The Development and Application of Mid-Infrared Spectroscopy as a Process Analytical Technology (PAT) tool for Cell Culture Applications

by

Róisín Foley, B.Eng. (Hons) Chemical & Process Engineering

This thesis is submitted to Dublin City University for the degree Doctor of Philosophy (PhD) in the School of Biotechnology

August 2013

Thesis Supervisor: Prof. Ian Marison

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy (PhD) is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ **(Candidate)** **ID No.: 59108185** **Date** _____

Table of Contents

Abstract

The objective of this thesis was to investigate the use of mid-infrared spectroscopy (MIR) as a PAT tool in bioprocessing. This was achieved through the development of chemometric models from MIR spectroscopic data. Models were applied to both upstream and downstream bioprocess steps to evaluate the potential of MIR as a PAT tool in each scenario.

The first study included a preliminary examination of 8 typical components found in a mammalian cell culture medium. A multivariate limit of detection (LOD) analysis was performed to establish the monitoring potential of the instrument for the given application. This initial work identified the components which were unlikely to be accurately detected, such as glutamine, but also highlighted the components that showed promise, such as glucose. A 7-level experimental design was used to develop partial least squares regression (PLS) models for each of the 8 components, with optimal model errors ranging from 6.03% for glucose to 63.06% for glutamine. An external influence investigation into the factors likely to impact model prediction ability was also performed. A statistical analysis on these influences enabled the significance of the effect to be determined. Finally, all investigative work performed in this study was completed using 2 MIR immersion probes; the first, a fixed conduit immersion probe and the second, a flexible fibre-optic immersion probe. This allowed for a comparative analysis of probes and identified the fixed conduit probe as the most suitable for the given application.

The second study applied models developed for glucose and lactate, in the first study, to a series of CHO DP12 cell cultures. The objective of this work was to investigate

how effective the previously developed models were at predicting glucose and lactate concentrations when applied in-situ, in a bioreactor. This study explored the at-line application of MIR and showed an improvement of 60.27% for glucose and 13.22% for lactate predictions, when applied at-line rather than in-situ. Central to this study was the optimisation of the models used, in an effort to improve their accuracy and robustness. The impact of the size and nature of the calibration set was investigated to identify the optimum calibration set for in-situ and at-line model development. When offline data was used as part of the optimised calibration training set, models consistently performed better than the original partial least squares regression (PLS) models, created solely from synthetic samples. Optimised glucose model results showed an improvement in RMSEP of 37.93% while optimised lactate model results had an improved RMSEP of 61.98%.

The third and final study presented here investigated the use of MIR as a qualitative and quantitative tool for total and recombinant protein detection. Exploratory work to establish the instruments capability in distinguishing between differing proteins was initially performed via principal component analysis (PCA). This showed that the instrument used could identify the recombinant protein of interest among a group of 5 other standard proteins. In-process samples, provided by an industrial collaborator, were used to develop PLS models for total and recombinant protein prediction. Results indicated that models performed better for total protein quantification, with the minimum percentage error of prediction, (PEP) 2.39%. The smallest PEP for the recombinant protein was found to be 6.66%. This study was completed with an investigation into the likely impact of 2 detergents on model performance. Due to the common usage of detergents in protein production, for e.g. protein solubilisation or

virus inactivation, the effect of these on MIR spectra and consequently models predictions was investigated. This analysis indicated that detergents were likely to impact results, particularly at concentrations at the higher end of the typically used range.

This thesis establishes and evaluates the potential of MIR spectroscopy as a PAT tool. It presents 3 studies which highlight development methodologies and outline possible applications, all the while seeking to optimise results obtained. Through systematic, novel investigations this thesis shows that MIR can be used as a PAT tool, but equally, it raises warnings of when the technique or analysis methods may fall short of the desired result.

Acknowledgements

I would like to thank my supervisor Prof. Ian Marison for offering me this opportunity. I would also like to thank my colleagues in the Laboratory for Integrated Bioprocessing, in Dublin City University, for their encouragement throughout the course of this project, in particular Dr. Siobhán Hennessy for her endless supply of advice, knowledge and enthusiasm. This research was funded by Science Foundation Ireland. I would like to express my gratitude to SFI for providing me with such an opportunity. I would especially like to acknowledge the exceptional support of my family and friends, in particular, Karl and Karen, for their patience and understanding.

Finally, I wish to dedicate this work to my parents, without whom, its completion would not have been possible. Their unshakable belief in my ability, coupled with their continued encouragement and support have enabled me to pursue my studies with zeal and confidence.

List of Abbreviations

ATR	Attenuated total reflectance
BSA	Bovine serum albumi
cGMP	current Good Manufacturing Practices
CHO	Chinese hamster ovary
CPP	Critical process parameter
CQA	Critical quality attributes
DNA	Deoxyribonucleic acid
FDA	Food and drugs administration
FT-IR	Fourier transform infrared
HPLC	High performance liquid chromatography
HSA	Human serum albumin
IRE	Internal reflectance element
LOD	Limit of detection
LV	Latent variable
MC	Mean Centred
MIR	Mid infrared
M^3C	Measurement, monitoring, modelling & control
NIR	Near infrared
PAT	Process analytical technology
PC	Principal component
PCA	Principal component analysis
PEC	Percentage error of calibration
PECV	Percentage error of cross validation
PEP	Percentage error of prediction
PLS	Partial least squares

PLS-DA	Partial least squares discriminant analysis
QbD	Quality by design
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross validation
RMSEP	Root mean square error of prediction
SIMCA	Soft independent modelling of class analogy
SNR	Signal-to-noise ratio
SNV	Standard normal variate

Chapter 1: Introduction

The field of biotechnology has undergone exceptional growth in the last decade. There is an ever increasing global demand for high quality biopharmaceuticals at lower costs. The pharmaceutical industry has previously experienced such mounting pressure; but with market focus shifting to biopharmaceuticals, it is now the turn of the biotechnology sector to accept the torch and take a leading role in embracing dynamic initiatives such as "Process Analytical Technology (PAT) and Quality by Design (QbD)". Although these initiatives are almost a decade old, their introduction to the biopharmaceutical industry is still in its infancy. However the search has begun to find robust and reliable techniques to enable the concept of Process Analytical Technology (PAT) to take root within the industry.

The production of recombinant proteins for therapeutic and diagnostic applications is developing at an extremely high rate, principally based on microbial and animal cell production systems. The efficient control of cell culture conditions such that very high cell densities may be attained, is hugely desirable. However, reaching high cell densities is not the only challenge. Maintaining and controlling this high cell concentration over the course of the cell culture is even more critical. Uncontrolled systems may lead to extreme loss in viability, high release of proteolytic enzymes and, hence lower protein productivity and increase protein degradation in the bioreactor.[1] Therefore systems capable of tight control to ensure high product yield and quality are imperative. This is generally achieved by careful formulation of the culture medium, coupled with controlled feeding of the medium through fed-batch or perfusion systems, to simultaneously maintain high cell viability and high production rates, while maintaining the required product quality.

The increasing demand for improved process efficiency and consistent product quality in bioprocessing has fuelled extensive research efforts in the area of on-line bioprocess monitoring and control. An important driving force for this growth was set in place by the Process Analytical Technology (PAT) initiative, proposed in 2002 by the Food and Drug Administration (FDA).[2] PAT endeavours to establish in the pharmaceutical industry, the ideology of Quality by Design, QbD, where continuous on-line process monitoring and supervision ensure optimum manufacturing conditions, allow complete product traceability and quality control, as well as improve early fault detection. In addition, the implementation of PAT to a process ultimately results in greater process understanding as it effectively monitors, and controls if necessary, all the Critical Process Parameters (CPP's) hence providing considerable insight into the Critical Quality Attributes (CQA's) of the process.[3] The availability of real-time analytical results has been shown to reduce production costs by reducing losses caused by unnecessary waiting periods between process steps, for example, results from in-process testing. Also, costs incurred due to batch failures and reworks are significantly reduced.[4]

Several 'conventional' bioprocess monitoring techniques are widely used; temperature, pH, dissolved oxygen, carbon dioxide etc. These analysers, most often in the form of in-situ probes, are used mainly for controlling the corresponding process parameters, although they do not provide much insight into the reaction itself. Analysers capable of providing information on the contents of a bioreactor at any point in time perhaps provide the greatest insight into the bioprocess. A major challenge currently in the area of bioprocess monitoring lies in ensuring the reliability of monitoring tools and

the accuracy of the measurements they provide in on-line conditions. As the concentrations of the major analytes and metabolites in biotechnological applications are typically very low, achieving the required reliability and precision of on-line measurements often proves a difficult task. On-line monitoring tools provide reliable, instantaneous analytical information about the bioprocess, which in turn can be used for advanced control schemes aimed at optimising the process in real-time. PAT instruments in a bioprocess application should meet certain criteria. They should be capable of rapid and accurate on-line monitoring with the ability to be integrated into reliable and robust calibration systems as well as being non-invasive and non-destructive. These applications should also provide dependable analytical feedback to continually maintain the desired operational parameters and allow for implementation of stringent process control mechanisms.

Vibrational spectroscopy has huge potential as a PAT tool in bioprocess monitoring. In-situ probes meet much, if not all of the criteria listed above. These probes can be sterilised in place, do not require any sample preparation and can produce spectroscopic data in seconds.[5] Infrared spectroscopy; near-infrared (NIR) and mid infra-red (MIR), work on the basis of detecting the vibration characteristics (stretching, contracting, bending, etc.) of various chemical functional groups over the specific infrared frequency range of light. Another vibrational spectroscopic technique, Raman spectroscopy, is complementary to infrared spectroscopy and is based on light scattering. Previous studies have demonstrated the application of NIR, MIR and Raman spectroscopy to cell cultures for the simultaneous on-line monitoring of media analytes and metabolites.[6-12] Other studies have highlighted the use of infra-red spectroscopy as a rapid technique for recombinant protein detection, outlining its

capabilities in protein characterisation and quantification.[13-16] However these uses have yet to be significantly demonstrated in an on-line capacity.

The useful information embedded within spectral data sets needs to be extracted if it is to be of use as part of a monitoring and/or control strategy. It is the development of multivariate calibration models which is the key to unlocking this information. In the case of spectroscopic data, a large number of independent variables are generated which relate to one predicted dependent variable e.g. absorbance values (independent variables) over a range of wavenumbers in the mid-infrared region can be used to predict the concentration (dependent variable) of a particular component. These variables are related to each other by a calibration model.

Chemometrics (multivariate analysis techniques) are used to establish correlations between a dependent variable, such as concentration, and absorbance. They are often used as data reduction techniques since chemometric analysis allows multivariate data to be transformed into a much smaller number of variables. The important information is maximised and system noise is minimised. A number of chemometric techniques can be employed, depending on the required information. In the studies presented in this thesis, the chemometric techniques used were principle component analysis (PCA) and partial least squares regression (PLS). PCA is often used for exploratory analysis and pattern recognition. In bioprocessing, PCA can be used as a qualification technique for raw materials or products where differences between samples may be highlighted. PLS is a supervised method which requires the use of a training or calibration set to develop a predictive model. It is frequently used in the generation of

calibration models capable of predicting the concentrations of various components present in a system.

The overall aim of this thesis was to establish the potential of MIR as a PAT tool in bioprocessing. This was achieved by employing chemometric techniques to develop calibration models capable of predicting the concentrations of analytes and metabolites present in a mammalian cell culture medium. These models were applied in real-time to a series of cultures and the efficacy of the technique for analyte and metabolite measurement was evaluated. The applicability of this technique in the monitoring of product, in the form of recombinant protein, was investigated by development of PLS models from spectral data obtained from in-process culture samples. Three separate studies were completed in order to carry out all aspects of the research outlined above.

Study 1 proposed a methodology on how to develop and evaluate MIR spectroscopy as a PAT tool in the quantification of media components. This work identified 8 components typically present in a mammalian cell culture medium and systematically examined each one in terms of the ability of MIR to detect and accurately predict its concentration.

Study 2 applied the PLS models developed for glucose and lactate in study 1, to mammalian cell cultures in real time to determine the accuracy of these models. This study also highlighted the differences when the technique was applied on-line and at-line.

Finally, study 3 focused on the ability of MIR to detect total and recombinant protein in samples taken directly from the primary recovery stage of an industrial process. This exploratory work investigated the potential of in-situ MIR as a PAT tool for qualitative and quantitative protein analysis.

All 3 studies presented in this thesis provide a cohesive approach to determining the ability of MIR to perform effectively as a PAT tool in bioprocessing. Both upstream (in the form of mammalian cell culture cultivation) and downstream (in the form of primary recovery) processing is considered. In addition to evaluating the potential of MIR as a PAT tool, each of the individual studies raises questions as to optimum methodologies, with limit of detection, calibration set type and means of evaluation, all scrutinised.

References

1. Ryll, T. *et al*. Performance of small-scale CHO perfusion cultures using an acoustic cell filtration device for cell retention: Characterization of separation efficiency and impact of perfusion on product quality. *Biotechnology and bioengineering* **69**, 440-449 (2000).

2. FDA. *Pharmaceutical cGMPS for the 21st Century — A Risk-Based Approach - Final Report*, 2004).

3. U.S. Department of Health and Human Services: Food and Drug Administration. in *Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance* , 2004).

4. Bakeev, K. Process Analytical Technology; Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries. **1**, 451 (2005).

5. Landgrebe, D. *et al*. On-line infrared spectroscopy for bioprocess monitoring. *Appl. Microbiol. Biotechnol.* **88**, 11-22 (2010).

6. Dabros, M., Amrhein, M., Gujral, P. & von Stockar, U. On-line Recalibration of Spectral Measurements using Metabolite Injections and Dynamic Orthogonal Projection. *Appl. Spectrosc.* **61**, 507-513 (2007).

7. 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. *J. Biotechnol.* **113**, 231-245 (2004).

8. 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. *J. Biotechnol.* **128**, 344-353 (2007).

9. Arnold, S. A., Gaensakoo, R., Harvey, L. M. & McNeil, B. Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch Escherichia coli process. *Biotechnol. Bioeng.* **80**, 405-413 (2002).

10. Cervera, A. E., Petersen, N., Lantz, A. E., Larsen, A. & Gernaey, K. V. Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnol. Prog.* **25**, 1561-1581 (2009).

11. Doak, D. L. & Phillips, J. A. In Situ Monitoring of an *Escherichia coli* Fermentation using a Diamond Composition ATR Probe and Mid-infrared Spectroscopy. *Biotechnol. Prog.* **15**, 529-539 (1999).

12. Whelan, J., Craven, S. & Glennon, B. In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnol. Prog.* **28**, 1355-1362 (2012).

13. Sellick, C. A. *et al.* Rapid monitoring of recombinant antibody production by mammalian cell cultures using fourier transform infrared spectroscopy and chemometrics. *Biotechnol. Bioeng.* **106**, 432-442 (2010).

14. Haris, P. I. & Severcan, F. FTIR spectroscopic characterization of protein structure in aqueous and non-aqueous media. *J Molec Catal B* **7**, 207-221 (1999).

15. Crowley, J., McCarthy, B., Nunn, N. S., Harvey, L. M. & McNeil, B. Monitoring a recombinant Pichia pastoris fed batch process using Fourier transform mid-infrared spectroscopy (FT-MIRS). *Biotechnol. Lett.* **22**, 1907-1912 (2000).

16. Barth, A. Infrared spectroscopy of proteins. *Biochim. Biophys. Acta-Bioenerg.* **1767**, 1073-1101 (2007).

Chapter 2: Literature Review

The choice of suitable on-line analytical techniques and data processing for monitoring of bioprocesses

Adapted from the Springer Berlin Heidelberg publication: Advances in Biochemical Engineering Biotechnology, Volume "Stage of the Art of M3C", 2012. Series ISSN 0724-6145

Abstract

With increasing pressure from regulatory authorities on industry to develop processes embracing 'Process Analytical Technology' (PAT) initiatives, there is a growing demand to establish reliable tools and systems capable of meeting this need. With regard to monitoring and control of bioprocesses, this need translates to a search for robust instrumentation capable of monitoring the critical process parameters in real time. The application of such technologies at all stages of the process, from the initial R&D phase, to process optimisation and production enhances process understanding and paves the way for the development of control platforms.

An examination of the PAT concept and selected tools are presented here. A description of each tool is given, with particular emphasis on the nature of the signal produced and how these relate to measurements of biomass, metabolites and product. A description of the signal processing that is necessary to gain meaningful results from the different tools is also given. Many techniques such as those based on vibrational spectroscopy are of particular interest, since they are capable of monitoring several critical process parameters which are typically controlled in a bioprocess. A window

of application for each of the techniques, when used in the area of bioprocessing, is suggested based on their uses and inherent limitations.

Keywords PAT, Critical Process Parameters, Bioprocess Monitoring, Spectroscopy, Calorimetry, Data Reconciliation, Bioprocess Control

2.1 Introduction

The last decade has introduced a significant number of changes to the pharmaceutical and biopharmaceutical industries, not least in the area of quality assurance and regulatory compliance. This new focus has largely been driven by the Food and Drug Administration (FDA). In 2002 the FDA announced a new initiative, "Pharmaceutical cGMPs for the 21st Century", the purpose of which was to modernise the regulation of pharmaceutical quality. The initiative supports and promotes the use of risk-based and science-based approaches for regulatory decision making, throughout the entire life-cycle of a product.[1] After 2 years in development, the final report outlines the envisioned direction in which the pharmaceutical and biopharmaceutical industries should be moving, but also provides guidance on how to make the proposed changes and embrace the new concepts put forward.

Central to the implementation of this new system is the use of science and engineering knowledge to establish boundaries and evaluate processes. Previously, a manufacturing process was developed and quality control and quality assurance tests were then applied to ensure compliance. This new initiative aims to use science and engineering knowledge to mitigate risk, by reducing process and product variability and applying continuous process improvement. Industry guidelines published by the

FDA in May 2006, "Q8 Pharmaceutical Development", state, "quality cannot be tested into the products, it should be built in by design".[2] Essentially this means that the rigorous testing of the past cannot improve product quality or enhance the process but rather quality should be pivotal throughout the lifecycle of a process and a key factor from the initial stages of development and process design. This introduces the concept of "Quality by Design" (QbD), whereby a "design space" is established, within which, the product quality profile is defined, the critical quality attributes (CQAs) and critical process parameters (CPPs) are identified and the manufacturing process is controlled. Process changes that occur within the design space are acceptable as the design space would have been subject to regulatory assessment and approval at the time the manufacturing process was filed. However movement outside the design boundary is considered a change and as such would most likely require regulatory post approval.

In order to establish a design space that will allow for maximum process flexibility while ensuring all CPPs and CQAs are identified and maintained, a large degree of process understanding is essential. Process analytical technology (PAT) is a "pillar/guiding principle" of the cGMP initiative.[1] The PAT framework published in September 2004 defines process understanding and highlights the tools required to achieve this standard of process knowledge:

"A process is generally considered well understood when (1) all critical sources of variability are identified and explained; (2) variability is managed by the process; and, (3) product quality attributes can be accurately and reliably predicted over the design space established"[3]

PAT provides in depth process understanding, but to implement PAT and operate under the principle of Quality by Design the process must be well understood. At the point of writing this chapter, the initiative is still in its infancy with regard to industrial implementation, and so a large degree of progress has been in the area of PAT tool development. With research and development focused on process analysers and data acquisition tools, many in the industry have applied these to processes to glean greater process knowledge. However, although PAT is a relatively new concept it has evolved over the last decade. It has transitioned from being an analysis in the process, to supplement quality control, to being an analysis of the process.[4]

As already mentioned the main driver of this initiative is the FDA, however the pharmaceutical and biopharmaceutical industries are a close second. They strive to ensure that products released to market are of the highest quality and compliant with regulations. This reason alone merits acceptance of these new guidelines, however there are other benefits. Live feedback and process control, reduced cycle times, laboratory test replacement and improved safety result in increased product yield and quality, reduction in batch failures and rework costs and increased throughput. Such changes result in continuous improvement and operational excellence, which in turn increases business value. As a result of PAT being embraced by industry, tools must be developed that are capable of real-time monitoring and control. Currently very few developed tools exist and even fewer have actually been implemented in a manufacturing environment.

This chapter explores the use of selected PAT tools which can be used in the context of M^3C in bioprocess applications and looks at the advantages and limitations of each.

Vibrational spectroscopic instruments as PAT analysers are examined. The theory behind the use of MIR, NIR and Raman spectroscopy for bioprocessing applications is stated and necessity of using multivariate data processing is explained. Reported uses of these techniques for bioprocess monitoring and control applications are summarised and the current state of the different technologies are compared.

2.2 Vibrational Spectroscopy

The energy of a molecule is quantised and can only exist in certain discrete energy levels, E_0, E_1, E_2 etc. When a photon (energy in the form of light) is emitted from a light source, a molecule will only absorb this energy if it equates to the difference in energy between these discrete energy levels. The energy emitted by the light or absorbed by the molecule is related to the frequency by the following equation:

$$E = h\upsilon \tag{2.1}$$

where; E is the energy emitted or absorbed, h is Plank's constant and υ is the frequency. As the energy is directly proportional to the frequency, it therefore follows that a photon with higher frequency has a higher energy. The energy absorbed by the molecule results in molecular vibrations. These vibrations can take any of the following forms: stretching, bending, rocking, wagging and twisting. In order for a molecule to absorb infrared radiation the frequency of the radiation must equal the frequency of the molecular vibration and this vibration must cause a change in the dipole moment of the molecule.

Infrared spectroscopy is concerned with the region of the electromagnetic spectrum between the visible and microwave regions. This region, the infrared region, is further broken down into the near-infrared region (12,500-4000cm^{-1}/800-2500nm), mid-

infrared region (4000-400cm^{-1}/2500-25000nm), and far-infrared region (<400cm^{-1}/>25000nm).

