modelling with two-dimensional fluorescence data for Claviceps purpurea bioprocess characterization. *Journal of Biotechnology* **2003**, *105*, 179-188.

[118] Haack, M. B.; Eliasson, A.; Olsson, L.; On-line cell mass monitoring of Saccharomyces cerevisiae cultivations by multi-wavelength fluorescence. *Journal of Biotechnology* **2004**, *114*, 199-208.

[119] Franz, C.; Jurgen, K.; Florentina, P.; Karl, B.; Sensor combination and chemometric modelling for improved process monitoring in recombinant E-coli fed-batch cultivations. *Journal of Biotechnology* **2005**, *120*, 183-196.

[120] Hantelmann, K.; Kollecker, A.; Hull, D.; Hitzmann, B.; Scheper, T.; Two-dimensional fluorescence spectroscopy: A novel approach for controlling fed-batch cultivations. *Journal of Biotechnology* **2006**, *121*, 410-417.

[121] Surribas, A.; Geissler, D.; Gierse, A.; Scheper, T.; Hitzmann, B.; Montesinos, J. L.; Valero, F.; State variables monitoring by in situ multi-wavelength fluorescence spectroscopy in heterologous protein production by Pichia pastoris. *Journal of Biotechnology* **2006**, *124*, 412-419.

[122] Rhee, J. I.; Kang, T. H.; On-line process monitoring and chemometric modeling with 2D fluorescence spectra obtained in recombinant E. coli fermentations. *Process Biochemistry* **2007**, *42*, 1124-1134.

[123] Ödman, P.; Johansen, C. L.; Olsson, L.; Gernaey, K. V.; Lantz, A. E.; On-line estimation of biomass, glucose and ethanol in Saccharomyces cerevisiae cultivations using in-situ multi-wavelength fluorescence and software sensors. *Journal of Biotechnology* **2009**, *144*, 102-112.

[124] Shope, T. B.; Vickers, T. J.; Mann, C. K.; The direct analysis of fermentation products by Raman spectroscopy. *Applied Spectroscopy* **1987**, *41*, 908-912.

[125] Gomy, C.; Jouan, M.; Dao, N. Q.; Observation of an alcoholic fermentation process by Raman-laser spectrometry and Raman-laser spectrometry using fiber optics (RLFO). *Comptes rendus de l'Academie des sciences, Série II* **1988**, *306*, 417-422.

[126] Gomy, C.; Jouan, M.; Dao, N. Q.; A quantitative laser-Raman spectrometric method with fiber optics for monitoring an alcoholic fermentation. *Analytica Chimica Acta* **1988**, *215*, 211-221.

[127] Spiegelman, C. H.; McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Cote, G. L.; Theoretical justification of wavelength selection in PLS calibration development of a new algorithm. *Analytical Chemistry* **1998**, *70*, 35-44.

[128] Shaw, A. D.; Kaderbhai, N.; Jones, A.; Woodward, A. M.; Goodacre, R.; Rowland, J. J.; Kell, D. B.; Noninvasive, on-line monitoring of the biotransformation by yeast of glucose to ethanol using dispersive Raman spectroscopy and chemometrics. *Applied Spectroscopy* **1999**, *53*, 1419-1428.

[129] Cannizzaro, C.; Rhiel, M.; Marison, I.; von Stockar, U.; On-line monitoring of Phaffia rhodozyma fed-batch process with in situ dispersive Raman spectroscopy. *Biotechnology and Bioengineering* **2003**, *83*, 668-680.

[130] Lee, H. L. T.; Boccazzi, P.; Gorret, N.; Ram, R. J.; Sinskey, A. J.; In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy. *Vibrational Spectroscopy* **2004**, *35*, 131-137.

[131] Arnold, A. S.; Wilson, J. S.; Boshier, M. G.; A simple extended-cavity diode laser. *Review of Scientific Instruments* **1998**, *69*, 1236-1239.

[132] Abu-Absi, N. R.; Kenty, B. M.; Cuellar, M. E.; Borys, M. C.; Sakhamuri, S.; Strachan, D. J.; Hausladen, M. C.; Li, Z. J.; Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. *Biotechnology and Bioengineering* **2011**, *108*, 1215-1221.