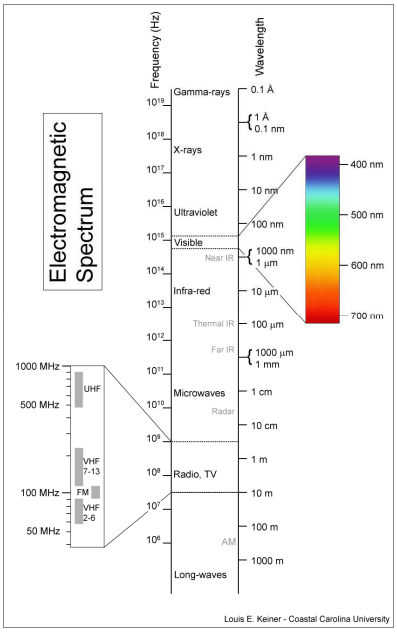


Fig. 2.1 Electromagnetic Spectrum

Energy in the mid-infrared region has lower frequency and so lower energy than that of the near-infrared region. Mid-infrared spectroscopy looks at the change in energy when the molecule is radiated and moves from the ground state to the next excited energy level. This is the fundamental energy change. As a result absorption bands within the mid-infrared region can be attributed to specific molecules or functional groups within the molecule; hence MIR spectroscopy is sometimes referred to as a "fingerprinting" technique. The higher energy of the near-infrared region results in the radiated molecule moving from the ground state to excited energy levels above the

fundamental absorption state, therefore NIR is focused on the combination bands and overtones.

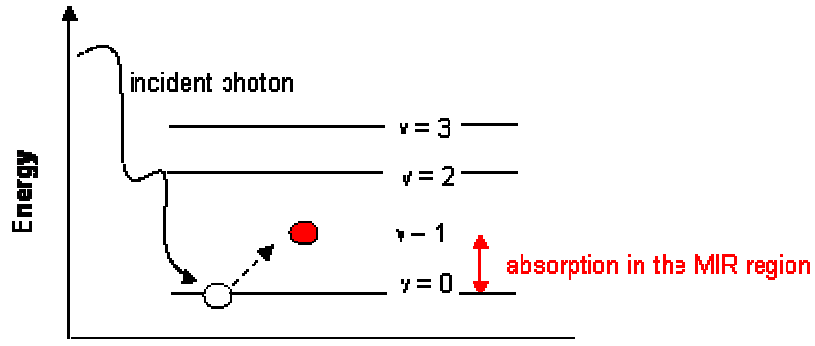


Fig. 2.2 MIR energy level diagram

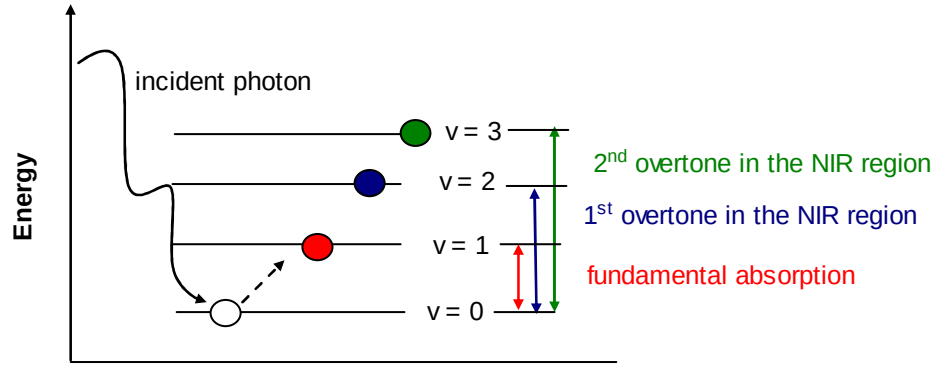


Fig. 2.3 NIR energy level diagram

Spectra are most commonly presented in terms of "intensity 'v' wavenumber" or "absorbance 'v' wavenumber". The absorbance and concentration can be calculated using the Beer-Lambert Law as follows:

$$A_v = -\log_{10}\left(\frac{I_v}{I_{v,0}}\right) = \varepsilon bc \qquad (2.2)$$

where :

A_v = absorbance at wavenumber v

I_v = intensity of the light emitted from the sample at wavenumber v

$I_{v,0}$ = intensity of the light emitted from the background (usually water or air) at wavenumber v

ε = Molar absorption coefficient

b = pathlength

c = concentration

Put more simply, the absorbance of a component at a particular wavenumber/wavelength is directly proportional to the concentration of that component. In order to transform spectral data into meaningful results it is necessary to develop a calibration model which relates the absorbance to concentration values, much the same as a mercury thermometer relates increase in height to temperature. In the case of the thermometer there is only one independent (measured) variable; the height of the mercury in the tube. This is referred to as a univariate model. Where spectral data are concerned, there are several independent variables; the multiple absorbance values over the range of the MIR or NIR spectrum. This is referred to as a multivariate model. Chemometric techniques are exploited to extract the relevant data, and in this way act as a key to unlocking the information buried within the spectral data.

Raman spectroscopy is a technique which is complementary to infrared spectroscopy and is based on the scattering of light. When a monochromatic light of given frequency

is directed at a molecule, most photons of energy will be elastically scattered and this is known as Rayleigh scattering. A small fraction (1 in 10^8 photons) of the light will exchange energy with the molecule. This is inelastic or Raman scattering. [5, 6] Figure 2.4 is an energy level diagram showing the different scattering phenomena. Scattering is a two photon process with one photon being absorbed and a second photon being emitted. With elastic Rayleigh scattering, the energy of the photon absorbed and the molecule is excited to a virtual state. The energy of the photon emitted is equal to the energy of the photon absorbed but with Raman scattering there is a difference between the energy of the absorbed and emitted photon [7] Energy can be transferred to a molecule excited to a virtual state and when the emitted light is scattered, the frequency of the photon is higher than the frequency of the original photon of light. This phenomenon is known as Stokes Raman scattering. An already excited molecule that interacts with another photon will have excessive energy so, when scattering occurs, the frequency of the emitted photon is lower than that of the original photon and this is known as anti-Stokes Raman scattering. In order for Raman scattering to occur, a change in polarizability of the molecule is necessary. Polarizability is the relative ease with which a dipole moment is induced when the oscillating electric field of a light source interacts with a molecule. [8]

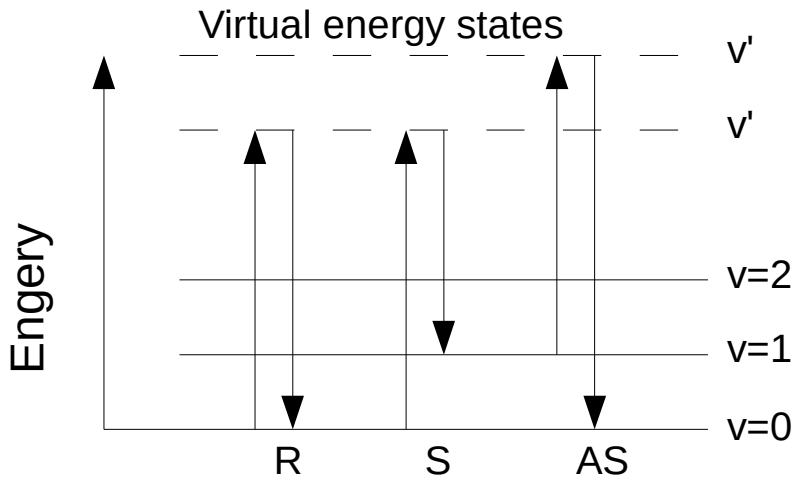


Fig. 2.4 Energy level diagram of scattering phenomena; Rayleigh scattering (R), Stokes Raman scattering (S) and anti-Stokes Raman scatting (AS)

In order to transform spectral data into meaningful results it is necessary to develop a calibration model which relates the spectra to a process parameter e.g. concentration values of a substrate. Chemometric techniques are exploited to extract the relevant data; this will be discussed further in following sections.

2.3 Development of MIR spectroscopy

Infrared radiation was discovered in 1800 by Sir William Herschel and following this, the first mid-infrared spectrometer was constructed by Melloni in 1833. In 1891 Albert Michelson invented the interferometer which produced an interference pattern by splitting a beam of light into two paths, bouncing the beams back and recombining them. A year later Lord Rayleigh proposed that this interference pattern could be converted into a spectrum using the Fourier Transformation mathematical technique. The first half of the 20th century saw little development in FT-IR spectroscopy and its potential as an analytical tool remained largely untapped until the late 1950's and early

1960's. With huge technological advances and the advent of the computer age fast approaching a need for highly sensitive, rapid detection devices over a broad range of applications arose. Early systems did not meet their goals, as they were hindered by memory size, poor stability and low resolution, so it was not until the late 1960's and early 1970's, when major technological improvements we made, that FT-IR instruments became commercially available. These systems were a welcome addition to their dispersive counterparts as they were more sensitive, had greater wavelength accuracy, and allowed for rapid spectral acquisition and manipulation in the form of spectral subtraction. Spectral subtraction allows the user to study mixtures of components without having to complete sample processing and separation prior to testing, thus expanding the boundaries of FT-IR spectroscopy and increasing it's applicability in a number of areas e.g. bioprocess applications.

Since the commercial debut of the FT-IR system in the 1970's the technology has been embraced by manufacturing industries and research communities alike. Instruments have been adapted and improved to meet the specific needs of the end user. Spectral measurements can be in several forms; transmission of radiation, internal reflectance (attenuated total reflectance), external reflectance, bulk diffuse reflectance and photoacoustic determinations.[9] In addition, the sampling configuration must be suitable for the instruments use. Given the variety of sampling techniques and sample interfacing available; from off–line transmission cells to in-situ fibre-optic reflectance probes, mid-infrared spectroscopy is adaptable to almost any area. The focus here is on its applicability as an on-line monitor in bioprocess applications.

MIR immersion probes have been available since the late 1980's. Improvements in the design and material of construction are on-going. At the early stages of development MIR fibres suffered from high material absorption and scattering and poor mechanical and chemical stability therefore "fixed" arm probes with parallel light pipes using internal reflection spectroscopy were found to be more suitable. However when placed in a process environment this design is far from ideal. These probes need to be precisely aligned and are highly sensitive to vibrations in the surrounding area, which can result in alignment changes and hence spectral differences.[10-12] There have been major advances in the development of fibre-optic materials over the last 10 years and these improvements have had far reaching consequences. In the case of MIR instrumentation it has resulted in flexible, more robust immersion probes which address many of the problems encountered with the rigid conduit probes. However regardless of probe type, process disturbances will regularly impact the spectra collected and these disturbances need to be accounted for when developing multivariate calibration models. The short pathlength of MIR, when compared to that of NIR means that from a sampling perspective the MIR does not penetrate as far into the material and may not be as representative of the sample as NIR would be, however, in the presence of particulate matter the shorter pathlength of MIR reduces light scattering, which is commonly experienced when NIR is used in a similar situation.[9]

2.4 Development of NIR spectroscopy

In 1800, Sir William Herschel separated the electromagnetic spectrum by passing light through a prism. He noted that the temperature increased significantly towards and beyond the red region, now referred to as the near-infrared region. This experiment marked the discovery of near-infrared radiation. However, as with mid-infrared

spectroscopy, near-infrared spectroscopy was not widely used in any applications until the second half of the 20th century when optical techniques and computers capable of complex analysis were rapidly developing.

In the 1950's NIR showed potential as a quantitative analysis tool, however it lagged behind the development of other optical devices such as MIR spectrometers, and many of its initial uses were as an accessory to these devices. The US Dept. of Agriculture, under the work of Karl Norris, began to investigate the use of NIR in the measurement of moisture content in cereals. Major difficulties existed such as interference and absorption of other constituents, and these were only resolved with the development of multivariate statistical methods which allowed the correlation of NIR spectral features. Following the introduction of such powerful computer aided tools; the first stand alone instrument became available in the early 1980's. Development of NIR spectroscopy as a quality and process control tool is largely due to the availability of efficient chemometric techniques and varying spectrometer configurations.[13] Its use within the pharmaceutical/biopharmaceutical industry is fast growing, but it remains relatively new in terms of its status as a process analytical tool.

Central to the development of NIR spectroscopy as a PAT tool is the availability of adequate sampling devices. Process environments vary considerably and selecting the most appropriate sampling interface is of paramount importance. A large degree of process understanding is vital prior to choosing the sampling interface. The physical, chemical and optical nature of the process stream should be known in order to determine if the results will be significantly impacted by light scattering. This is particularly important where NIR is concerned as the strong light source and the weak

absorbance allow infrared radiation to penetrate further into the sample, allowing particulates present in a suspension or slurry to cause light scattering. Other aspects such as the potential of the process fluid to foul the probe or sample system should also be evaluated.[9] Typically NIR sampling systems for in-situ process monitoring are in the form of one of the following two types: extractive sampling systems, where the sample is removed from the main process line analysed and returned to the process or immersion probe sampling, where a probe is placed in the process stream of interest and a number of intermittent scans taken using fibre-optics to connect the analyser to the probe interface. A number of variations of each of the above two types exist. Invariably, it is the process conditions which will dictate the system of choice.

2.5 Development of Raman spectroscopy

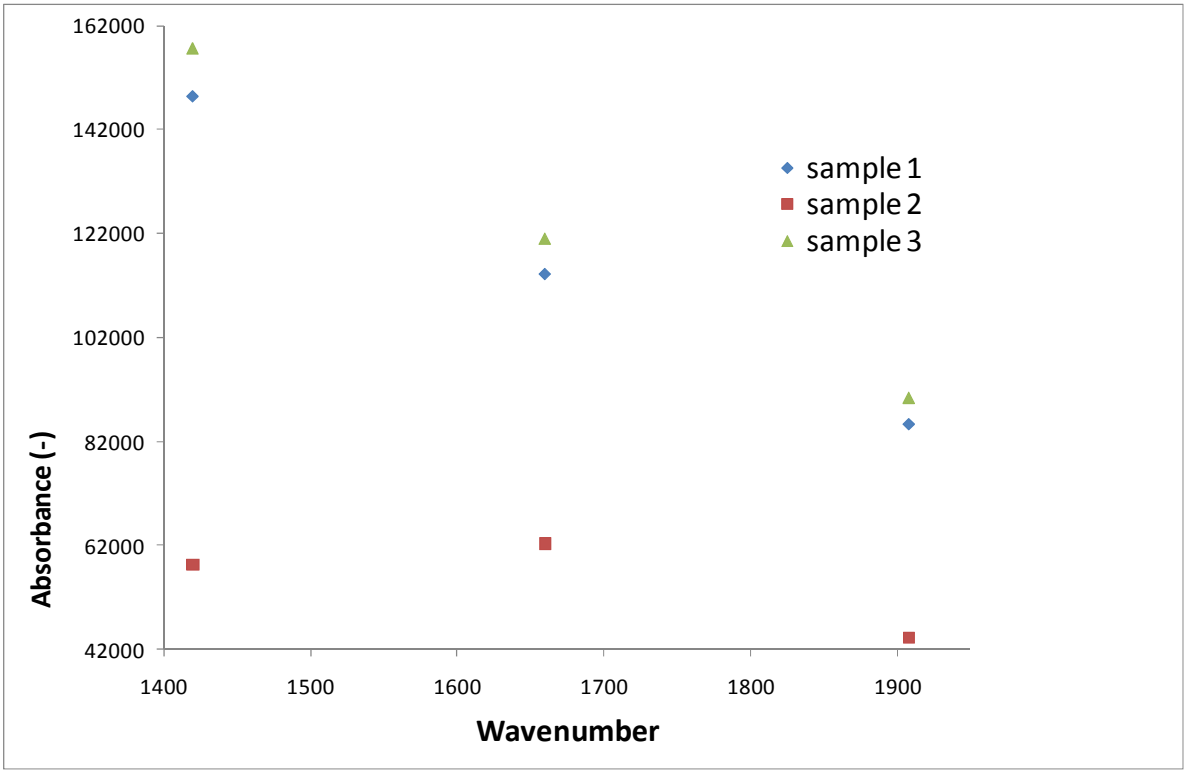
During the 1920s the scattering effect theory was investigated by a number of researchers including physicist C. V. Raman, who in 1928 was the first to experimentally demonstrate the Raman effect in liquids. [14] Originally instruments consisted of a mercury lamp passed through a filter to produce monochromatic light which was then used to excite the sample. Scattered radiation was observed at 90^{o} degrees from the incident radiation, dispersed using a glass prism and recorded on a photographic plate. [7] Such a system is known as dispersive Raman spectroscopy. A modified FT-IR instrument was also used to collect Raman spectra and the use of such instruments is now known as FT-Raman spectroscopy. Dispersive Raman spectroscopy and FT-Raman spectroscopy each have their own specific advantages and both continue to be used for different applications.[7, 15]

The basic components of any Raman spectrometer are the excitation source, the spectrometer and the detector. Significant advances in all three areas [16] have lead to the possibility of using Raman spectroscopy as a PAT tool. A major development in modern instruments was the incorporation of the laser as the monochromatic light source since the 1960s. As such a small fraction of light is Raman scattered, a powerful excitation source is necessary and powerful laser light has led to the possibility of measuring smaller samples in a shorter amount of time.[5, 7] Optical filtering devices are used to filter out a large portion of the Rayleigh scattered photons and so maximise the amount of Raman scattered photons which can be detected. In the early 1990s holographic notch filters were introduced. Their efficacy is due to the fact that the optical density of the notch filter is high and the spectral bandwidth of the notch can be extremely narrow. They are also free from extraneous reflection bands and provide significantly higher laser damage thresholds than standard interference filters.[16, 17] Silicon based charged couple devices (CCD) are replacing photomultipliers as detectors in dispersive Raman instruments and this has allowed simultaneous measurement of multiple locations. Detectors in FT-Raman instruments have also been improved.[5, 18] Fibre-optics allow probes to be inserted directly into a reactor and cables up to 100 meters may be used to connect the instrument to the measurement point. Non-contact sampling where a probe can make measurements through a sight glass is also possible.[8] All of these advances mean that Raman spectroscopy has developed into a versatile PAT tool.

2.6 Interpretation of spectral data using chemometrics

It is almost impossible to discuss the use of spectral data without a discussion on concept of chemometrics as the various chemometric techniques maximise the

information available from the spectroscopic instruments. Chemometrics can be defined as the chemical discipline that uses mathematical, statistical and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant chemical information by analysing chemical data.[19] When applied to the spectra collected on-line during a bioprocess by MIR or NIR spectrometers, it is the second function of chemometrics that is of most interest. Process data from a spectrometer is analysed in a multivariate rather than a univariate way, i.e. for each sample, the response at multiple wavenumbers are taken into account. If the spectrum of a sample was recorded at three wavenumbers using any spectroscopic technique, a simple 2-dimensional plot of response versus wavenumber could be used to visualise the data. The same data can also represented by a single point in 3 dimensions where each dimension is corresponds to a wavenumber.



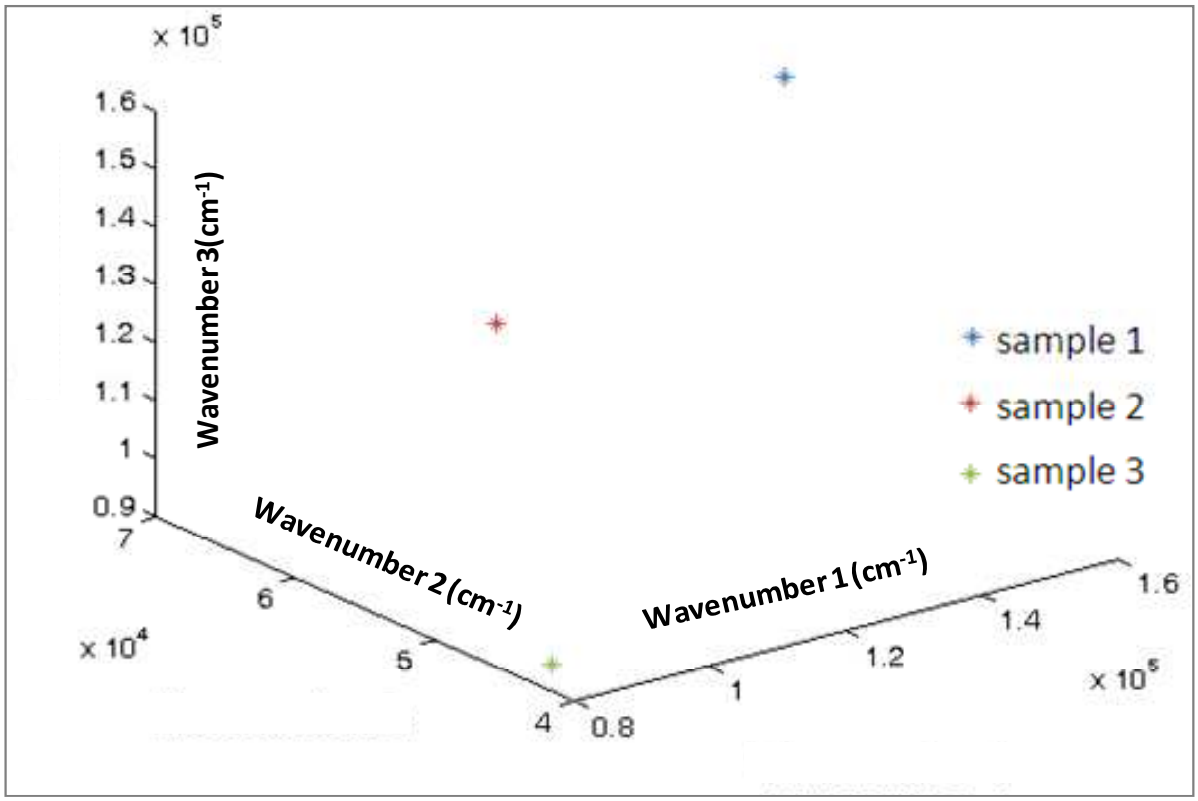


Fig. 2.5 Samples represented in 2-dimensional with the same samples represented by 1 point 3-dimensional space

An individual spectrum recorded on a spectrometer can have hundreds of data points and a single component can have a response in multiple places within the one region making the data highly correlated. Rather than representing the spectral data in 2-dimensional space, chemometric techniques use multi-dimensional space or hyperspace to represent the same spectrum by a single point. As there is usually much redundant information in spectra due to variables being highly correlated, data does not need to be represented in space with as many dimensions as there was original data points. The spectral data containing hundreds of data points can be fully characterised

in as few as twenty dimensions.[20] Chemometric or multivariate calibration techniques allow the concentration of a given analyte to be related to spectral features. They are also useful for distinguishing real chemical information from instrument noise.

2.6.1 Pre-treatments

Prior to analysing spectral data, a mathematical pre-treatment may be necessary. Common pre-treatments include mean centering, mean normalisation and using the first or second derivative of the spectra. To perform mean centring on a data set, the mean spectrum of the set is computed and then is subtracted from each spectrum in the set. This is done to prevent data points that are further from the origin from exerting an undue amount of leverage over the points that are closer to the origin.[20] Leverage is a measure of how extreme a data point is compared to the majority. A data point with high leverage will have a high influence on any model created. Mean normalisation is an adjustment to a data set that equalizes the magnitude of each sample. When the spectra have been normalized, qualitative information that distinguishes one sample from another is retained but information that would separate two samples of identical composition but different concentration is removed. A standard normal variate (SNV) pre-treatment is one which centres and scales individual spectra. The effect of this pre-treatment is that on the vertical scale each spectrum is centred on zero and varies roughly from -2 to +2. This effectively removes the multiplicative interferences of scatter and particle size in spectral data.[21]

The first derivate of a spectrum is the slope of the curve at every point. It has peaks where the original has maximum slope and crosses zero where there was a peak in the original spectrum. As the slope is not affected by additive baseline offsets in the

spectrum; calculating the first derivative is an effective method of removing baseline effects. The second derivative is the slope of the first derivative. It has peaks in roughly the same places as the original spectrum but these peaks are in the inverted direction. Calculating the second derivative of a spectrum will remove additive baseline effects and as well as multiplicative baseline effect.[22, 23] Small spectral differences are enhanced and overlapping peaks are separated by the use of derivative pre-treatments.

As a measured spectrum is not a continuous mathematic curve, but rather a series of equally-spaced points, traditional derivative calculation performed by using the difference in values between two adjacent points has the effect of reducing the signal to noise ratio in the data. It is necessary therefore to include some form of smoothing in the calculation. One method of calculating the derivate of spectra is to use the algorithm described by Savitzky and Golay.[24] This works by taking a narrow window centred at the wavelength of interest, and fitting a low order polynomial to the data points in this window using least squares. The calculated polynomial is a continuous curve of the form $y=a+bx+cx2$...where x is the wavelength and y is the spectral response. The first and second derivative of this fitted curve are then used as estimate of the derivatives of the underlying spectrum.

The choice of pre-treatment can depend on the type of spectra being analysed e.g. NIR will often have derivative pre-treatments applied.[25-27] A multi-component mixture or a sample collected on-line which may be subject to instrument drift will also be pre-treated with a procedure such as SNV or derivatives. Many other pre-treatments are

possible and the nature of the application will dictate the most suitable one or indeed combination to choose.

2.6.2 Quantitative analysis

Interpretation of spectra can be a challenge as many different components can have a response in similar regions of the electromagnetic spectrum. This becomes an issue when you want to indentify and quantify individual components in a mixture. The first step in developing a calibration model is to do a simple feasibility study such as that described in the ASTM international standards[28] for each component of interest. The procedure described involves the collection of spectra from 30-50 samples incorporating the expected variations in particle size, sample presentation, and process conditions which are expected during analysis. If the results from this simple study are favourable as judged by error values from cross validation methods and the precision required was obtained, the study can be expanded to see if multi-component mixtures can be adequately modelled.

In order to make a good calibration model, a suitable experimental design must be employed. The samples used for developing the model are known as the training or calibration set and should ideally comprise several uniformly distributed concentrations for each component of interest. The factors in an experimental design for a multi-component mixture are the individual components and these factors should be mutually independent or orthogonal, i.e. the correlation coefficient between each pair of factors is zero.[29] There has been some discussion in the literature on the importance of using uncorrelated samples in the development of chemometric models for on-line metabolite monitoring.[30-33] As the performance of any model is directly

affected by the training set used in its development, the training set should fulfil certain criteria. It should:

- contain all expected components
- span the concentration ranges of interest
- span the conditions of interest
- contain mutually independent samples

The calibration should also be validated using a set of samples (validation set) which is independent of the training set. Strategies on how to determine an experimental design which will achieve these aims can be found elsewhere.[29, 34, 35]

2.6.3 Partial least squares regression

An often used chemometric calibration technique for bioprocessing applications is partial least squares regression (PLS). This is a multivariate statistical technique developed from classical least squares and inverse least squares regression by Swedish statistician Herman Wold for use in economic forecasting. His son Svante Wold along with other Scandinavian scientists including Harold Martens promoted its use in chemical applications.[34]

As mentioned earlier, spectra can be described by wavenumbers and responses in 2-dimensional space or as single points in hyperspace. In this way PLS works as a variable reduction system and new axes in hyperspace are computed using both the chemical and spectral data. These new axes are computed in the direction of the most variance within the data and with PLS, the axis is the best compromise between the spectral variance and the concentration variance.

In an ideal case of noise free spectra, the factor space for the spectral data and the corresponding factor space for the concentration data are congruent i.e. the scores of the spectral data points are proportional to the corresponding eigenvector of the scores of the concentration data points. This relationship can be expressed by equation 2.3, where Y_f is projection or score of a single concentration point onto the f^{th} concentration factor, X_f is the score of a single spectral point onto the f^{th} spectral factor and B_f is the proportionality constant for the f^{th} pair of concentration and spectral factors.

$$Y_f=B_fX_f \tag{2.3}$$

The aim of PLS is to find a vector W that represents the best compromise between the spectral factor and the concentration factor. This vector is a factor that maximises the covariance between the concentration data matrix and all possible linear functions of the spectral data matrix. The factor W will have the same number of elements as there were wavelengths in the original spectra and the elements are called the loading weights. The first vector W_1, is the most significant optimum factor and the portion of the variance in the spectral data spanned by this factor is removed as is the spanned variance in the concentration data. The next factor W_2 is found for the spectral and concentration residuals that are not spanned by W_1. The process is continued until all possible factors have been found. The first new axis or factor is the most significant and accounts for the largest amount of variance in the spectral and concentration data. A graphical representation of this can be seen in figure 2.6.[20]

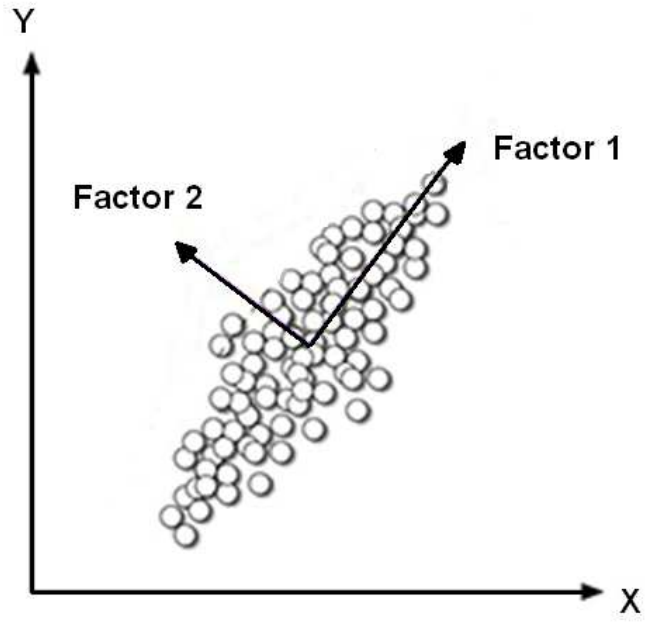


Fig. 2.6 Graphical representation of PLS

In physical terms PLS assumes that there are errors in both blocks which are of equal importance. The concentrations used in a calibration model are subject to error (e.g. dilution and weighing) just as much as the spectra or chromatograms. An important feature of PLS is that it is possible to determine how well the data have been modelled either by using *x (spectral data)* or *y (concentration data)* blocks. Fig. 2.7 illustrates the change in training set error as different numbers of components are calculated for both *x* and *y* in a typical dataset. This means that two different answers for the optimal number of components can be obtained, one based on the spectral data and the other based on the concentration data.[34]

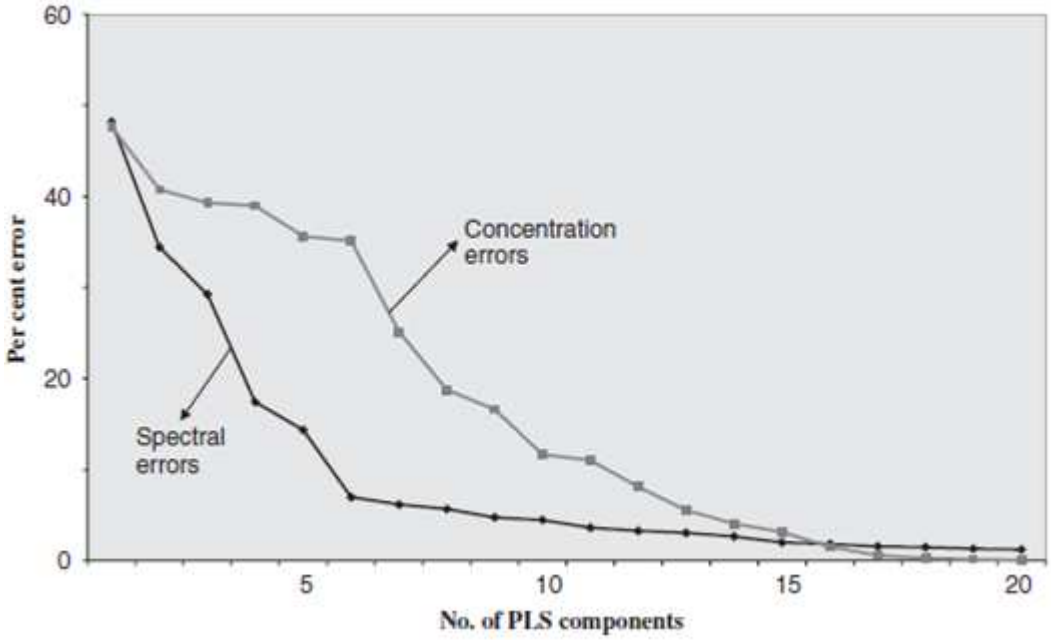


Fig. 2.7 PLS training set errors using both the concentration and spectral estimates[34]

2.7 PAT applications of vibrational spectroscopy in bioprocessing

The applications or potential applications of vibrational spectroscopy in bioprocessing are largely dependent on the sampling interfaces available. A number of instruments exist and sample interfaces vary from sample cavities using cuvettes or vials to immersion probes. Where real-time data is required for monitoring and control purposes, the type of available instruments is very much reduced as all off-line techniques are eliminated. Bioprocess applications to date have either used flow cells, where the sample of interest is passed through a measuring chamber, or immersion probes, where a probe is inserted into a reactor and the sample is scanned *in-situ* by transflectance, transmission or reflectance methods. The development of high quality fibre-optics and autoclavable probes has increased the capabilities of these techniques. The most common applications in bioprocessing are analyte, metabolite and biomass monitoring, with monitoring systems in some cases further developed to enable process control.

2.7.1 MIR applications

MIR lags behind its infrared counterpart, NIR, when it comes to its applications in bioprocessing. Despite the fact that MIR can detect and quantify components in aqueous solutions at significantly lower levels than that of NIR,[27] MIR is less extensively used. Only in the last decade has MIR been considered a potentially useful tool for bioprocess monitoring. Work to date has mainly focused on detection of substrates and metabolites in yeast and bacterial cultures but it has also been applied to suspended and immobilised animal cell cultures.[36] Most methods use synthetic samples or samples taken from cell cultures to build multivariate models capable of predicting changing concentration values.

The most common component modelled is glucose. This is the predominant substrate in cell culturing and so, is of most interest from a detection and monitoring point of view.[37-39] Other substrates detected using on-line MIR techniques include fructose, lactose, galactose, ammonia and methyl oleate. [40-42] Accuracy values vary between studies with standard prediction errors ranging from 0.26g/L to 0.86g/L for glucose. Subtle differences exist between the various techniques developed. The sample presentation method is of some importance for this application as many cell cultures require aeration resulting in gas bubbles forming on the probe tip. Automated flow systems can help mitigate this problem, while the recessed geometry of the probe tip can facilitate the formation of pockets on the crystal surface.[27] In addition to the sampling interface, the models employed are specific to each individual set up. Although multivariate chemometric modelling is used to develop these models, each model is unique.

This technique has also been applied to determine the rate of product formation. Cell culture products that have been successfully detected using MIR include ethanol, lactic acid and glucuronic acid.[38, 39, 43]

On-line MIR measurements have been used not just to detect or monitor cell culture substrates and metabolites, but also to control cultures. Kornmann et al used *Gluconacetobactor xylinus* to develop a control strategy based on the depletion of two substrates, fructose and ethanol.[44] Real-time spectroscopic scans were collected every 5 minutes, concentrations were sent to an adaptive control algorithm and fructose and ethanol were fed to the culture in controlled volumes. Schenk et al showed that a similar system could be used to control methanol feeding to *Pichia pastoris* cultures.[12]

2.7.2 NIR applications

NIR spectroscopy can provide on-line information on substrate, biomass, product and metabolite concentrations.[45, 46] This information can be further used to control and optimise cell cultures. Extensive work has been carried out in this area to date. NIR has been used to monitor concentration changes in yeast, bacterial and even mammalian cell cultures. Crowley et al used NIR to monitor the main substrates, glycerol and methanol, as well as biomass, in a *Pichia pastoris* culture,[46] Petersen et al used NIR to predict the changing concentrations of glucose, ammonium and biomass in a *Streptomyces coelicolor* culture,[31] while Rodrigues et al developed an NIR model to monitor clavulanic acid, the product of a fed batch process with *S. clavuligerus.*[47]

The technique has also been applied to monitoring of mammalian cell cultures. Four key analytes of a CHO-K1 mammalian cell culture, glucose, lactate, glutamine and ammonia, were monitored by Arnold *et al.* [48] and this work was further developed by Roychoudhury *et al.* [49] where a multiplexed calibration technique was used.

As with MIR, NIR predictive models have also been applied to control systems in order to allow fed-batch cultures to react in "real time". As early as 1994 Vaccari *et al*. proposed using NIR to control the glucose feed in the production of lactic acid by *Lactobacillus casei.*[50] Many others have developed control strategies for various yeast and microbial cultures. [51, 52]

2.7.3 Raman applications

The reported use of Raman spectroscopy for monitoring bioprocesses *in-situ* and in real time is limited and this is most likely due to the need for low frequency lasers to avoid fluorescence which can have heating effects due to the long exposure times necessary for such lasers. Most reported studies describe the use of Raman spectroscopy to monitor yeast cultures. One of the earliest applications of *in-situ* Raman spectroscopy was monitoring the production of ethanol in yeast fermentations.[53] In this study the concentrations of fructose and glucose were also measured. Shaw *et al.* used a dispersive Raman instrument to monitor the change in substrate and metabolite concentrations as well as product formation in yeast fermentation and found it necessary to include a by-pass filter to remove cells as they were causing interference to the photon scattering process.[54] The production of carotenoids in *Phaffia rhodozyma* cultures has also been monitored by dispersive Raman spectroscopy.[55] Bacterial cultures have also been monitored with *in-situ*

measurements of glucose, acetate, formate, lactate and phenylalanine being reported.[56] In a more recent study Raman spectra have been collected *in-situ* in a mammalian cell bioreactor. As well as monitoring substrates and metabolites, the spectra were correlated to total cell density and viable cell density showing that it may be possible for Raman spectroscopy to distinguish between live and dead cells.[57] While these studies all demonstrate the potential of Raman spectroscopy as a monitoring tool, it has yet to be proved capable of control in industrial bioprocesses.

Although separate techniques, both MIR and NIR have similar applications in bioprocessing; both have been used for monitoring and control purposes. Raman spectroscopy has been used to monitor bioprocesses but to a lesser degree than the other vibrational spectroscopies. The manner in which these techniques are exploited is similar. In all cases multivariate chemometric models are developed based on synthetic, semi-synthetic or actual samples from a cell culture. Typically these models are then validated and applied to a culture on-line. These techniques all have their benefits and limitations, but to date NIR has been the subject of more investigation and as a result is more developed in terms of applications in bioprocessing. However, the potential of MIR and Raman should not be underestimated or overshadowed.

2.8 Conclusions

Choosing a suitable on-line analytical technique and data processing method for bioprocess applications is essential if reliable monitoring and control are to be achieved. Each of the process analysers described here has the potential to be used for on-line measurement but it is only through proper understanding of their specific

advantages and limitations that they can be applied to monitor the appropriate process variables. The relationship between the measureable parameters and critical process parameters needs to be recognised in order to develop calibrations for the critical process parameters of interest and knowledge of the likely signal interferences will allow the employment of data treatments which can minimise or even eliminate their effects. Advanced data processing methods such as data reconciliation and artificial neural networks can also enhance the accuracy of the measured variables by using inputs from a number of on-line sensors. The combination of suitable analytical techniques and data processing methods should provide an increase in bioprocess knowledge which will in turn allow the process to be tightly controlled and operate within a previously established design space.

2.9 References

1. U.S. FDA. Pharmaceutical cGMPs for the 21st Century - A Risk-Based Approach - Final Report. (2004).

2. U.S. FDA. Guidance for Industry, Q8 Pharmaceutical Development. (2006).

3. U.S. FDA. in *Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance* , 2004).

4. Baughmann, E. in *Process analytical technology: Spectroscopic tools and implemenation strategies for chemical and pharmaceutical industries* (ed Bakeev, K. A.) 1 (Blackwell Publishing Limited, Oxford, 2005).

5. Günzler, H. & Gremlich, H. in *IR spectroscopy: An introduction* 309 (Wilcy-VCH, Weinheim, 2002).

6. Brown, J. M. in *Molecular spectroscopy* (Oxford University Press, New York, 1998).

7. Barańska, H. in *Laser Raman spectroscopy* (eds Barańska, H., Łabudzińska, A. & Terpiński, J.) (Ellis Horwood Limited, Warsaw, 1987).

8. Jestel, N. J. in *Process analytical technology: Spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries* (ed Bakeev, K. A.) (Blackwell Publishing Limited, Oxford, 2005).

9. Bakeev, K. in *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries* (ed Bakeev, K. A.) (Blackwell Publishing Limited, Oxford, 2005).

10. Schenk, J., Viscasillas, C., Marison, I. W. & von Stockar, U. On-line monitoring of nine different batch cultures of E. coli by mid-infrared spectroscopy, using a single spectra library for calibration. *J. Biotechnol.* **134**, 93-102 (2008).

11. Schenk, J., Dabros, M., Marison, I. W. & von Stockar, U. Simple and quick in-situ calibration of a FTIR instrument to control fed-batch fermentations of Pichia pastoris. *J. Biotechnol.* **118**, S37-S37 (2005).

12. 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* **128**, 344-353 (2007).

13. Gabriele, R. Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Adv. Drug Deliv. Rev.* **57**, 1109-1143 (2005).

14. Singh, R. C. V. Raman and the Discovery of the Raman Effect. *Physics in Perspective (PIP)* **4**, 399-420 (2002).

15. Das, R. S. & Agrawal, Y. K. Raman spectroscopy: Recent advancements, techniques and applications. *Vibrational Spectroscopy*.

16. Chase, B. A New Generation of Raman Instrumentation. *Appl. Spectrosc.* **48**, 14A-19A (1994).

17. Clementschitsch, F., Jürgen, K., Florentina, P. & Karl, B. Sensor combination and chemometric modelling for improved process monitoring in recombinant E. coli fed-batch cultivations. *J. Biotechnol.* **120**, 183-196 (2005).

18. Vankeirsbilck, T. *et al*. Applications of Raman spectroscopy in pharmaceutical analysis. *TrAC Trends in Analytical Chemistry* **21**, 869-877 (2002).

19. Massart, D. L., Vandeginste, B. G. M., Deming, B. M., Michotte, Y. & Kaufman, L. in
Chemometrics: a textbook. Data handling in science and technology (Elsevier, Amsterdam, 1988).

20. Kramer, R. in *Chemometric Techniques for Quantitative Analysis* 216 (Dekker, New York, 1998).

21. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* **43**, 772-777 (1989).

22. Naes, T., Isaksson, T., Fearn, T. & Davies, T. in *A User-Friendly Guide to Multivariate Calibration and Classification* 344 (NIR Publications, Chichester, UK, 2002).

23. Gabrielsson, J. *et al*. OPLS methodology for analysis of pre-processing effects on spectroscopic data. *Chemometrics Intellig. Lab. Syst.* **84**, 153-158 (2006).

24. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* **36**, 1627-1639 (1964).

25. Abu-Absi, N. R. *et al*. Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. *Biotechnol. Bioeng.* **108**, 1215-1221 (2011).

26. Henriques, J., Buziol, S., Stocker, E., Voogd, A. & Menezes, J. Monitoring Mammalian Cell Cultivations for Monoclonal Antibody Production Using Near-Infrared Spectroscopy. **116**, 73-97 (2010).

27. Landgrebe, D. *et al*. On-line infrared spectroscopy for bioprocess monitoring. *Applied Microbiology and Biotechnology* **88**, 11-22 (2010).

28. ASTM. in *Annual Book of ASTM Standards* , Philadelphia, PA, 2004).

29. Brereton, R. G. Multilevel multifactor designs for multivariate analysis. *Analyst* **122**, 1521-1529 (1997).

30. Petiot, E. *et al*. In situ quantification of microcarrier animal cell cultures using near-infrared spectroscopy. *Process Biochemistry* **45**, 1427-1431 (2010).