[133] Moretto, J.; Smelko, J. P.; Cuellar, M.; Berry, B.; Doane, A.; Ryll, T.; Wiltberger, K.; Process Raman spectroscopy for in-line CHO cell culture monitoring. *American Pharmaceutical Review* **2011**, *14*, 18-25.

[134] Whelan, J.; Craven, S.; Glennon, B.; In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnology Progress* **2012**, *28*, 1355-1362.

[135] Berry, B.; Moretto, J.; Matthews, T.; Smelko, J.; Wiltberger, K.; Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis. *Biotechnology Progress* **2015**, *31*, 566-577.

[136] Mehdizadeh, H.; Lauri, D.; Karry, K. M.; Moshgbar, M.; Procopio-Melino, R.; Drapeau, D.; Generic Raman-based calibration models enabling real-time monitoring of cell culture bioreactors. *Biotechnology Progress* **2015**, *31*, 1004-1013.

[137] De Gelder, J.; De Gussem, K.; Vandenabeele, P.; Moens, L.; Reference database of Raman spectra of biological molecules. *Journal of Raman Spectroscopy* **2007**, *38*, 1133-1147.

[138] Pecul, M.; Rizzo, A.; Leszczynski, J.; Vibrational Raman and Raman Optical Activity Spectra of d-Lactic Acid, d-Lactate, and d-Glyceraldehyde: Ab Initio Calculations. *The Journal of Physical Chemistry A* **2002**, *106*, 11008-11016.

[139] Culka, A.; Jehlicka, J.; Edwards, H. G. M.; Acquisition of Raman spectra of amino acids using portable instruments: Outdoor measurements and comparison. *Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy* **2010**, *77*, 978-983.

[140] Dhamelincourt, P.; Ramirez, F. J.; Polarized micro-Raman and FT-IR spectra of L-glutamine. *Applied Spectroscopy* **1993**, *47*, 446-451.

[141] Ujike, T.; Tominaga, Y.; Raman spectral analysis of liquid ammonia and aqueous solution of ammonia. *Journal of Raman Spectroscopy* **2002**, *33*, 485-493.

[142] McAughtrie, S.; Lau, K.; Faulds, K.; Graham, D.; 3D optical imaging of multiple SERS nanotags in cells. *Chemical Science* **2013**, *4*, 3566-3572.

[143] Kang, S.; Mullen, J.; Miranda, L. P.; Deshpande, R.; Utilization of tyrosine- and histidine-containing dipeptides to enhance productivity and culture viability. *Biotechnology and Bioengineering* **2012**, *109*, 2286-2294.

[144] Notingher, I.; Verrier, S.; Romanska, H.; Bishop, A. E.; Polak, J. M.; Hench, L. L.; In situ characterisation of living cells by Raman spectroscopy. *Spectroscopy-an International Journal* **2002**, *16*, 43-51.

[145] Palonpon, A. F.; Ando, J.; Yamakoshi, H.; Dodo, K.; Sodeoka, M.; Kawata, S.; Fujita, K.; Raman and SERS microscopy for molecular imaging of live cells. *Nature Protocols* **2013**, *8*, 677-692.

[146] Wang, S. Y.; Hasty, C. E.; Watson, P. A.; Wicksted, J. P.; Stith, R. D.; March, W. F.; Analysis of metabolites in acqueous solutions by using laser Raman spectroscopy. *Applied Optics* **1993**, *32*, 925-929.

[147] Grubbs, F. E.; Sample criteria for testing outlying observations. *Annals of Mathematical Statistics* **1950**, *21*, 27-58.

[148] Shapiro, S. S.; Wilk, M. B.; An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591-611.

[149] Bogaert, P. *Probabilités pour scientifiques et ingénieurs*; De Boeck: Paris, 2006.

[150] Wilk, M. B.; Gnanades.R; Probability plotting methods for analysis of data. *Biometrika* **1968**, *55*, 1-17.

[151] Box, G. E. P.; Non-normality and tests on variances. *Biometrika* **1953**, *40*, 318-335.