31. Petersen, N. *et al*. In situ near infrared spectroscopy for analyte-specific monitoring of glucose and ammonium in streptomyces coelicolor fermentations. *Biotechnol. Prog.* **26**, 263-271 (2010).

32. Schenk, J., Marison, I. W. & von Stockar, U. Simplified Fourier-transform mid-infrared spectroscopy calibration based on a spectra library for the on-line monitoring of bioprocesses. *Analytica Chimica Acta* **591**, 132-140 (2007).

33. Rhiel, M. H., Amrhein, M., 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. *Anal. Chem.* **74**, 5227-5236 (2002).

34. Brereton, R. G. in *Applied Chemometrics for Scientists* (John Wiley & Sons Ltd, Chichester, UK, 2007).

35. Munoz, J. A. & Brereton, R. G. Partial factorial designs for multivariate calibration: extension to seven levels and comparison of strategy. *Chemometrics Intell. Lab. Syst.* **43**, 89-105 (1998).

36. 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. *Biotechnol. Bioeng.* **77**, 174-185 (2002).

37. Doak, D. L. & Phillips, J. A. In Situ Monitoring of an *Escherichia coli* Fermentation using a Diamond Composition ATR Probe and Mid-infrared Spectroscopy. *Biotechnol. Prog.* **15**, 529-539 (1999).

38. Franco, V. G., Perín, J. C., Mantovani, V. E. & Goicoechea, H. C. Monitoring substrate and products in a bioprocess with FTIR spectroscopy coupled to artificial neural networks enhanced with a genetic-algorithm-based method for wavelength selection. *Talanta* **68**, 1005-1012 (2006).

39. Mazarevica, G., Diewok, J., Baena, J. R., Rosenberg, E. & Lendl, B. On-line fermentation monitoring by mid-infrared spectroscopy. *Appl. Spectrosc.* **58**, 804-810 (2004).

40. Schenk, J., Viscasillas, C., Marison, I. W. & von Stockar, U. On-line Monitoring of Nine Different Batch Cultures of E.coli by Mid-infrared Spectroscopy, Using a Single Spectra Library for Calibration. *Journal of Biotechnology* (2007).

41. Roychoudhury, P., Harvey, L. M. & McNeil, B. At-line monitoring of ammonium, glucose, methyl oleate and biomass in a complex antibiotic fermentation process using attenuated total reflectance-mid-infrared (ATR-MIR) spectroscopy. *Analytica Chimica Acta* **561**, 218-224 (2006).

42. Dabros, M., Amrhein, M., Bonvin, D., Marison, I. W. & von Stockar, U. Data reconciliation of mid-infrared and dielectric spectral measurements for improved on-line monitoring of bioprocesses. *Journal of Biotechnology* (2007).

43. Fayolle, P., Picque, D. & Corrieu, G. On-line monitoring of fermentation processes by a new remote dispersive middle-infrared spectrometer. *Food Control* **11**, 291-296 (2000).

44. 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* **113**, 231-245 (2004).

45. Cervera, A. E., Petersen, N., Lantz, A. E., Larsen, A. & Gernaey, K. V. Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnology Progress* **25**, 1561-1581 (2009).

46. Crowley, J., Arnold, S. A., Wood, N., Harvey, L. M. & McNeil, B. Monitoring a high cell density recombinant Pichia pastoris fed-batch bioprocess using transmission

and reflectance near infrared spectroscopy. *Enzyme and Microbial Technology* **36**, 621-628 (2005).

47. Rodrigues, L., Vieira, L., Cardoso, J. & Menezes, J.
The use of NIR as a multi-parametric in situ monitoring technique in filamentous fermentation systems. *Talanta* **75**, 1356 (2008).

48. 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. *Biotechnol. Bioeng.* **84**, 13-19 (2003).

49. 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. *Anal. Chim. Acta* **590**, 110-117 (2007).

50. Vaccari, G. *et al*. A near-infrarod spectroscopy technique for the control of fermentation processes: An application to lactic acid fermentation. *Biotechnol. Bioeng.* **43**, 913-917 (1994).

51. González-Vara y R., A. *et al*. Enhanced production of L-(+)-lactic acid in chemostat by Lactobacillus casei DSM 20011 using ion-exchange resins and cross-flow filtration in a fully automated pilot plant controlled via NIR. *Biotechnol. Bioeng.* **67**, 147-156 (2000).

52. Cimander, C. & Mandenius, C. Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling. *J. Chem. Technol. Biotechnol.* **77**, 1157-1168 (2002).

53. Gomy, C., Jouan, M. & Dao, N. Q. Methode d'analyse quantitative par spectrometrie Raman-laser associee aux fibres optiques pour le suivi d'une fermentation alcoolique. *Anal. Chim. Acta* **215**, 211-221 (1988).

54. Shaw, A. D. *et al*. Noninvasive, On-Line Monitoring of the Biotransformation by Yeast of Glucose to Ethanol Using Dispersive Raman Spectroscopy and Chemometrics. *Appl. Spectrosc.* **53**, 1419-1428 (1999).

55. Cannizzaro, C., Rhiel, M., Marison, I. & von Stockar, U. On-line monitoring of Phaffia rhodozyma fed-batch process with in situ dispersive raman spectroscopy. *Biotechnol. Bioeng.* **83**, 668-680 (2003).

56. 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* **35**, 131-137 (2004).

57. Abu-Absi, S. F. *et al*. Defining process design space for monoclonal antibody cell culture. *Biotechnology and Bioengineering* **106**, 894-905 (2010).

Chapter 3

Study 1: Potential of Mid-infrared spectroscopy for on-line monitoring of mammalian cell culture medium components

As published in Applied Spectroscopy, 2012, Vol. 66 (1), 33-39

Abstract

This study proposed a methodology to evaluate the potential of mid-infrared spectroscopy (MIR) as a process analytical technology (PAT) tool for in-situ (in-line) monitoring of cell culture media constituents, paving the way for on-line bioprocess monitoring and control of mammalian cell cultures. The methodology included a limit of detection (LOD) analysis and external influence investigation in addition to the calibration model development. The LOD analysis in the initial step provided a detailed procedure by which to evaluate the monitoring potential of the instrument of choice, for the application in question. The external influence study highlighted the potential difficulties when applying this technique to a typical mammalian cell culture.

A comparative investigation between a fixed conduit immersion probe and flexible fibre-optic immersion probe was also carried out. Limitations associated with the use of MIR in the cell culture environment were also examined. A preliminary investigation, on components typically found in mammalian cell cultures, involving spectral characterisation and limit of detection analysis was completed. It was evident at this initial stage that glutamine, could not be accurately detected at levels typically found in a mammalian cell culture medium. Results for glucose and ammonia,

however, proved promising. A 7-concentration level experimental design was used, and partial least squares regression employed, to develop calibration models. Optimized model results echoed the results of the preliminary analysis with the percentage error of prediction for glucose as low as 6.03% with the fixed conduit probe, and glutamine having a higher error of 63.06% for the same probe. Comparison of model results obtained from both probes supported the fixed conduit as the more accurate of the two probes, for this experimental set up. The effect of external influences on the MIR spectra and hence the concentrations predicted by the model were also examined. These were subjected to statistical analysis to determine the significance of the effect. This study demonstrates that MIR as a PAT tool, has limited potential for mammalian cell culture monitoring due to low concentrations of analytes present and outlines a method to allow the system to be evaluated.

3.1 Introduction

With increasing pressure from regulatory authorities on industry to develop processes embracing 'Quality by Design' initiatives, there is a growing demand to establish reliable tools and systems capable of meeting this need. [1, 2] With regard to monitoring and control of bioprocesses, this need translates to a search for robust instrumentation capable of monitoring the key process analytes and metabolites in real time. Such information could potentially be used in the development of process control tools and hence would meet the fundamental principles of 'Quality by Design' and 'Design Space'.[3] Achieving on-line or real-time measurement and control allows for instantaneous analysis of the results and correction of offsets before the process moves outside of its design space.[4] In a typical bioprocess a quantitative and qualitative analysis of all the major analytes in real time will provide vital information on the

process and facilitate the identification of key parameters capable of improving process outputs e.g. biomass, product secreted etc.[5] The first step in developing such a system lies in the identification of a reliable monitoring technique, which could further be used as an integral part of an advanced control system.

Infrared spectroscopy has the ability to monitor several of the analytes present in the culture media at any one time and as such is potentially a powerful tool in bioprocess monitoring.[5-7] Such techniques can be used in-situ (in-line) making them non-invasive and eliminating the need for sample removal thereby reducing the risk of culture contamination due to possible compromised sterility.[8] Sample preparation is not required and spectral information is obtained instantaneously. These features make infrared techniques suitable for inclusion in control systems developed to function within a Process Analytical Technology (PAT) environment, an initiative proposed by the FDA in 2004[9] and further supported by the International Conference on Harmonisation[10] in 2006. The spectral data gleaned from such sensors must undergo some form of multivariate analysis in order to extract the desired information.[11] With respect to the monitoring of bioprocesses, both the chemometric and infrared techniques serve as a lock and key to releasing bioprocess data. Both near infrared (NIR) and mid infrared (MIR) spectroscopy have been used for bioprocess monitoring but with mammalian cell culturing, sterility is critical, so in-situ probes are preferable.[12] In-situ sampling is possible with both techniques and is well documented for NIR,[12-17] but less so for MIR spectroscopy, when applied to mammalian cell culture media components.[18, 19]

A preliminary step to developing a multi-analyte calibration model is to identify the instrument detection limits for each of the analytes under investigation. The limit of detection (LOD) is the lowest concentration of analyte in a sample that can be detected but not necessarily quantified under the experimental conditions of the method.[20] Determining the LOD for a multivariate calibration is significantly more complex than establishing that of a univariate model. To date, few studies have included this preliminary step in similar works despite its importance in an initial feasibility study.

A number of external influences are likely to be present throughout the course of a cell culture. The effect of these influences on the collected spectra and the media concentration values predicted by the model are of interest when considering the implementation of a calibration model in a monitoring or control application. The significance of the impact of the external influence can be examined using hypothesis testing. This can provide an insight into the effect of changing environmental factors on cell culture measurements using MIR. It also highlights when and where influences should be incorporated into the model, so as to ensure optimum results.

The aim of this work was to outline a method to examine the potential of MIR as a PAT tool to measure the varying media component concentrations of mammalian cell cultures in real time. This was achieved by developing a partial least squares (PLS) calibration model using synthetic samples, which ensure that the model remained generic and non-specific to any single bioprocess. The methodology also included a logical sequence of preliminary steps, including a detailed limit of detection analysis, prior to the application of MIR to a mammalian cell culture. In addition, two probe

types were used to determine the effect, if any, on results when the sampling device was varied, but the detector remained the same.

3.2 Materials and Methods

3.2.1 Samples

Eight individual components made up of typical cell culture medium substrates, buffers and products of cell metabolism were examined. These were, glucose (Sigma Aldrich Ireland Ltd.), L-glutamine (Sigma Aldrich Ireland Ltd.), HEPES (Sigma Aldrich Ireland Ltd.); glutamate, (Oxoid Ltd.); calcium lactate pentahydrate (Fisher Scientific UK); sodium hydrogen carbonate (Fisher Scientific UK); potassium phosphate (Sigma Aldrich Ireland Ltd.) and ammonium sulphate (Fisher Scientific U.K.). Five of these (glucose, lactate, ammonia, glutamine and glutamate), are mammalian cell metabolites or by-products. Such metabolites and by-products are likely to vary considerably throughout the course of a cell culture and therefore are of most interest from a monitoring and potentially, control, standpoint. The remaining three elements, (HEPES, potassium phosphate and sodium hydrogen carbonate) are typically used to buffer a cell culture medium. These do not vary considerably over the course of a cell culture; however they may vary over a range of different media. In order to make the model generally applicable to a range of mammalian cell culture media these were included in its development. Although the precise concentration of these buffers is not known for each commercial medium they tend to be present at high enough concentrations to impact the spectrum and so were included in the model. Other potential elements likely to be present in the media e.g. amino acids/vitamins are at lower concentrations and hence the impact on the spectrum and on the prediction ability of the model is minimal. A cell culture medium, EX-CELL CHO DHFR$^-$

Medium AF (Sigma Aldrich, Ireland Ltd.), was used in model validation steps and external influence investigation. Table 3.1 outlines the concentration ranges used for model development for each of the 8 components selected.

Table 3.1 Component concentration ranges

Component	Concentration Range g/L
Glucose	0-6
Lactate	0-1
Ammonia	0-1
Glutamine	0-0.6
Glutamate	0-0.7
Phosphate	0-1
Sodium Bicarbonate	0-1.25
HEPES	0-4.5

3.2.2 Instrumentation

All scans were taken using a Fourier transform mid-infrared ReactIR iC10 instrument with MCT detector (Mettler Toledo AutoChem, Inc., Columbia, US). Two immersion probes were tested in all cases (i) AgX 9.5mm x 2m fibre silver halide probe and (ii) K6 conduit 16 mm probe, both of which have a fixed optical pathlength of approx. 1 – 2 µm and use a diamond ATR crystal with 6 internal reflections. Fig. 3.1 shows both probe types. The sampling procedure was as follows: a background scan of deionised water at 37^{o}C was taken followed by 3 replicates of 128 co-added scans of each sample also at 37^{o}C as this temperature is optimum for mammalian cell growth.

(i) (ii)

Fig. 3.1 (i) Flexible Fibre-Optic and (ii) Fixed Conduit Probes

3.2.3 Preliminary analysis

Concentrated solutions of each of the 8 main components were made up using deionised water. The spectra for each component were collected and then plotted. The wavenumber ranges over which each of the components absorbed was noted, to be later used in the development of the calibration model. Fig. 3.2 shows the raw spectra of the 8 components modelled and indicates their main regions of absorbance.

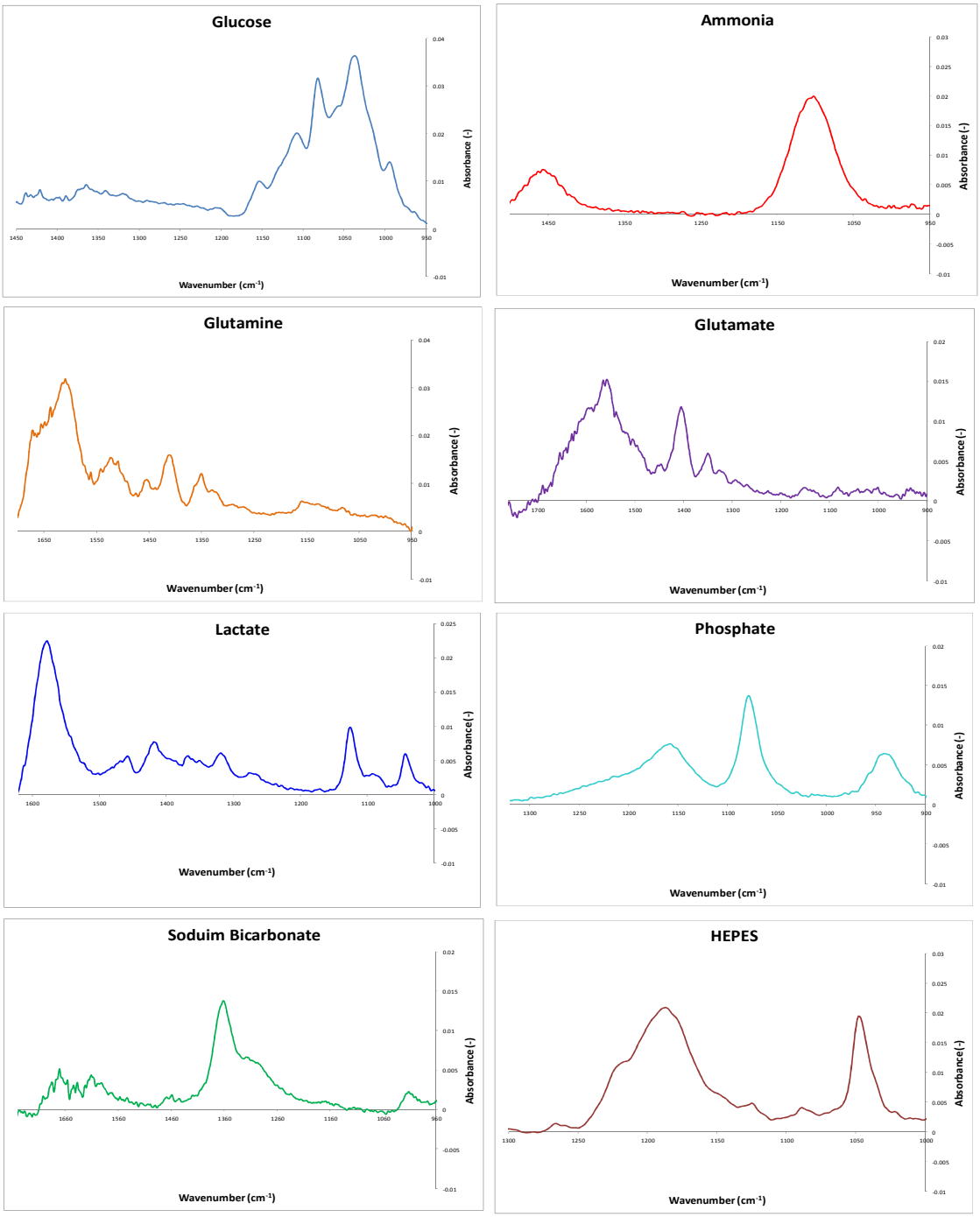


Fig. 3.2 Raw spectra of the 8 matrix components over the usable spectral regions

The concentrated stock solutions described above were used to prepare a series of dilute solutions. These were then scanned in triplicate, with the K6 conduit probe and the collected spectra plotted and examined. This allowed for a preliminary investigation into the presence of outliers and also an investigation into the LOD based on a simple visual test. The point where the component peaks could not be clearly distinguished from the instrument noise was defined as the observed LOD. This is only possible for MIR spectroscopy as it is based on fundamental vibrations and the peaks can be directly related to a chemical bond but such a step could not be done for NIR spectroscopy. Fig. 3.3 below indicates how peak height increases with increasing concentration.

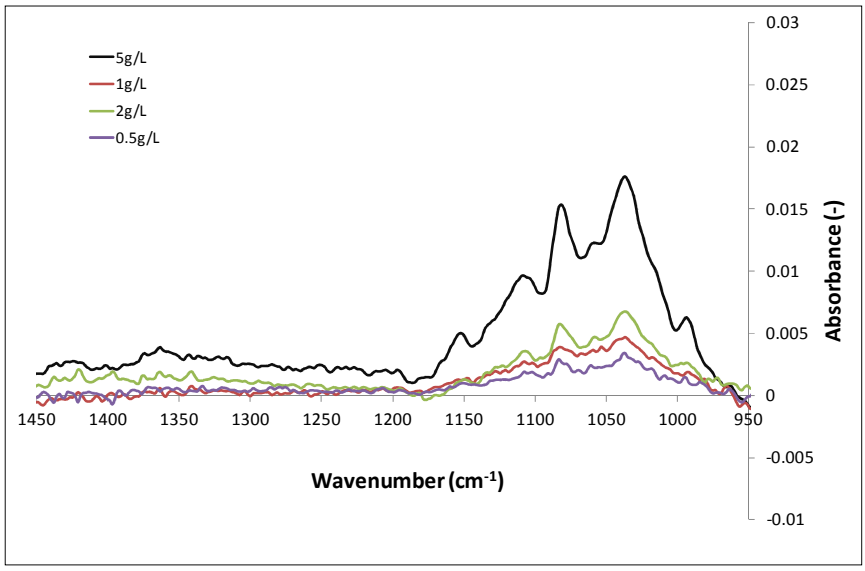


Fig. 3.3 Glucose spectra at varying concentrations

As multivariate analysis techniques are usually necessary to fully exploit vibrational spectroscopy data, a method for determination of LOD for multivariate spectral data is needed. Unfortunately there is no generally accepted method for this but an approach used by many authors is to transform the multivariate data to univariate data by the use

of chemometric techniques[21] such as PLS regression.[22] The method outlined by Ortiz *et al*[23, 24] is used here. It allows for the assessment of the LOD of an analytical method by evaluating false positive and false negative probabilities after data has been transformed by PLS. The false positive probability follows a Student's *t*-distribution with (*n*-2) degrees of freedom where *n* is the number of samples while the false negative probability has a non-central Student's *t*-distribution with (*n*-2) degrees of freedom and a parameter of non-centrality which needs to be established. Critical values for both types of Student's *t*-tests and the parameter of non-centrality were determined using the statistical toolbox in MATLAB (v7.9.0.529 (R2009b), The MathWorks Inc., Cambridge, UK). Development of PLS models was carried out using the PLS toolbox (V6.2 Eigenvector Research Inc, Washington, US) for MATLAB. In all cases the spectral data were mean centered and first and second derivative pre-treatments using a segment size of 15 points were calculated using quadratic Savitzky-Golay[25] filters. Statistical analysis was applied to the regression between actual concentration values and those predicted by leave-one-out cross validation.

3.2.4 Calibration model development

In order to develop a robust reliable model for a multivariate calibration, several concentrations are necessary, which require all possible combinations of levels for all the factors, i.e. a full factorial multilevel design. This is impractical as it would result in a huge number of samples, therefore, for this calibration model, a partial factorial design for a multivariate calibration was employed. A 7 concentration-level experimental design was chosen, which accounts for the orthogonality between successive factors. This resulted in the generation of 49 samples, each containing

varying concentrations of the 8 components.[26] Concentrated solutions of the 8 components were made and based on the experimental design, specified units of each solution were used to create each of the 49 samples. The concentration of each bulk solution was such that it reached the expected maximum concentration of that component in a typical cell culture medium. The purpose of this was to ensure that the calibration model under development catered for the entire concentration range of each of the 8 components in the medium, but at the same time did not compromise the accuracy of the model by using too wide a concentration range in its development. Each of the 49 samples was scanned with both probe types.