[152] Student; The probable error of a mean. *Biometrika* **1908**, *6*, 1-25.

[153] Sellick, C. A.; Hansen, R.; Jarvis, R. M.; Maqsood, A. R.; Stephens, G. M.; Dickson, A. J.; Goodacre, R.; Rapid monitoring of recombinant antibody production by mammalian cell cultures using Fourier transform infrared spectroscopy and chemometrics. *Biotechnology and Bioengineering* **2010**, *106*, 432-442.

[154] Ashton, L.; Xu, Y.; Brewster, V. L.; Cowcher, D. P.; Sellick, C. A.; Dickson, A. J.; Stephens, G. M.; Goodacre, R.; The challenge of applying Raman spectroscopy to monitor recombinant antibody production. *Analyst* **2013**, *138*, 6977-6985.

[155] André, S.; Saint Cristau, L.; Gaillard, S.; Devos, O.; Calvosa, E.; Duponchel, L.; In-line and real-time prediction of recombinant antibody titer by in situ Raman spectroscopy. *Analytica Chimica Acta* **2015**, *892*, 148-152.

[156] Chi, Z. H.; Chen, X. G.; Holtz, J. S. W.; Asher, S. A.; UV resonance Raman-selective amide vibrational enhancement: Quantitative methodology for determining protein secondary structure. *Biochemistry* **1998**, *37*, 2854-2864.

[157] Wu, Q.; Hamilton, T.; Nelson, W. H.; Elliott, S.; Sperry, J. F.; Wu, M.; UV Raman spectral intensities of E. coli and other bacteria excited at 228.9, 244.0, and 248.2 nm. *Analytical Chemistry* **2001**, *73*, 3432-3440.

[158] Lopez-Diez, E. C.; Goodacre, R.; Characterization of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Analytical Chemistry* **2004**, *76*, 585-591.

[159] Wen, Z. Q.; Raman spectroscopy of protein pharmaceuticals. *Journal of Pharmaceutical Sciences* **2007**, *96*, 2861-2878.

[160] Tuma, R.; Raman spectroscopy of proteins: from peptides to large assemblies. *Journal of Raman Spectroscopy* **2005**, *36*, 307-319.

[161] Thomas, G. J.; Raman spectroscopy of protein and nucleic acid assemblies. *Annual Review of Biophysics and Biomolecular Structure* **1999**, *28*, 1-27.

[162] Vanwart, H. E.; Scheraga, H. A.; Agreement with the disulfide stretching frequency-conformation correlation of Sugeta, Go, and Miyazawa. *Proceedings of the National Academy of Sciences of the United States of America* **1986**, *83*, 3064-3067.

[163] Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B.; Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy* **2000**, *54*, 413-419.

[164] Höskuldsson, A.; Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems* **2001**, *55*, 23-38.

[165] Centner, V.; Massart, D. L.; deNoord, O. E.; deJong, S.; Vandeginste, B. M.; Sterna, C.; Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry* **1996**, *68*, 3851-3858.

[166] Cai, W. S.; Li, Y. K.; Shao, X. G.; A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems* **2008**, *90*, 188-194.

[167] Shamsipur, M.; Zare-Shahabadi, V.; Hemmateenejad, B.; Akhond, M.; Ant colony optimisation: a powerful tool for wavelength selection. *Journal of Chemometrics* **2006**, *20*, 146-157.

[168] Ulber, R.; Frerichs, J. G.; Beutel, S.; Optical sensor systems for bioprocess monitoring. *Analytical and Bioanalytical Chemistry* **2003**, *376*, 342-348.

[169] Alford, J. S.; Bioprocess control: Advances and challenges. *Computers & Chemical Engineering* **2006**, *30*, 1464-1475.

[170] Mercier, S. M.; Diepenbroek, B.; Wijffels, R. H.; Streefland, M.; Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations. *Trends in Biotechnology* **2014**, *32*, 329-336.

[171] Macgregor, J. F.; Kourti, T.; Statistical process control of multivariate processes. *Control Engineering Practice* **1995**, *3*, 403-414.