The spectral data were exported from the iCIR software (Mettler Toledo AutoChem, Inc., Columbia, US) and imported into MATLAB. Mean values of the triplicate scans were used for model development. Pre-processing in the form of mean-centering was initially applied to all spectra. A second derivative pre-treatment using a segment size of 15 points was also examined as second derivative pre-treatments can eliminate the effect of linear baseline spectral off-sets[27] that are likely to occur over the course of a culture. Partial least squares regression, optimized by leave-one-out cross-validation, was used to develop separate calibration models for each of the 8 components.

The calibration model was validated in two ways. Firstly, a 4-level multivariate design was employed. This was similar to the 7-level design used for the creation of the model, however only 16 samples of varying concentration were generated. There are a number of suggested validation methods in use but the “Rule of 3” is widely accepted as a suitable technique for evaluating the accuracy of a model.[22] This rule proposes that the calibration set be 3 times larger than the validation set. In this case the

calibration set, a 7-level design, resulted in the generation of 49 samples, therefore, the 4-level design, resulting in a sample set of 16 samples and almost one third the size, was chosen to validate the models. Each of these 16 samples was scanned with both probes as before. The spectral data resulting from these scans was then inputted into the model and the component concentrations in each of the 16 samples were predicted by the model. These predicted values were then compared to the actual values. Secondly the cell culture medium was spiked with known concentrations of different components and the model was used to predict the resulting increase in concentration. The results of all models were evaluated by the root mean square errors of calibration, cross validation and prediction (RMSEC, RMSECV and RMSEP) as well as the LOD for a multicomponent mixture.

3.2.5 External influence investigation

In order to establish if the model would accurately predict cell culture concentrations when exposed to the conditions typically found in a cell culture environment, a series of experiments were carried out. The effect of varying temperature, agitation, biomass, and pH were investigated. In addition, the presence or lack of antifoam and pluronic was also examined. For each external influence examined, a background of water was initially scanned within a bioreactor; the bioreactor was then drained and dried and filled with EX-CELL CHO DHFR$^-$ medium. The bioreactor environment was varied as outlined in Table 3.2. Both the agitation and temperature were varied by adjusting the appropriate settings on the reactor control system. The pH of the media was varied by the addition of hydrochloric acid and sodium hydroxide to create acidic and alkaline conditions respectively. The biomass was altered by simulating the growth of a cell culture during the exponential growth phase; with the cumulative

addition of a fixed cell density over time. Each variation in bioreactor conditions was scanned in triplicate using 128 co-added scans. These scans were ratioed against the initial background of deionised water.

TABLE 3.2 External influence investigation summary

Exp. #	Temperature (^{o}C)	pH (-)	Agitation (rpm)	Biomass (cells/ml)	Antifoam (% v/v)	Pluronic (g/L)
1	**30 – 44 ± 0.2**	7.9	150	-	-	-
2	37 ± 0.2	**5 - 10**	150	-	-	-
3	37 ± 0.2	7.9	**40 - 400**	-	-	-
4	37 ± 0.2	7.9	150	**0 - 4.5x10^{6}**	-	-
5	37 ± 0.2	7.9	150	-	**0.05**	-
6	37 ± 0.2	7.9	150	-	-	**0.1**

PLS regression was the multivariate technique employed to assess the effect of an influence when the influence was varied (temperature, pH, agitation, and biomass) and PLS-DA (Partial least squares discriminant analysis)[28] was the multivariate method used for examination if an influence impacted the spectra when present (antifoam and pluronic). The data was subjected to two tests, a spectral test and a relevance test. The spectral test involved the data undergoing a multivariate transformation followed by hypothesis testing. The relevance test examined the effect (if any), on the predicted glucose concentration values. Pre-treatments in the form of mean centering followed by second derivative using the Savitzky-Golay method were applied to all spectra prior to modeling, and in all cases, cross validation was used to optimize the models.

A PLS model was constructed to determine whether there was a relationship between the spectra and the external influence variables, temperature, pH, agitation, and biomass. Plots of predicted magnitude of influence applied versus actual values for each of the tests yielded the coefficient of determination (r^2) and hence the correlation coefficient (r) was calculated. Based on a Student's *t*-test, Pearson's correlation[29] was used to test the significance between the influence and the spectral measurement. The null hypothesis (H_0) was that the influence had no effect on the spectra. Where $t>t_{critical}$ at a significance level (α) of 0.05, H_0 was rejected and this implied that the external influence under investigation had a significant effect on the spectra. Where $t<t_{critical}$ at $\alpha=0.05$, H_0 was accepted and this implied that the external influence under investigation did not have a significant effect on the spectra.

The same statistical test was employed when determining the relevance of the impact of each influence on the predicted glucose concentration values. In this case, the optimum PLS model for glucose, established in the calibration model development section, was used to predict the glucose concentration. A plot of magnitude of external influence versus predicted glucose concentration provided r^2 values for each of the external influences, and as before, a Student's *t*-test was used to establish if the influence had a significant impact on the predicted glucose concentration values at α=0.05.

Multivariate discriminant analysis was used for tests involving antifoam and pluronic as these were not varied. In both cases concentrations typically used in cell culture media were added. A PLS-DA model was used to transform the data with respect to the presence or lack of influence applied. A Student's *t*-test was used to interpret the results of the PLS-DA and determine the significance of the effect of both antifoam and pluronic on the spectra. In this case, H_0 was that all samples were the same and the alternative hypothesis (H_a) was that samples with the influence applied were different. The same hypothesis test was also applied to the predicted glucose concentrations for the same sample set.

All experiments were carried out using both the AgX 9.5mm x 2m fibre silver halide probe (flexible) and the K6 conduit 16 mm probe (fixed).

3.3 Results and Discussion

3.3.1 Preliminary analysis

The spectrum of commercial media exhibited a number of peaks. The region of highest absorbance (1250 – 1000 cm^{-1}) is as a result of the overlapping absorbances of the constituent components. This highlighted the problems associated with attributing the absorbance to a particular component given the overlapping nature of the pure component spectra.

The spectra of concentrated solutions of each of the pure components were examined to determine the wavenumber ranges over which absorbance occurs. Concentrated stock solutions, within the water solubility limits of each of the components were used as all absorbance features may not be clearly evident in dilute solutions. The wavenumber ranges established at this stage are shown in Table 3.3. These were used when developing and applying the calibration model. This ensured that the predictions were based on the underlying chemical principle and not metabolism induced correlations.[19, 30]

The observed LOD for all components can be found in Table 3.3 alongside the LOD determined using a PLS model which has been subjected to hypothesis testing ensuring false positive and false negative probabilities of 0.05. The PLS model details used in the calculation of these LODs have also been included in Table 3.3. The observed and calculated LODs differ by an order of magnitude for all components with the exception of glutamine and glutamate which differ by a factor of 4. This shows that a visual inspection is not sufficient to determine such a parameter and more

information can be gleaned using a chemometric technique which is as expected for a multivariate data set. Glucose, ammonia, phosphate, lactate, HEPES and bicarbonate could all be detected to levels of 0.09 g/L or lower. Glutamine and glutamate were not detected to as low a level as the other components. This is most likely due to the fact that the molecular bonds present in these components tend to have weak absorbances in the MIR region detectable with the instrument used in this study.[31] These LOD values are based on pure component solutions and are only applicable to single component measurements. This is a simple feasibility study and it establishes at an early stage the ability of the MIR instrument to measure the components at the low levels found in mammalian cell cultures.

TABLE 3.3 Preliminary analysis results

Component	Wavenumber Range (cm^{-1})	Pre-treatment	# LVs	Calculated LOD (g/L)	Observed LOD (g/L)
Glucose	950-1450	Mean centered	5	0.09	1.0
Glutamine	950-1700	Mean centered, 1st derivative	6	0.30	1.2
Ammonia	950-1500	Mean centered, 1st derivative	2	0.01	0.1
Phosphate	900-1320	Mean centered	5	0.03	0.3
Glutamate	900-1760	Mean centered	5	0.35	1.3
Lactate	1000-1620	Mean centered	6	0.03	0.5
HEPES	1000-1250	Mean centered, 1st derivative	6	0.05	0.6
Bicarbonate	960-1750	Mean centered, 1st derivative	2	0.06	0.5

3.3.2 Calibration model development and validation

As stated previously, calibration model development was completed using both probes, the K6 conduit fixed probe and the flexible fibre-optic probe. The detector used in both cases was the same. Previous studies have used fixed conduit ATR probes[18, 32] and discussed the importance of alignment of such probes as well as spectral differences which can occur when the alignment is changed.[33-35] The development of infrared fibre-optic immersion probes is relatively new,[13] in particular for mid-infrared probes;[36] hence the evaluation of such a probe for the monitoring of mammalian cell culture media components is quite relevant. The fibre-optic configuration should eliminate the alignment issues inherent in the fixed conduit configuration.

The results obtained using both probe types varied considerably and as the experimental design and detector were unchanged, a comparison between both probes was carried out in an attempt to establish possible causes for the differing results. The MIR range is generally considered to lie between 4000-400 cm^{-1}. Both probes have a reduced effective range due to absorbance of the diamond ATR crystal over the range of 2250-1950 cm^{-1}. In addition the fibre-optics of the flexible probe themselves absorb infrared radiation, further reducing the effective range of this probe to wavenumbers less than 1950 cm^{-1}. While the K6 conduit fixed probe measured absorbance between 4000-2250 cm^{-1}, spectra in this region were extremely noisy and therefore unusable. At lower wavenumbers and also, close to the ATR absorbance region, the spectra exhibited a large degree of noise, so this meant that the two probes under investigation had the same usable wavenumber range (1800-900 cm^{-1}) therefore any discrepancies in results between the two probes cannot be attributed to differing wavenumber ranges.

The most significant difference between both probes is the geometry of the probe tip. The K6 conduit probe contains a recessed diamond crystal. The outer probe casing forms a gradual slope to the ATR crystal, thus allowing for shearing of bubbles, which may adhere to the crystal surface. The silver halide fibre-optic probe also contains a recessed diamond crystal, however in this case the outer casing does not form a gradual slope, but rather the crystal is set at a 90^{o} angle to the casing and hence a "pocket" at the probe tip allows for the entrapment of bubbles. Removing bubbles from the tip of this probe proves more difficult as the shear forces at the probe tip do not reach the bubble trapped inside. In addition, the high surface tension of water results in bubbles in aqueous solutions adhering to the probe tip,[21] making this application (to an aqueous based cell culture media), more problematic than typical applications in reaction chemistry. Fig. 3.4 shows the problems encountered with bubble entrapment.

Fig. 3.4 Air bubble on probe tip

Calibration model results for all samples are shown in Table 3.4. These models are based on calibration samples which contain varying amounts of all 8 components of interest. In the majority of cases, the RMSEC values are lower for the K6 conduit probe than those for the silver halide fibre-optic probe. All the RMSEC values are less than 0.25 g/L, and in the case of ammonia, it was as low as 0.02 g/L. Different concentrations of each of the 8 components were used, corresponding to typical values in cell culture media. In order to put the RMSEC in perspective, a percentage error of calibration (PEC) was calculated for each of the models, by dividing the RMSEC by the average concentration used in the calibration. While the RMSEC for glutamate and glutamine appear quite low for both probes, they represent quite high percentage errors, indicating that problems are likely with the prediction ability of the glutamate and glutamine models. The percentage errors of cross validation and prediction, (PECV and PEP), calculated in a similar manner to the PEC, were greater than 60% for glutamate and glutamine, thereby indicating that the predicted concentrations of glutamate and glutamine are not reliable. This supports the results of the preliminary analysis, where the LODs were higher than those of the other components and close to the maximum concentrations typically found in a cell culture environment. Based on this information, these components models were not further validated using the spiking test.

All other components had lower percentage errors than glutamine and glutamate with glucose and ammonia showing the lowest percentage errors of approximately 15% or less regardless of probe type or validation set used (Table 3.5). A plot of predicted glucose concentration versus actual glucose concentration as measured with the K6

conduit probe can be seen in Fig. 3.5. This model had low percentage errors and it can be seen that all samples lie close to the 1:1 calibration line.

TABLE 3.4 Calibration model results

Component	Probe	# LVs	RMSEC	PEC	RMSECV	PECV
Glucose	K6 conduit	4	0.25	8.64	0.34	11.65
	Fibre-optic	5	0.31	10.38	0.44	14.81
Glutamine	K6 conduit	12	0.02	7.48	0.24	77.24
	Fibre-optic	3	0.12	40.07	0.25	82.89
Ammonia	K6 conduit	3	0.02	4.49	0.03	5.91
	Fibre-optic	4	0.03	5.04	0.04	7.16
Phosphate	K6 conduit	4	0.14	30.38	0.19	40.60
	Fibre-optic	5	0.11	24.28	0.15	32.82
Glutamate	K6 conduit	6	0.09	26.43	0.22	62.88
	Fibre-optic	3	0.12	35.39	0.27	77.68
Lactate	K6 conduit	4	0.04	9.30	0.06	13.84
	Fibre-optic	3	0.20	45.37	0.24	54.60
HEPES	K6 conduit	4	0.21	9.94	0.29	13.86
	Fibre-optic	4	0.21	10.23	0.27	13.05
Bicarbonate	K6 conduit	8	0.07	12.25	0.23	37.26
	Fibre-optic	5	0.12	19.05	0.31	51.39

TABLE 3.5 Validation results

Component	Probe	RMSEP matrix	PEP matrix	RMSEP spiked	PEP spiked	LOD
Glucose	K6 conduit	0.17	6.03	0.41	13.53	0.41
	Fibre-optic	0.35	12.42	0.24	15.40	0.80
Glutamine	K6 conduit	0.19	63.06	-	-	1.31
	Fibre-optic	0.31	104.25	-	-	5.58
Ammonia	K6 conduit	0.06	13.17	0.09	16.14	0.14
	Fibre-optic	0.06	11.53	0.10	18.23	0.12
Phosphate	K6 conduit	0.14	31.58	0.07	10.49	0.33
	Fibre-optic	0.18	41.85	0.07	10.57	0.45
Glutamate	K6 conduit	0.21	64.45	-	-	0.67
	Fibre-optic	0.32	96.93	-	-	1.93
Lactate	K6 conduit	0.07	16.77	0.32	78.56	0.16
	Fibre-optic	0.13	30.81	0.18	44.69	0.28
HEPES	K6 conduit	0.30	14.97	0.08	8.26	0.71
	Fibre-optic	0.26	12.91	0.49	48.47	0.66
Bicarbonate	K6 conduit	0.37	45.72	0.60	24.49	1.64
	Fibre-optic	0.52	64.76	0.72	30.13	2.68

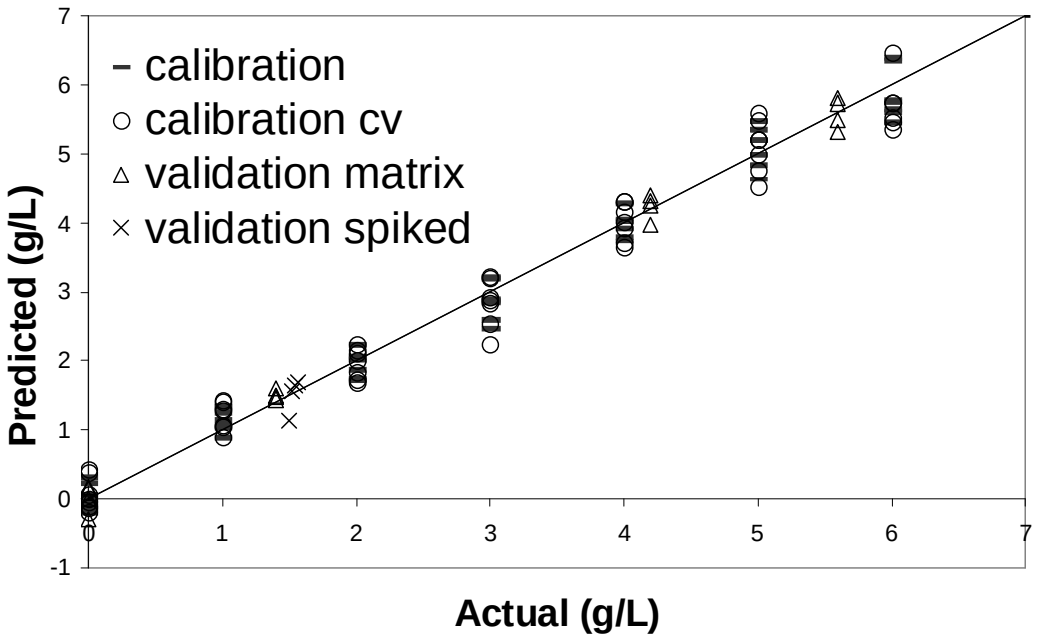


Fig. 3.5 Plot of predicted versus actual glucose concentrations for calibration and validation data, as measured with K6 conduit probe

Table 3.6 provides a summary of literature examples of infrared spectroscopy instruments with different types of sampling being used to measure glucose concentration; with glucose being the most common metabolite measured in bioprocesses. For the sake of comparison with the results in this study, any values reported as mM were converted to g/L and reduced to 2 significant figures.

TABLE 3.6 Infrared spectroscopy instruments being used to measure glucose in bioprocesses

Technique	Sampling Accessory		Culture type	#Cal	#Val	RMSEC g/L	RMSECV g/L	RMSEP g/L	% error	Ref.
NIR	Off-line	Transmission	Mammalian	58	14	0.60		0.53		[37]
NIR	Off-line	Transmission	Mammalian			0.04		0.07	1.86	[38]
NIR	In-line	Transmission/Fibre optic		104	24	0.10		0.15	3.98	[14]
NIR	In-line	Reflectance/Fibre optic	Bacteria	80	30	1.79	4.70	2.90		[13]
NIR	In-line	Transmission/Fibre optic	Mammalian	217		0.13	0.07	0.10		[12]
NIR	On-line	Transmission	Mammalian					0.17		[39]
NIR-Vis	Off-line	Transmission	Yeast	126	70	0.79		0.80		[40]
NIR	In-line	Transflectance/Fibre optic	Mammalian	50	16	0.10		0.20		[15]
NIR	In-line	Transflectance/Fibre optic	Bacteria				1.90	2.00	9.70	[16]
NIR	In-line	Transflectance/Fibre optic	Mammalian				0.23	0.19		[17]

Technique	Sampling	Accessory	Culture type	#Cal	#Val	RMSEC g/L	RMSECV g/L	RMSEP g/L	% error	Ref.
NIR	In-line	Transflectance/Fibre optic	Mammalian	73	12	0.30	0.36	0.36		[30]
MIR	In-line	ATR/Fibre optic	Bacteria			2.80		3.50		[36]
MIR	In-line	ATR/Conduit	Bacteria	91			0.26			[32]
MIR	In-line	ATR	Mammalian	60	225	0.09		0.11		[18]
MIR	Off-line	Transmission	Yeast/Bacteria				0.38			[41]
MIR	On-line	ATR	Yeast			0.35	0.40	0.27		[42]
MIR	Off-line	HATR	Bacteria	70	20	0.69		0.56		[43]

When results from this study are compared to those listed in Table 3.6, it can be seen that the RMSEC values are of the same order of magnitude but without a percentage error value, it is not always appropriate to compare these. The most comparable study was that of Rhiel *et al*,[18] where an RMSEC value of 0.09 g/L was obtained. While this value is lower than that achieved in this study, models were developed with a larger number of calibration samples which could improve the error values. As stated previously, an experimental design requiring a minimal number of samples (49) was used in this study.[26] When the model was applied to the validation matrix samples, phosphate, lactate, and bicarbonate all had lower errors when the K6 conduit probe was used. The errors for HEPES were lower for the fibre-optic probe but only by a small amount. Samples of media spiked with known amounts of a given component represent a slightly more difficult test for the model as these samples had unknown components present as the exact composition of the commercial medium used was not known. Spiked sample PEPs for glucose, ammonia and lactate were all higher than for the validation matrix samples. The results for the other components did not give consistent results for the different validation sets. Of the 9 studies with in-situ probes in Table 3.6, only 3 of these achieve a lower RMSEP than that reported in this study but as stated above, a direct comparison is not always possible without details on the percentage error.

In a similar way to the calibration models of single components in the preliminary analysis section, the models made from multicomponent samples were subjected to the same hypothesis test to determine an LOD with a false positive and false negative probability of 0.05. Every component had a higher LOD value in a multicomponent

mixture than that calculated using single components samples. In almost all cases the value was an order of magnitude higher and for glutamine, glutamate and bicarbonate, the LOD value was larger than the maximum concentration of each component that was used to make the calibration samples.

3.3.3 *External Influence investigation*

The results of the external influence tests can be seen in Table 3.7. In all cases of the agitation test, H_0 was accepted, strongly suggesting that variation in impeller speed had little effect on the spectra and hence the predicted concentration of the model. Hypothesis testing of the antifoam results indicated that the presence of antifoam did not appear to impact the spectra significantly; therefore it is not necessary to account for this when developing a calibration model.

Based on the results of the statistical tests applied to variation in pH, as expected, H_0 was rejected both for the spectral test and the relevance test, for both probes. Varying pH changes the chemical constitution of the media and based on the underlying principle of infrared spectroscopy,[33] it follows that the spectra will also change. The cell culture environment requires tight control of the pH; therefore major shifts in pH, resulting in inaccurately predicted values are unlikely to occur. However, for the development of a calibration model where the system is likely to experience pH changes, pH should be used as an additional factor in the experimental design.

TABLE 3.7 Hypothesis test results for external influences

		Agitation	Biomass	pH	Temperature	Antifoam	Pluronic
Multivariate	**Fibre optic**	accept H_o	reject H_o	reject H_o	reject H_o	accept H_o	reject H_o
test	**K6 conduit**	accept H_o	reject H_o	reject H_o	reject H_o	accept H_o	reject H_o
Relevance	**Fibre optic**	accept H_o	accept H_o	reject H_o	reject H_o	accept H_o	reject H_o
test	**K6 conduit**	accept H_o	reject H_o	reject H_o	reject H_o	accept H_o	accept H_o

Statistical analysis of biomass concentration indicates that the spectra and hence the predicted concentration values will change as the biomass concentration increases. By simulating the growth of a cell culture the sample media could possibly have been altered with the addition of small amounts of spent media in which the cells were suspended prior to addition. This potential change which would not occur over the course of an actual cell culture could have been confounded with the effect of the increasing biomass concentration. The relevance test indicated that this is not a significant result for the fibre optic probe while for the K6 conduit probe it was calculated to be significant at α=0.05. Previous studies have shown that biomass concentration can have an effect on NIR spectra due to light scatter and the fact that biomass absorbs in the NIR region.[14, 16] The ATR sample method for MIR instruments results in a short penetration depth of the MIR light source[44, 45] so scatter effects will not be present.[36]

The effect of varying temperature was calculated to have a significant effect for both the spectral and relevance tests, for both probes. This is not unexpected as temperature impacts the bonds between the molecules which is the underlying principle of vibrational spectroscopy including MIR.[46] As with pH, temperature is tightly controlled in mammalian cell cultures so unless deliberate temperature shifts are necessary, this factor does not need to be accounted for in the model. If the same model was to be applied to a cell culture with identical parameters with the exception of temperature, it would need to be recalibrated accounting for the temperature change by including temperature as a factor in the experimental design.

Pluronic can sometimes be added to mammalian cell cultures and the effect of addition at typical concentrations appears to have a significant effect on a spectral level. This result was

found using both probes. The relevance test indicated that it was not significant for the K6 conduit probe but that it was a significant effect for the fibre-optic probe. Given this difference, it would be recommended that if pluronic is to be added to a cell culture, it should be included in the experimental stage. This would not be difficult to do as it would be present at the same concentration in each sample.

3.4 Conclusion

This study outlines a methodology for evaluating the potential of mid-infrared spectroscopy as an on-line tool for monitoring mammalian cell culture media constituents. This method is beneficial as it identifies at an early stage where the technique may be best applied. A detailed comparison of two sampling systems is also outlined. This highlights any issues due to differing design of 'sample to crystal' interface areas. The importance of a chemometric technique, in the treatment of the data for the LOD analysis, is clearly evident as the observed LOD was always found to be greater than the calculated LOD. The LOD results indicate that at concentration levels found in cell cultures, certain components e.g. glutamine, lie below the detection ability of the instrument. Also, the LOD for each component is significantly higher in the multicomponent mixture than in the single component mixture.

An experimental design using a sample set of 49 and concentrations typically found in a mammalian cell culture were used in the development of the calibration models.

The study indicates the applicability of the technique in the monitoring of glucose and ammonia, both of which are major media constituents in mammalian cell cultures. Although all 8 components investigated cannot be reliably monitored, due to their presence at such low concentration levels, there is potential for the development of a control platform, of a mammalian cell

culture, based on glucose and ammonia. It should be noted that the accuracy of the prediction ability of a model is very much dependent on the sample set size and the concentration level of each of the components present. An increase in the number of samples in the sample set should improve the accuracy of the model.

Finally the external influence series of experiments indicate that changes in certain environmental conditions will impact spectra. However is must be noted that maintaining these environmental conditions relatively constant is crucial to the overall bioprocess. Therefore changes will impact results, however, the range over which these changes can occur, without impacting the bioprocess is so tight that the process will be affected before the spectra are impacted.

3.5 References

1. Mandenius, C. *et al*. Quality-by-Design for biotechnology-related pharmaceuticals. *Biotechnology Journal* **4**, 600-609 (2009).

2. Rathore, A. S. Roadmap for implementation of quality by design (QbD) for biotechnology products. *Trends Biotechnol.* **27**, 546-553 (2009).

3. Drennen, J. Quality by Design—What Does it Really Mean? *Journal of Pharmaceutical Innovation* **2**, 65-66 (2007).

4. Vojinovic, V., Cabral, J. M. S. & Fonseca, L. P. Real-time bioprocess monitoring. Part I: In situ sensors. *Sensors Actuators B: Chem.* **114**, 1083-1091 (2006).

5. Landgrebe, D. *et al*. On-line infrared spectroscopy for bioprocess monitoring. *Appl. Microbiol. Biotechnol.* **88**, 11-22 (2010).

6. Dabros, M., Amrhein, M., Gujral, P. & von Stockar, U. On-line Recalibration of Spectral Measurements using Metabolite Injections and Dynamic Orthogonal Projection. *Appl. Spectrosc.* **61**, 507-513 (2007).

7. Kornmann, H., Valentinotti, S., Marison, I. & von Stockar, U. Real-time update of calibration model for better monitoring of batch processes using spectroscopy. *Biotechnol. Bioeng.* **87**, 593-601 (2004).

8. Pollard, D. J. *et al*. Real-time analyte monitoring of a fungal fermentation, at pilot scale, using in situ mid-infrared spectroscopy. *Bioprocess. Biosyst. Eng.* **24**, 13-24 (2001).

9. U.S. Department of Health and Human Services: Food and Drug Administration. in *Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance* , 2004).

10. International Conference of Harmonisation. in *ICH Harmonised Tripatite Guideline: Q8(R1) Pharmaceutical Developement* , 2008).

11. 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* **37**, 371-378 (2001).

12. 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. *Biotechnol. Bioeng.* **84**, 13-19 (2003).

13. Lewis, C. B., McNichols, R. J., Gowda, A. & Coté, G. L. Investigation of Near-Infrared Spectroscopy for Periodic Determination of Glucose in Cell Culture Media in Situ. *Appl. Spectrosc.* **54**, 1453-1457 (2000).

14. Tamburini, E., Vaccari, G., Tosi, S. & Trilli, A. Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe. *Appl. Spectrosc.* **57**, 132-138 (2003).

15. 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. *Anal. Chim. Acta* **590**, 110-117 (2007).

16. Petersen, N. *et al*. In situ near infrared spectroscopy for analyte-specific monitoring of glucose and ammonium in streptomyces coelicolor fermentations. *Biotechnol. Prog.* **26**, 263-271 (2010).

17. Henriques, J., Buziol, S., Stocker, E., Voogd, A. & Menezes, J. in *Optical Sensor Systems in Biotechnology* (ed Rao, G.) 73-97 (Springer Berlin / Heidelberg, 2010).

18. 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. *Biotechnol. Bioeng.* **77**, 174-185 (2002).

19. Rhiel, M. H., Amrhein, M., 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. *Anal. Chem.* **74**, 5227-5236 (2002).

20. Ewing. in *Ewing's analytical instrumentation handbook* (Marcel Dekker, New York, 2005).

21. Olivieri, A. C. *et al*. Uncertainty estimation and figures of merit for multivariate calibration. *Pure Appl. Chem.* **78**, 633–661 (2006).

22. Brereton, R. G. in *Applied Chemometrics for Scientists* (John Wiley & Sons Ltd, Chichester, UK, 2007).

23. Ortiz, M. C. *et al*. Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares. *Chemometrics Intellig. Lab. Syst.* **69**, 21-33 (2003).

24. Sarabia, L. & Ortiz, M. C. DETARCHI: a program for detection limits with specified assurance probabilities and characteristic curves of detection. *Trac-Trends in Analytical Chemistry* **13**, 1-6 (1994).

25. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627-1639 (1964).

26. Munoz, J. A. & Brereton, R. G. Partial factorial designs for multivariate calibration: extension to seven levels and comparison of strategy. *Chemometrics Intell. Lab. Syst.* **43**, 89-105 (1998).

27. Naes, T., Isaksson, T., Fearn, T. & Davies, T. in *A User-Friendly Guide to Multivariate Calibration and Classification* 344 (NIR Publications, Chichester, UK, 2002).

28. Hennessy, S., Downey, G. & O'Donnell, C. P. Attempted Confirmation of the Provenance of Corsican PDO Honey Using FT-IR Spectroscopy and Multivariate Data Analysis. *J. Agric. Food Chem.* **58**, 9401-9406 (2010).

29. Jobson, J. D. in *Applied Multivariate Data Analysis* (Springer, New York, 1999).

30. Petiot, E. *et al*. In situ quantification of microcarrier animal cell cultures using near-infrared spectroscopy. *Process Biochemistry* **45**, 1427-1431 (2010).

31. Dhamelincourt, P. & Ramirez, F. J. Polarized Micro-Raman and FT-IR Spectra of L-Glutamine. *Appl. Spectrosc.* **47**, 446-451 (1993).

32. Doak, D. L. & Phillips, J. A. In Situ Monitoring of an *Escherichia coli* Fermentation using a Diamond Composition ATR Probe and Mid-infrared Spectroscopy. *Biotechnol. Prog.* **15**, 529-539 (1999).

33. Schenk, J., Marison, I. W. & von Stockar, U. pH prediction and control in bioprocesses using mid-infrared spectroscopy. *Biotechnol. Bioeng.* **100**, 82-93 (2008).

34. 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. *J. Biotechnol.* **128**, 344-353 (2007).

35. Schenk, J., Marison, I. W. & von Stockar, U. Simplified Fourier-transform mid-infrared spectroscopy calibration based on a spectra library for the on-line monitoring of bioprocesses. *Anal. Chim. Acta* **591**, 132-140 (2007).

36. Fayolle, P., Picque, D. & Corrieu, G. On-line monitoring of fermentation processes by a new remote dispersive middle-infrared spectrometer. *Food Control* **11**, 291-296 (2000).

37. Chung, H., Arnold, M. A., Rhiel, M. & Murhammer, D. W. Simultaneous Measurements of Glucose, Glutamine, Ammonia, Lactate, and Glutamate in Aqueous Solutions by Near-Infrared Spectroscopy. *Appl. Spectrosc.* **50**, 270-276 (1996).

38. McShane, M. J. & Cote, G. L. Near-Infrared Spectroscopy for Determination of Glucose, Lactate, and Ammonia in Cell Culture Media. *Appl. Spectrosc.* **52**, 1073-1078 (1998).

39. Rhiel, M. H., Cohen, M. B., Arnold, M. A. & Murhammer, D. W. On-line monitoring of human prostate cancer cells in a perfusion rotating wall vessel by near-infrared spectroscopy. *Biotechnol. Bioeng.* **86**, 852-861 (2004).

40. Finn, B., Harvey, L. M. & McNeil, B. Near-infrared spectroscopic monitoring of biomass, glucose, ethanol and protein content in a high cell density baker's yeast fed-batch bioprocess. *Yeast* **23**, 507-517 (2006).

41. Franco, V. G., Perín, J. C., Mantovani, V. E. & Goicoechea, H. C. Monitoring substrate and products in a bioprocess with FTIR spectroscopy coupled to artificial neural networks enhanced with a genetic-algorithm-based method for wavelength selection. *Talanta* **68**, 1005-1012 (2006).

42. Mazarevica, G., Diewok, J., Baena, J. R., Rosenberg, E. & Lendl, B. On-line fermentation monitoring by mid-infrared spectroscopy. *Appl. Spectrosc.* **58**, 804-810 (2004).

43. Roychoudhury, P., Harvey, L. M. & McNeil, B. At-line monitoring of ammonium, glucose, methyl oleate and biomass in a complex antibiotic fermentation process using attenuated total reflectance-mid-infrared (ATR-MIR) spectroscopy. *Anal. Chim. Acta* **561**, 218-224 (2006).

44. Wilson, R. H. Fourier-Transform Midinfrared Spectroscopy for Food Analysis. *Trac-Trends Anal. Chem.* **9**, 127-131 (1990).

45. Wilson, R. H. & Tapp, H. S. Mid-infrared spectroscopy for food analysis: recent new applications and relevant developments in sample presentation methods. *Trac-Trends Anal. Chem.* **18**, 85-93 (1999).

46. Wulfert, F., Kok, W. T. & Smilde, A. K. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Anal. Chem.* **70**, 1761-1767 (1998).

Chapter 4

Study 2: Application and optimisation of in-situ MIR calibration models for the prediction of glucose and lactate in mammalian cell cultures

As submitted to Applied Spectroscopy, January 2013

Abstract

The primary aim of this study was to apply multivariate calibration models to data collected in real time during the course of a CHO DP12 cell culture in bioreactors to predict glucose and lactate concentrations. To achieve this aim, an investigation into the optimisation of these models was performed in an effort to improve their accuracy and robustness. The study comprised a series of 3 cultures which were monitored in-situ using mid-infrared (MIR) spectroscopy. Samples for reference HPLC analysis were taken daily to compare to the in-situ MIR predicted data. Aliquots of these daily samples were scanned using the same instrument once all cultures were complete and the same partial least squares regression (PLS) models applied in order to evaluate the set- up when applied in an offline or at-line scenario. The PLS models used exhibited expected trends when applied in-situ, with glucose depletion and lactate formation clearly evident. The accuracy of prediction however was low, with RMSEP values for glucose and lactate, 0.73g/L and 1.21g/L respectively. When the same models were applied to the spectra of culture samples taken offline the glucose and lactate errors were reduced by 60.27 and 13.22% respectively. Optimisation of glucose and lactate models for in-situ use was achieved by developing PLS models using spectral data generated in-situ and offline from all 3 cell cultures. When offline data was used as

part of the optimised calibration training set, such models consistently performed better than the original PLS models. Optimised glucose model results showed an improvement in RMSEP as high as 37.93% while optimised lactate model results had an improved RMSEP of 61.98%.

4.1 Introduction

Over a decade since its conception in August 2002, the Food and Drug Administration's (FDA) initiative, "Pharmaceutical Current Good Manufacturing Practices", (cGMP's), has had a significant impact on the pharmaceutical and biopharmaceutical industries.[1-3] The purpose of this initiative is to modernise the regulation of pharmaceutical quality through the support and promotion of risk-based and science-based approaches.[4] Central to the implementation of this initiative is the concept of Quality by Design (QbD) which utilises Process Analytical Technologies (PAT) to gain in-depth process understanding, leading to the eventual application of these technologies to monitor and control processes, thus mitigating risk and reducing variability.[5]

Quantitative analysis of the key analytes and metabolites in a bioprocess can provide a plethora of information, which, when applied, can yield powerful results such as improved product quality and enhanced biomass production.[6] PAT tools should be capable of providing rapid yet reliable measurements, be possible to calibrate and preferably be non-invasive. Infrared (IR) spectroscopy meets these criteria and has been shown to monitor key analytes both in-situ and at-line.[6-8] Development of chemometric models for quantitative measurement of cell culture components via IR has been documented, with the greater portion of this work focused on the use of near-

infrared (NIR) spectroscopy.[9-14] The development of such models using mid-infrared (MIR) spectroscopy has also been reported, but evaluations of its online application to mammalian cell cultures are minimal.[15,16]

A number of studies have been carried out using multivariate analysis of spectroscopic data to develop calibration models capable of simultaneous monitoring of several key analytes in a bioreactor.[17-20] Emphasis has been placed on the importance of the calibration set and Cervera et al. outline various techniques employed for NIR spectroscopy.[21] There has been much debate on the selection of the optimum calibration sample set, with some researchers choosing purely synthetic sample matrices arguing that the use of such samples provides accurate results over wide concentration ranges and proves a more difficult validation test for the model.[22] Others favour the use of real fermentation samples highlighting fermentation are broths impossible to simulate.[9,11] The final option is that of a combined sample set, one containing spectra obtained from real samples and/or spiked samples and/or synthetic samples. This method has also been used by researchers when developing calibration matrices.[23,24] The study presented here performs a comparative investigation by examining the efficacy of various MIR models for glucose and lactate prediction both in-situ and at-line. Models were created using spectra collected from synthetic samples only, a combination of synthetic and real cultures samples (collected offline) and finally from all sample types; synthetic samples and real culture samples (collected online and offline).

The purpose of this study was to evaluate the capability of MIR as both an online in-situ and at-line tool for glucose and lactate monitoring when applied to a mammalian cell culture. This was achieved by application to a series of CHO DP12 cultures in a

bioreactor, but also by further analysis of spectroscopic data collated. A direct comparison between models, with various calibration sets and also spectral pre-treatments is provided in order to identify an optimum methodology for MIR calibration model development for the purpose of monitoring media constituents within a bioreactor.

4.2 Materials & Methods

4.2.1 Cell culture

A series of 3 batch cultures was completed in a bio-reaction calorimeter, RC1e (Mettler-Toledo AutoChem Inc., Columbia MD, USA). CHO DP12 cells were cultivated in suspension in 1.6L of media (ExCell^{TM} 325 PF CHO, Sigma Aldrich Ireland Ltd.) which was supplemented with Antifoam C, glutamine, insulin & MTX (Sigma Aldrich Ireland Ltd.). Samples were removed every 24 hours and cell counts performed manually. The pH of the culture was controlled and maintained between pH 7.0 and pH 7.2 using 2 M NaOH and CO_2. Dissolved oxygen was also monitored and entered the reactor via a ring sparger. The reactor temperature was controlled at 37 ^{o}C for the duration of each of the 3 cultures.

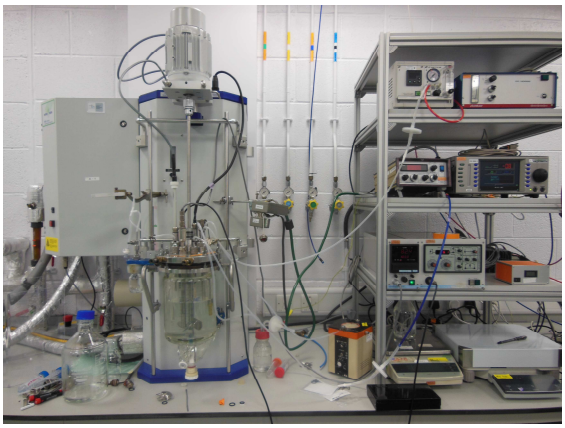


Fig. 4.1 Bio-reaction calorimeter, RC1e, with controller set up

4.2.2 Instrumentation and real time monitoring

The cell culture environment within the reactor was monitored in real time using a Fourier transform mid-infrared ReactIR iC10 instrument with MCT detector (Mettler Toledo AutoChem, Inc., Columbia, US). A K6 conduit 16 mm immersion probe was used with a fixed path length of approx. 1-2 μm and a diamond ATR crystal with 6 internal reflections. The detector and probe were purged with nitrogen gas continuously. Prior to media transfer to the reactor, the vessel was filled with deionised water and heated to 37 ^{o}C. A background scan was taken at this point. Under sterile conditions the water was removed, the media added and the reactor inoculated. The instrument was set to scan every 10 minutes, with each scan comprising 128 co-added scans. All data was collected on the instrument software, iC IR^{TM} (Mettler-Toledo AutoChem Inc., Columbia MD, USA). As the spectral files were generated, they were exported to a MATLAB program (v7.9.0.529 (R2009b), The MathWorks Inc., Cambridge, UK), where previously developed PLS models[25] developed on the PLS toolbox for MATLAB (V6.2 Eigenvector Research Inc, Wenatchee, WA, US) were used to predict the concentration of glucose and lactate present inside the bioreactor. A check was included in the program in an attempt to eliminate predicted concentrations that were impossible, i.e. sudden drops or increases in concentration which could not reasonably be obtained within the 10 minutes between each scan.

4.2.3 Reference analysis

The first of the 3 cultures ran until day 7, after which time cell counts indicated that the culture had moved from stationary to death phase. The following 2 cultures ran to day 6. All cultures were sampled daily. Once cell counts were completed the samples

were prepared for offline analysis. Raw samples were centrifuged and the supernatant retained. Glucose and lactate concentrations were determined by HPLC analysis. The supernatant was filtered using a 0.22 μm hydrophilic PTFE filter (Millipore Corporation, Billerica, MA, USA). A sample volume of 12 μl was injected onto a SUPELCOGEL C-610H column (Sigma-Aldrich Corporation, St. Louis, MO, USA) equilibrated with 0.01 M sulphuric acid solution at a flowrate of 0.5 ml/min

4.2.4 Offline/at-line spectra collection and analysis

The supernatant samples (total of 22) were stored at -80 ^{o}C until all 3 cultures were complete, after which time they were thawed, heated to 37 ^{o}C and scanned offline and in triplicate against a background of deionised water at 37 ^{o}C. The calibration models used to predict the concentrations of the 2 components of interest in real time (see Table 4.1) were then applied to the spectral data generated from the offline scans in order to determine if the accuracy was improved when the instrument was not subject to the environmental conditions of a cell culture. This study was performed in order to investigate the at-line capabilities of both the instrument and the models.

4.2.5 Model development and optimisation

The development of the multivariate models applied to the spectral data collected during the 3 cell cultures has been fully described in a previous study.[25] All models created were based on a 7-level partial factorial design.[26] Synthetic samples (49) were generated to reflect the conditions of a cell culture environment, however actual culture samples, were not included in the model development. Table 4.1 summaries the PLS regression models used for in-situ application. These models, developed in a previous study, both use mean centering and 2nd derivative Savitzky-Golay as pre-

treatments. Mean centering is often performed as a default pre-treatment. It was selected here as it prevents data points that are further from the origin exerting an undue amount of leverage over the points that are closer to the origin. In this case, as the cultures were run over the course of a week, it minimised the effect of instrument drift on the model results. Second derivative Savitzky-Golay or S-G smoothing, as it is also known; enhances small spectral differences and separates overlapping peaks. As the culture medium contains several components, some of which are unknown, this form of pre-treatment was selected to separate merging spectral peaks occurring as a result of the many components present in the culture.

TABLE 4.1 PLS regression models for in-situ & at-line application

Model #	Component	Pre-treatment	No. Latent Variables	Wavenumber Range (cm^{-1})
1	Glucose	Mean centered 2^{nd} derivative Savitzky-Golay, filter width 15	4	950-1450
1	Lactate	Mean centered 2^{nd} derivative Savitzky-Golay, filter width 15	4	1000-1620

Models used to predict glucose and lactate concentrations were further processed in an effort to optimise such models and create more robust and reliable predictions. The calibration sets were expanded to include spectra taken in-situ during the culture, and/or culture sample spectra collected offline. These models are referred to throughout this text as "hybrid models" as the x-data required in generating these regression models is a composite of spectral data obtained from synthetic and real culture samples.