[172] Tomasi, G.; van den Berg, F.; Andersson, C.; Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* **2004**, *18*, 231-241.

[173] Gnanadesikan, R.; Kettenring, J. R.; Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **1972**, *28*, 81-124.

[174] Wise, B. M.; Gallagher, N. B.; Butler, S. W.; White, D. A.; Barna, G. G.; MSPC tools for a semiconductor etch process. *Abstracts of Papers of the American Chemical Society* **1997**, *213*, 365-COMP.

Communications scientifiques

Publications

André, S.; Saint Cristau, L.; Gaillard, S.; Devos, O.; Calvosa, É.; Duponchel, L.; “In-line and real-time prediction of recombinant antibody titer by in situ Raman spectroscopy”. *Analytica Chimica Acta* **2015**, *892*, 148-152.

André, S.; Lagresle, S.; Hannas, Z.; Calvosa, É.; Duponchel, L.; “Mammalian cell culture monitoring using *in situ* spectroscopy: is your method really optimized?”. En cours de soumission.

Liu, Y.-J.; André, S.; Saint Cristau, L.; Gaillard, S.; Calvosa, É.; Duponchel, L.; "Multivariate statistical process control (MSPC) using Raman spectroscopy for in-line culture cell monitoring considering time-varying batches synchronized with correlation optimized warping (COW)”. En cours de soumission.

Communications orales

European Biotechnology Congress, Bucarest, 7 – 9 mai 2015 : "*On-line prediction of antibody titer by Raman spectroscopy and chemometrics*”, André, S.; Devos, O.; Calvosa, É.; Duponchel, L.

XXI^{èmes} journées du Groupe Français de Spectroscopie Vibrationnelle (GFSV), Reims, 17 – 19 juin 2015 : "*Mise en évidence de contraintes liées au suivi de bioprocédé par spectrométrie Raman et chimiométrie*", André, S.; Devos, O.; Duponchel, L.

Communication par affiche

Journée Découverte de la Recherche en chimie et en physique, Villeneuve d’Ascq, 25 mars 2015 : "Suivi de bioprocédés par spectrométrie Raman et chimiométrie", André S.; Devos, O.; Duponchel, L.

Apports de la spectroscopie Raman et de la chimiométrie au suivi *in situ* de cultures cellulaires : nouvelles perspectives en biotechnologie

Dans les années 2000, les grandes instances de sécurité sanitaire ont proposé l'initiative PAT (Process Analytical Technology) pour inciter les acteurs des milieux pharmaceutiques et agroalimentaires à améliorer leurs méthodes de production et de contrôle des produits manufacturés en utilisant de nouvelles techniques innovantes. Cette thèse s'inscrit dans cette démarche et propose d'exploiter la spectroscopie Raman, couplée aux outils chimiométriques pour le suivi en temps réel de cultures cellulaires à visée pharmaceutique. Acquis à l'aide d'une sonde optique à immersion, les spectres Raman in situ permettent de bénéficier d'une vue d'ensemble de l'état biochimique du bioprocédé au cours du temps. Ainsi, en appliquant des outils chimiométriques adéquats, il est possible de tirer profit des informations spectrales générées, notamment pour prédire les concentrations métaboliques au cours du temps.

Les travaux de recherche présentés dans cette thèse proposent tout d'abord de démontrer la nécessité d'optimiser l'acquisition spectrale et le traitement statistique des données pour différentes cultures cellulaires. Des modèles de régression robustes intégrant plusieurs sources de variabilité sont alors développés. Ils prennent en compte les variations inter-cultures, les changements de paramètres « procédés », le repositionnement des sondes optiques, voire le changement de lignée cellulaire. Enfin, ces mêmes spectres Raman sont utilisés pour développer des outils de contrôle statistique des procédés afin de détecter en temps réel des états anormaux comme par exemple la contamination ou encore pour prédire la concentration d'un produit d'intérêt comme les anticorps.

Mots clés : spectroscopie Raman, chimiométrie, cellules – cultures et milieux de cultures, microbiologie pharmaceutique, processus biotechnologiques – surveillance, sondes optiques.