Four model categories are presented, the first of which has been outlined in detail in Table 4.1. PLS regression, optimised by leave-one-out cross validation was used to develop all calibration models. All data was mean centered and second derivative pre-treatments, of filter widths 15 and 21 were applied, calculated using Savitzky-Golay filters.[27] Second derivative pre-treatments were chosen as they can eliminate the effect of linear baseline off-sets[28] that are likely to occur over the course of a culture and smooth noisy spectra. Standard normal variate (SNV) pre-treatment was also applied to mean centred data. This pre-treatment is predominantly used for NIR spectra as it removes multiplicative interferences of scatter and particle size[29] however it was also

examined in this study as a potential pre-treatment due to the nature of the cell culture environment (increasing turbidity due to biomass growth).

The second model category was divided into 2 sub-models: Model 2a and Model 2b, as they were largely similar, however the calibration set size of Model 2b was smaller and the validation sets used for both models were different. Model 2a consisted of the 49 original spectra obtained from scans of the synthetic samples and also spectra collected offline from all 3 cultures (22 culture samples) creating a training set of 71 samples. Using this data 3 models were developed for both glucose and lactate. These models differed in pre-treatments applied and/or the number of latent variables used. Table 4.2 provides specific details for all models developed. All versions of Model 2a were applied to the in-situ data generated during cultures 1, 2 and 3.

TABLE 4.2 Hybrid model details

Model Type	Component	Pre-treatments & Latent Variables	Calibration data set	Validation data set
2a	Glucose	MC_2der15_4LV[a]	49 synthetic samples, 22 offline spectra (cultures 1,2 &3)	In-situ spectra: Cultures 1,2 & 3
2a	Glucose	MC_2der21_5LV	49 synthetic samples, 22 offline spectra (cultures 1,2 &3)	In-situ spectra: Cultures 1,2 & 3
2a	Glucose	MC_SNV_5LV	49 synthetic samples, 22 offline spectra (cultures 1,2 &3)	In-situ spectra: Cultures 1,2 & 3
2a	Lactate	MC_2der15_4LV	49 synthetic samples, 22 offline spectra (cultures 1,2 &3)	In-situ spectra: Cultures 1,2 & 3
2a	Lactate	MC_2der21_3LV	49 synthetic samples, 22 offline spectra (cultures 1,2 &3)	In-situ spectra: Cultures 1,2 & 3
2a	Lactate	MC_SNV_4LV	49 synthetic samples, 22 offline spectra (cultures 1,2 &3)	In-situ spectra: Cultures 1,2 & 3
2b	Glucose	MC_2der15_4LV	49 synthetic samples, 11 offline spectra (cultures 1,2 &3)	11 offline spectra: Cultures 1,2 & 3
2b	Glucose	MC_2der21_6LV	49 synthetic samples, 11 offline spectra (cultures 1,2&3)	11 offline spectra: Cultures 1,2 & 3
2b	Glucose	MC_SNV_4LV	49 synthetic samples, 11 offline spectra (cultures 1,2&3)	11 offline spectra: Cultures 1,2 & 3
2b	Lactate	MC_2der15_3LV	49 synthetic samples, 11 offline spectra (cultures 1,2&3)	11 offline spectra: Cultures 1,2 & 3
2b	Lactate	MC_2der21_3LV	49 synthetic samples, 11 offline spectra (cultures 1,2&3)	11 offline spectra: Cultures 1,2 & 3
2b	Lactate	MC_SNV_4LV	49 synthetic samples, 11 offline spectra (cultures 1,2&3)	11 offline spectra: Cultures 1,2 & 3

Model Type	Component	Pre-treatments & Latent Variables	Calibration data set	Validation data set
3	Glucose	MC_2der15_4LV	49 synthetic samples, in-situ spectra (cultures 1&2)	In-situ spectra: Culture 3
3	Glucose	MC_2der21_4LV	49 synthetic samples, in-situ spectra (cultures 1&2)	In-situ spectra: Culture 3
3	Glucose	MC_SNV_6LV	49 synthetic samples, in-situ spectra (cultures 1&2)	In-situ spectra: Culture 3
3	Lactate	MC_2der15_3LV	49 synthetic samples, in-situ spectra (cultures 1&2)	In-situ spectra: Culture 3
3	Lactate	MC_2der21_3LV	49 synthetic samples, in-situ spectra (cultures 1&2)	In-situ spectra: Culture 3
3	Lactate	MC_SNV_4LV	49 synthetic samples, in-situ spectra (cultures 1&2)	In-situ spectra: Culture 3
4	Glucose	MC_2der15_4LV	49 synthetic samples, in-situ spectra (cultures 1&2), offline spectra (cultures 1,2&3)	In-situ spectra: Culture 3
4	Glucose	MC_2der21_5LV	49 synthetic samples, in-situ spectra (cultures 1&2), offline spectra (cultures 1,2&3)	In-situ spectra: Culture 3
4	Glucose	MC_SNV_5LV	49 synthetic samples, in-situ spectra (cultures 1&2), offline spectra (cultures 1,2&3)	In-situ spectra: Culture 3

Model Type	Component	Pre-treatments & Latent Variables	Calibration data set	Validation data set
4	Lactate	MC_2der15_3LV	49 synthetic samples, in-situ spectra (cultures 1&2), offline spectra (cultures 1,2&3)	In-situ spectra: Culture 3
4	Lactate	MC_2der21_3LV	49 synthetic samples, in-situ spectra (cultures 1&2), offline spectra (cultures 1,2&3)	In-situ spectra: Culture 3
4	Lactate	MC_SNV_4LV	49 synthetic samples, in-situ spectra (cultures 1&2), offline spectra (cultures 1,2&3)	In-situ spectra: Culture 3

a: MC = mean centered; 2der15 = Savitzky-Golay 2nd derivative pre-treatment with filter with of 15; 2der21 = Savitzky-Golay 2nd derivative pre-treatment with filter with of 21; LV = latent variables

The calibration set used in the development of Model 2b also contained the 49 original spectra collected from the synthetic samples, but in this case only 11 of the 22 spectra obtained from the offline culture samples were added. Model 2b was then applied to the remaining 11 offline spectra in order to establish if the addition of spectral data obtained from an actual cell culture sample enhanced the predictive ability of an at-line model. As with model 2a, 3 PLS regression models were developed for each of the 2 components of interest.

Model 3 comprised the original 49 spectra and spectra obtained each day, in-situ, at the time of sampling, during the first 2 cultures. This model was then applied to the online spectra generated over the course of the third culture. The training sets for glucose and lactate for Model 3 differed slightly. Based on the in-situ application analysis in this study and the limit of detection investigation completed in an earlier study,[25] the glucose scans were found to be unreliable once the glucose concentration fell below its limit of detection. Only spectra collected prior to this point were included in the development of the new glucose models. Therefore only spectra collected at the sample time from day 0 to day 3 were used. Lactate was formed and from day 0 lactate levels lay above the minimum detection limit, therefore all spectra collected at the sample time during the first two cultures were used in the newly developed lactate models.

In the final model, Model 4, the 22 spectra collected offline were added to the training sets of Model 3, for glucose and lactate. This model was then applied to the in-situ spectral data of the third culture.

4.3 Results and Discussion

4.3.1 In-situ application

Glucose and lactate concentrations were predicted in real time as all 3 cultures progressed. Reference analysis via HPLC showed that the actual glucose concentration in each of the cultures began at 3.0g/L ± 0.2g/L and fell to 0g/L. The glucose versus time plots in Fig. 4.1 show that the online predicted glucose concentrations followed this trend in depletion. This trend has been reported in other studies.[30-32] However to the authors' knowledge, only two other studies exist where measurements of a mammalian cell culture were taken in real time using an MIR immersion probe, and therefore directly comparable.[15,16] As the glucose approached its limit of detection (LOD) of 0.41g/L, established in a previous study,[25] the results became unreliable. In cultures 1 and 3, in-situ results predicted an increase in glucose, having reached a minimum of 0g/L, and culture 2 predicted negative concentrations. (Note: due to unplanned instrument downtime during culture 2 there are 15 hours on day 5 over which in-situ data was not collected). These spurious spectra, exhibited by all 3 cultures may be as a result of increased biomass concentration from day 4, causing probe fouling, however, the lactate results did not appear to be impacted by possible physical changes in the cell culture environment, and so it is most likely that in-situ, predicted glucose values after day 4 were unreliable as the glucose concentration in the bioreactor had fallen below a detectable limit.

The root mean squared error of prediction, (RMSEP), was used to evaluate each of the models.[33] The RMSEP was calculated based on data collected in-situ from all 3 cultures. This was found to be 0.73g/L; quite a large error given the maximum value of glucose at anytime was 3.0g/L ±0.2g/L. This value was recalculated using only data

generated from day 0 to day 3 of all cultures and was reduced to 0.6g/L, thereby improving the error by 17.80%. An obvious trend in glucose depletion was observed, however the level of accuracy required for application to real time monitoring and control of mammalian cell cultures was not reached using this set up and method of model development. Therefore further optimisation of the model was investigated.

HPLC results for lactate showed that the minimum starting concentration was 0g/L. The lactate concentration reached a maximum on day 4 where concentrations of all 3 cultures were 2.78g/L ± 0.06g/L with a slight drop in concentration over the remaining days of each culture. The lactate versus time plots in Fig. 4.1 show that the in-situ predicted lactate concentrations followed the observed trend, with in-situ concentrations peaking on day 4 and then dropping by ~0.35g/L. However all 3 cultures indicate that predicted lactate concentrations consistently fell below actual lactate values. The average maximum difference between actual and predicted concentrations, over the course of all 3 cultures was 1.65g/L, a considerable difference and over 50% of the maximum lactate concentration reached for all 3 cultures. However the clear and precise trending exhibited by all 3 cultures suggests that further optimisation of the model used may further reduce the deficit and create more robust and reliable models. An RMSEP value of 1.21g/L was calculated for lactate. Spectral data and predicted concentrations did not indicate unreliable, spurious results after day 4, therefore unlike glucose, a second RMSEP value based on days 0-3 was not calculated.

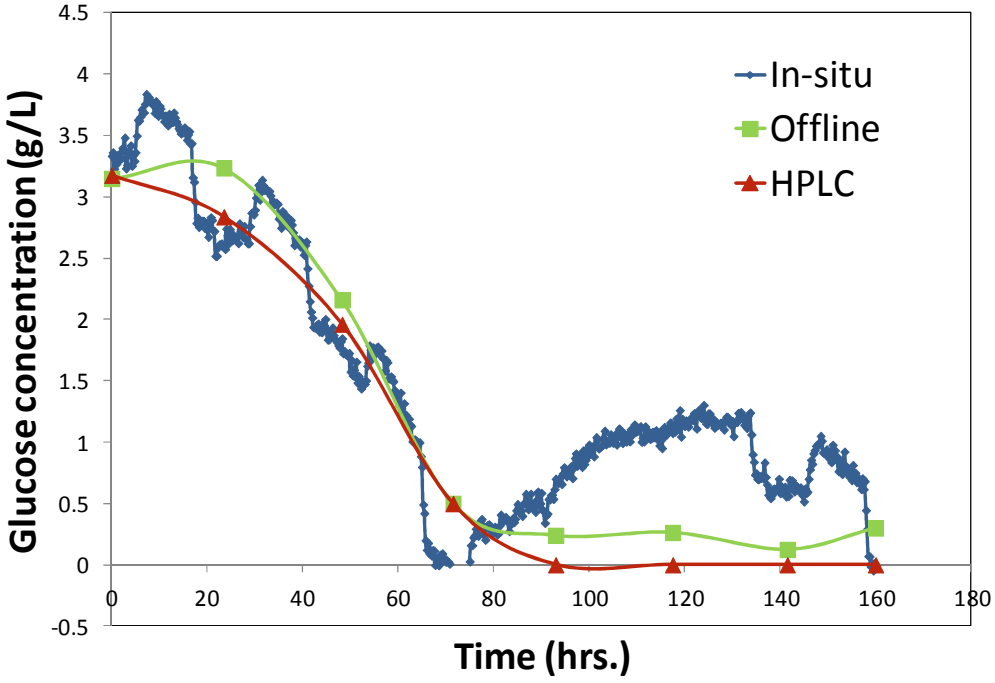


Fig. 4.2 (a)

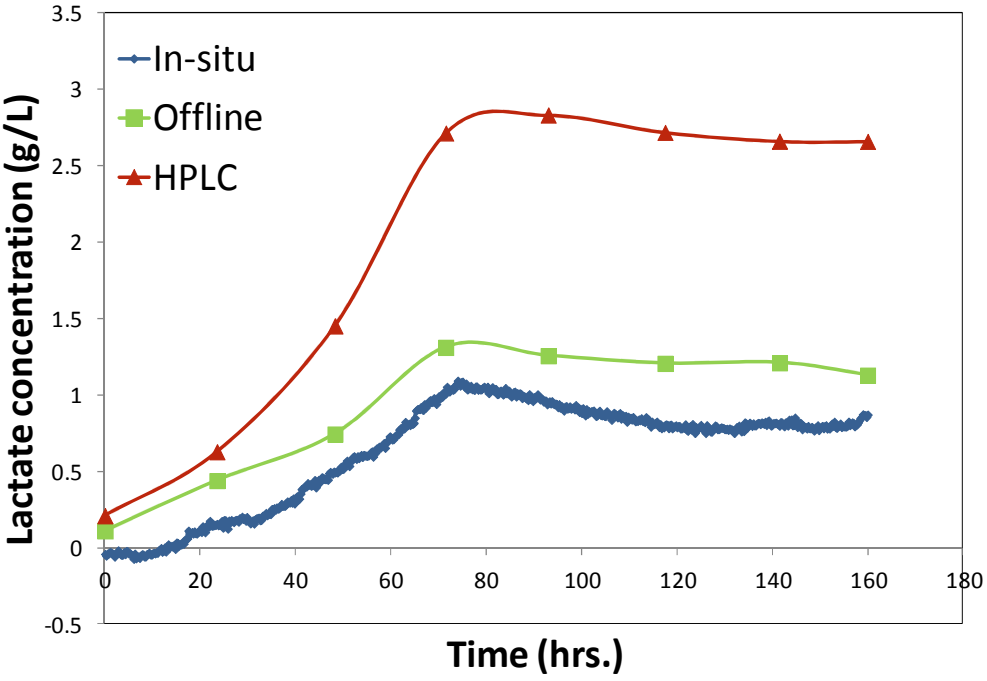


Fig. 4.2 (b)

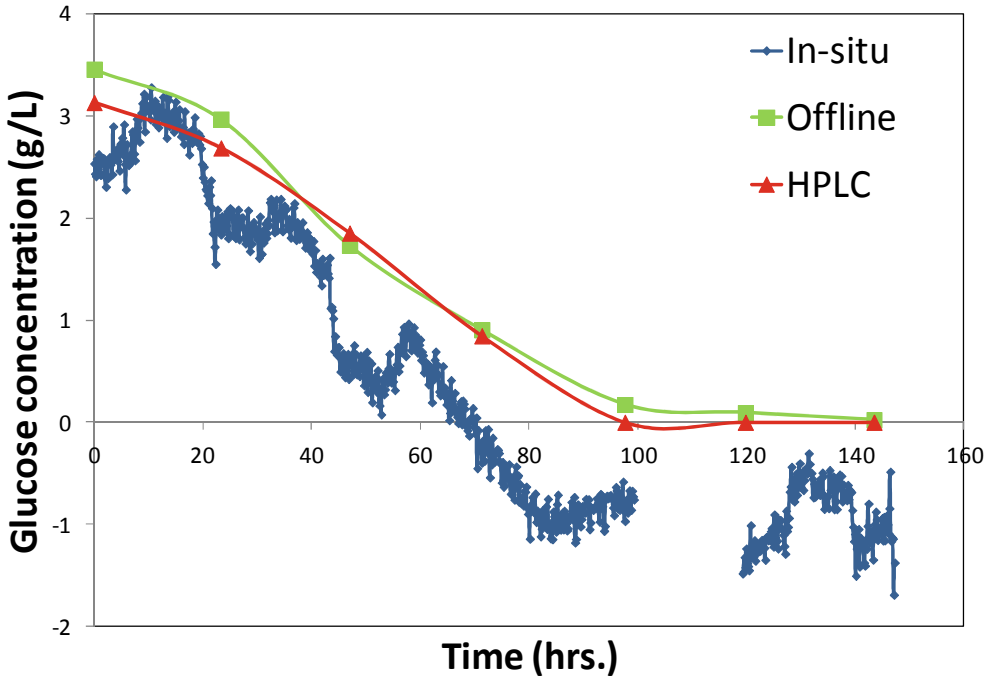


Fig. 4.2 (c)

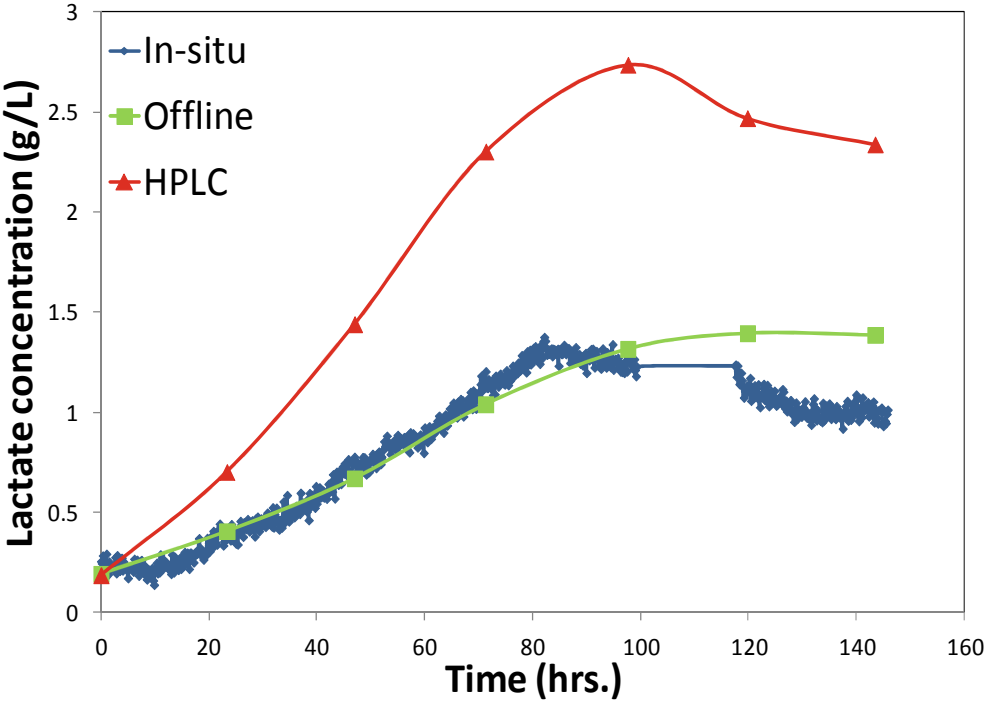


Fig. 4.2 (d)

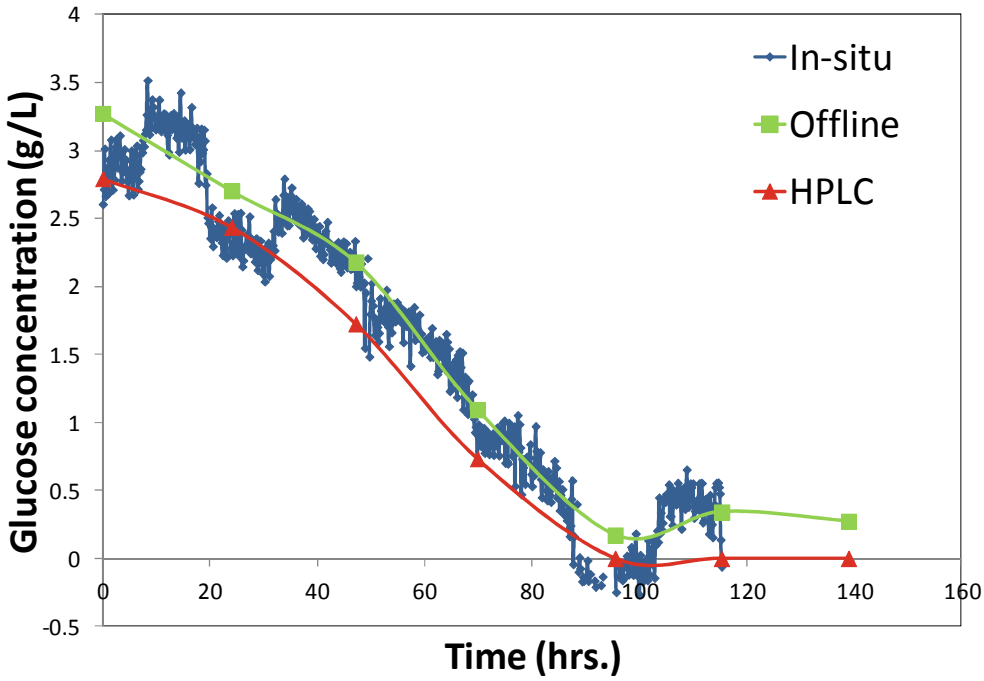


Fig. 4.2 (e)

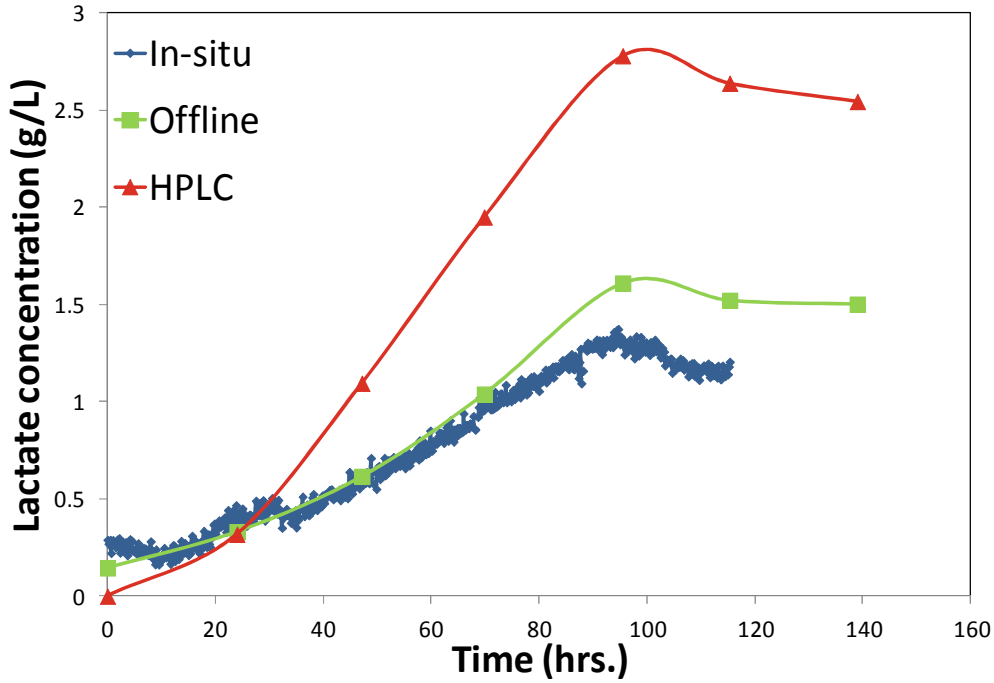


Fig. 4.2 (f)

Fig. 4.2 (a-f) Online, offline & HPLC results of glucose & lactate 'v' time for cultures 1, 2 & 3

4.3.2 At-line application

Daily samples from all 3 cultures were retained to perform offline/at-line analysis. Glucose and lactate concentrations for all 22 samples were predicted using the PLS regression models outlined in Table 4.1. Fig. 4.2 summarises the in-situ, offline and HPLC results for glucose and lactate trends for all 3 cultures. Based on these plots it is evident that the offline method is more effective. RMSEP values were calculated as a means of evaluating the comparison between both in-situ and at-line applications. At-line prediction yielded a glucose RMSEP value of 0.26g/L, a 60.27% decrease on the real time value of 0.73g/L, while the lactate RMSEP fell to 1.02g/L, a 13.22% reduction on the in-situ RMSEP of 1.21g/L. Table 4.3 summarises the comparison of both methods.

Table 4.3 In-situ versus at-line prediction comparison

Glucose		**Lactate**	
In-situ RMSEP (g/L)	At-line RMSEP (g/L)	In-situ RMSEP (g/L)	At-line RMSEP (g/L)
0.73	0.29	1.21	1.05

The improvement in the predictive abilities of the instrument may be attributed to a number of factors: the constant flux within the reactor and the possible minor

vibrations of the probe due to instrumentation mechanics were eliminated. All samples were scanned in a controlled environment without agitation and impact of external physical influences. A known cause of spurious scans was removed by application of this technique; that being the formation of gaseous bubbles on the probe tip. In the aqueous cell culture environment of a bioreactor, real time monitoring using an immersion probe is subject to frequent and unpreventable formation of bubbles on monitoring devices.[34] This is further compounded by the required aeration of the cell culture. Such bubbles, either partially or fully covering the probe tip, skew the results as the IR may penetrate into the gas within the bubble thereby producing an inaccurate "snapshot" of the contents of the reactor. The effect of this may be mitigated by programming the data acquisition system to reject scans that are not plausible, as was done in this case. At-line analysis allows for the removal of bubbles prior to initiation of scanning.

Though not appearing to be problematic in this instance, potential fouling of the probe in the latter stages of the cell culture, due to high cell densities is also eliminated. The at-line method removed cellular matter by centrifugation of samples and scanning of the resulting supernatant.

4.3.3 Model optimisation

Spectral data generated both in-situ and offline was used to further develop optimised PLS models. The details of these hybrid models are outlined in Table 4.2. Second derivative pre-treatments consistently resulted in improved predictions; however the optimum filter window varied depending on the model type. The SNV pre-treatment did not show the same improved prediction, see Fig. 4.3 and Fig. 4.4, and in the case of

glucose, the RMSEP values of hybrid models using SNV pre-treatment failed to improve upon the RMSEP values of the original model (Table 4.1). Hybrid model results are outlined in Table 4.4 and Table 4.5.

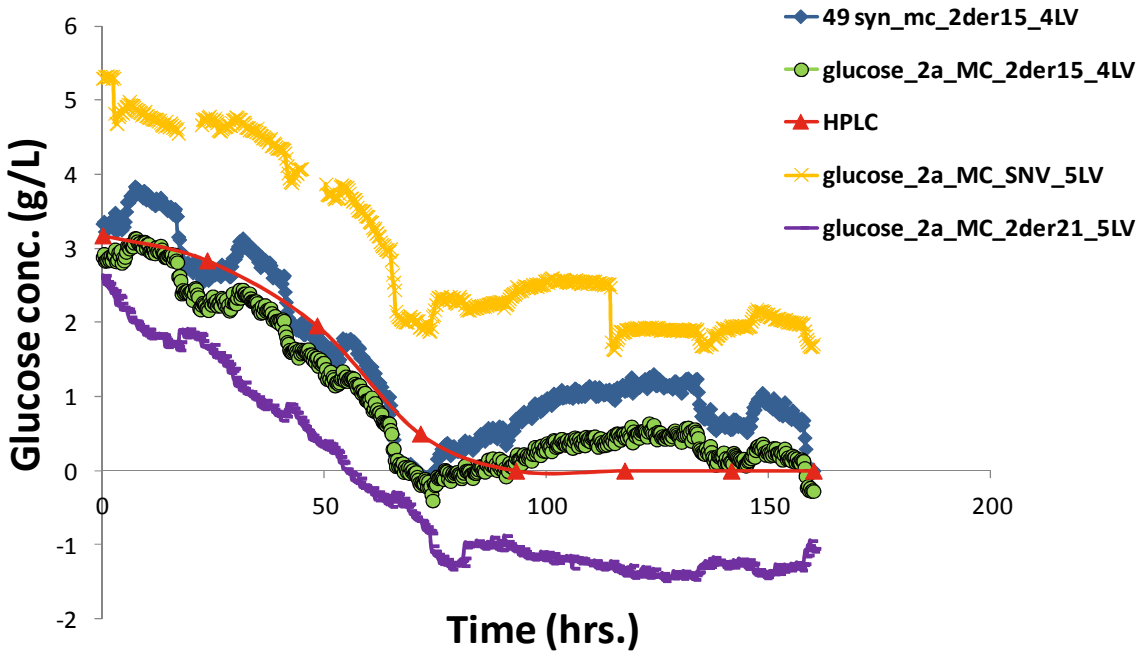


Fig. 4.3 (a) Model 2a; Culture I

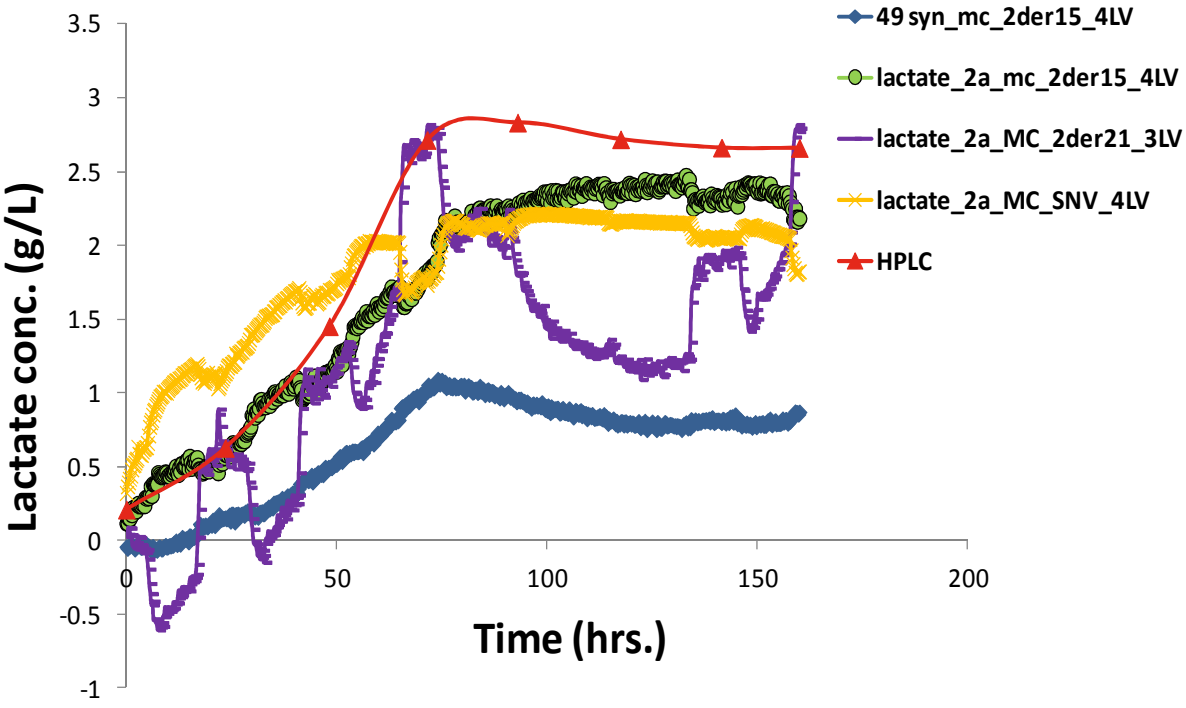


Fig. 4.3 (b) Model 2a; Culture I

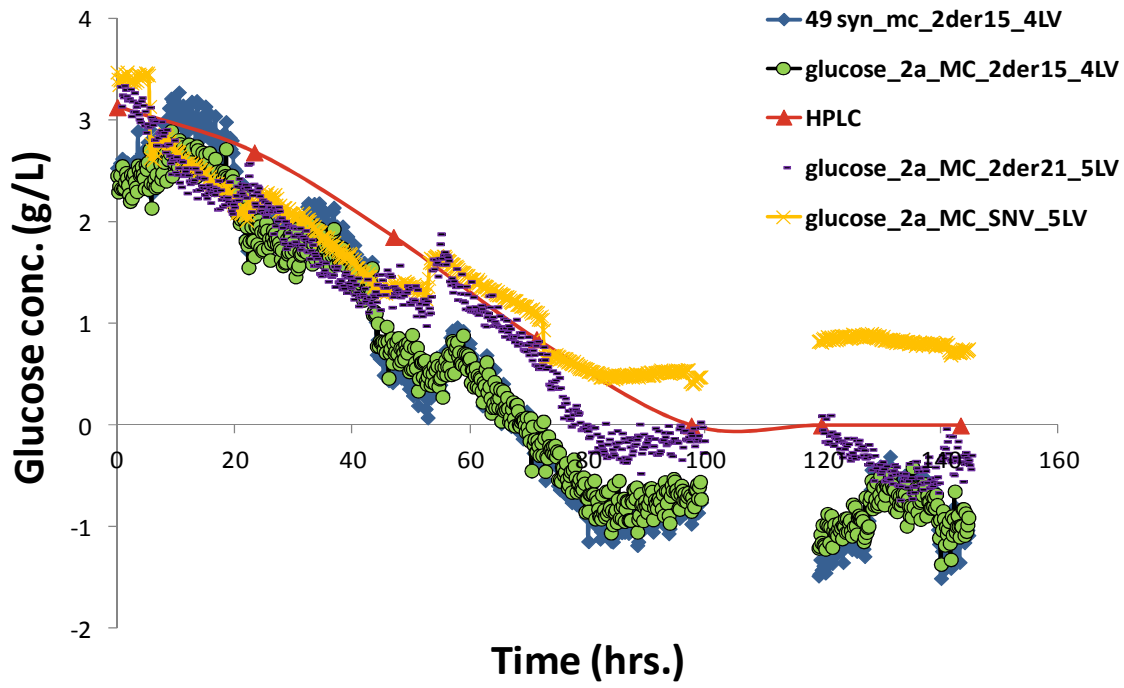


Fig. 4.3 (c) Model 2a; Culture II

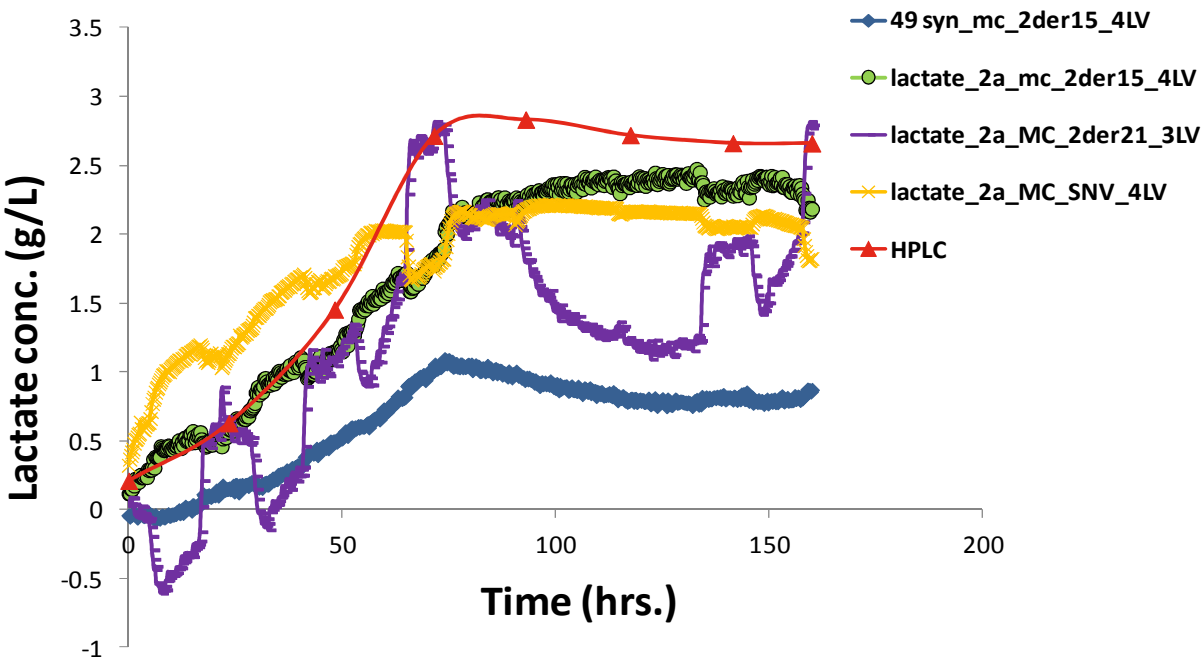


Fig. 4.3 (d) Model 2a; Culture II

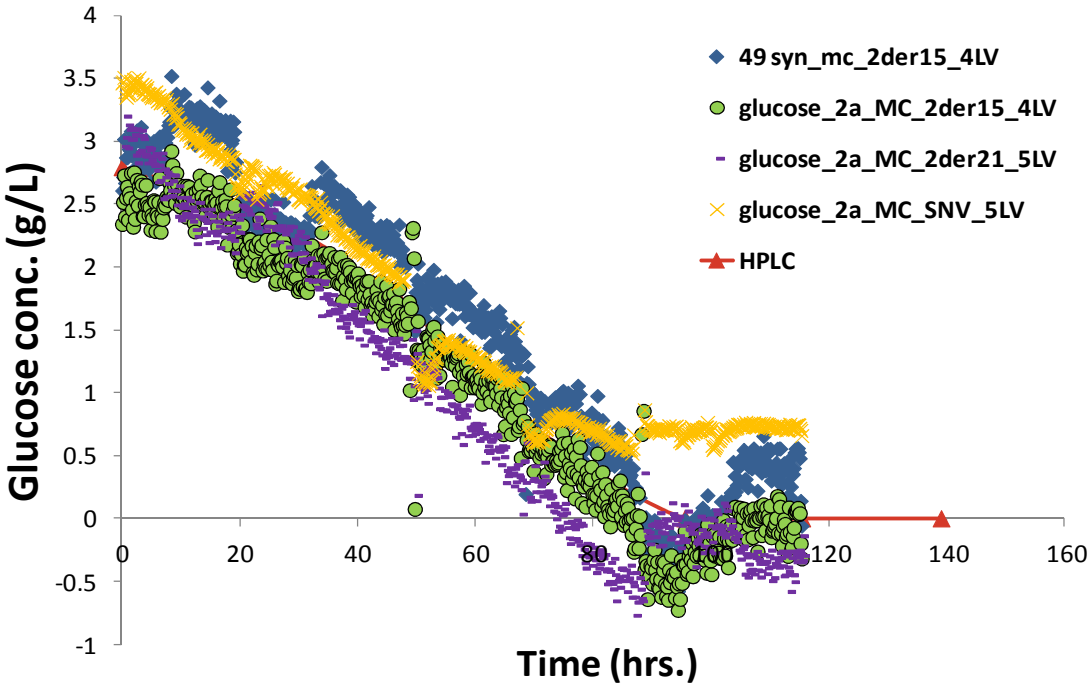


Fig. 4.3 (e) Model 2a; Culture III

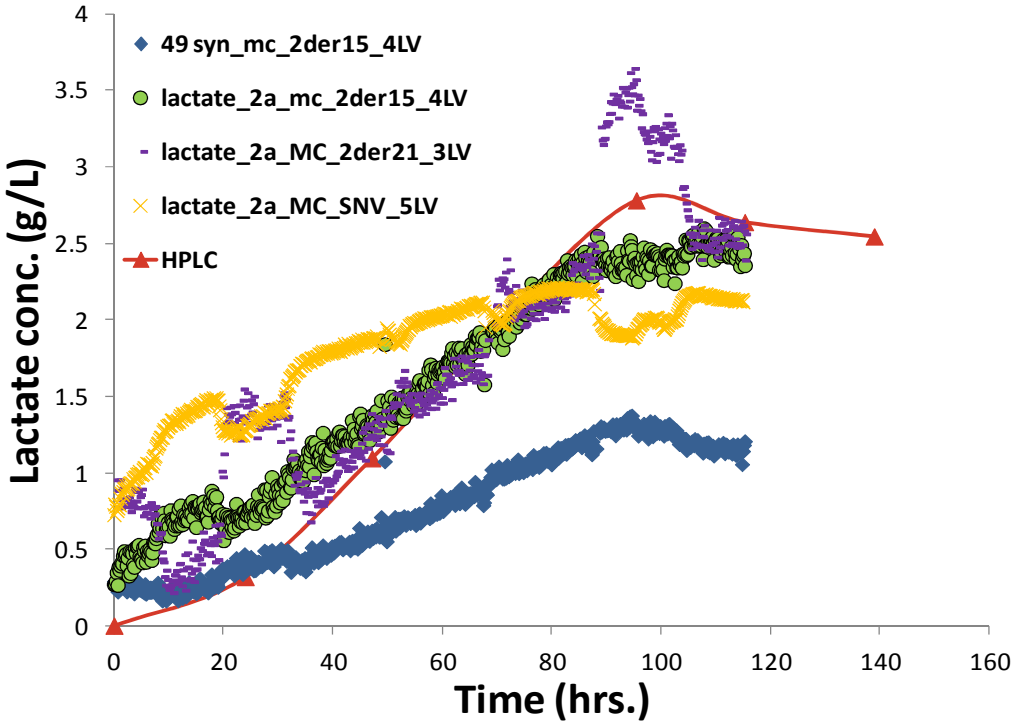


Fig. 4.3 (f) Model 2a; Culture III

Fig. 4.3 (a-f) Hybrid models type "2a" applied to cultures 1, 2 &

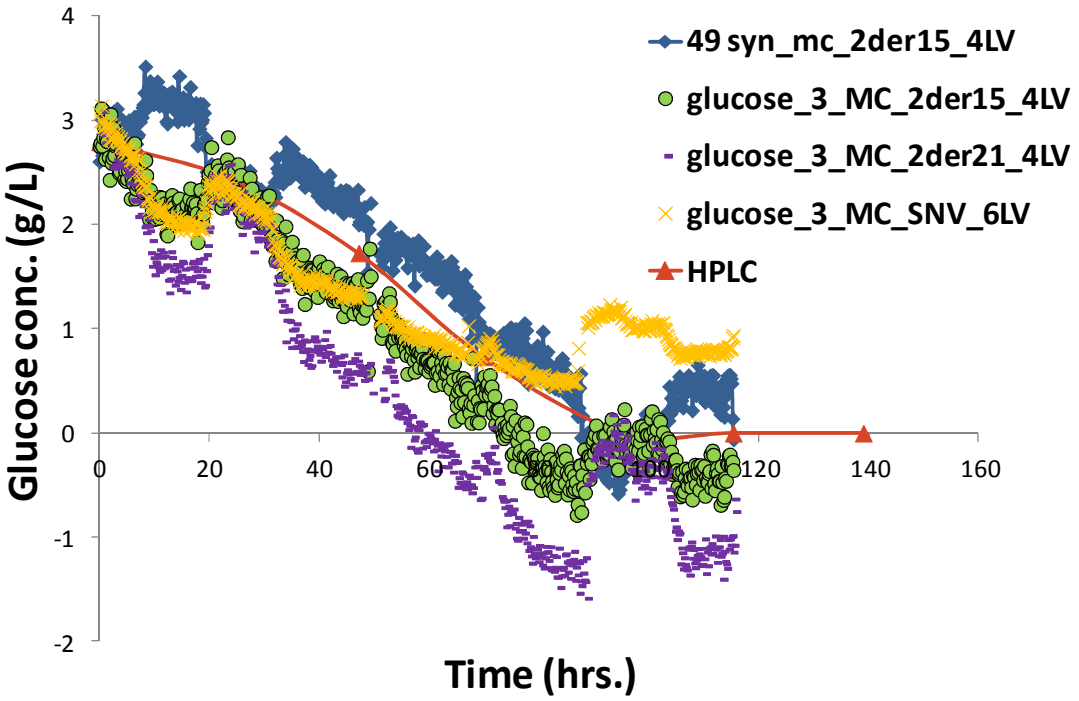


Fig. 4.4 (a) Model 3; Culture III

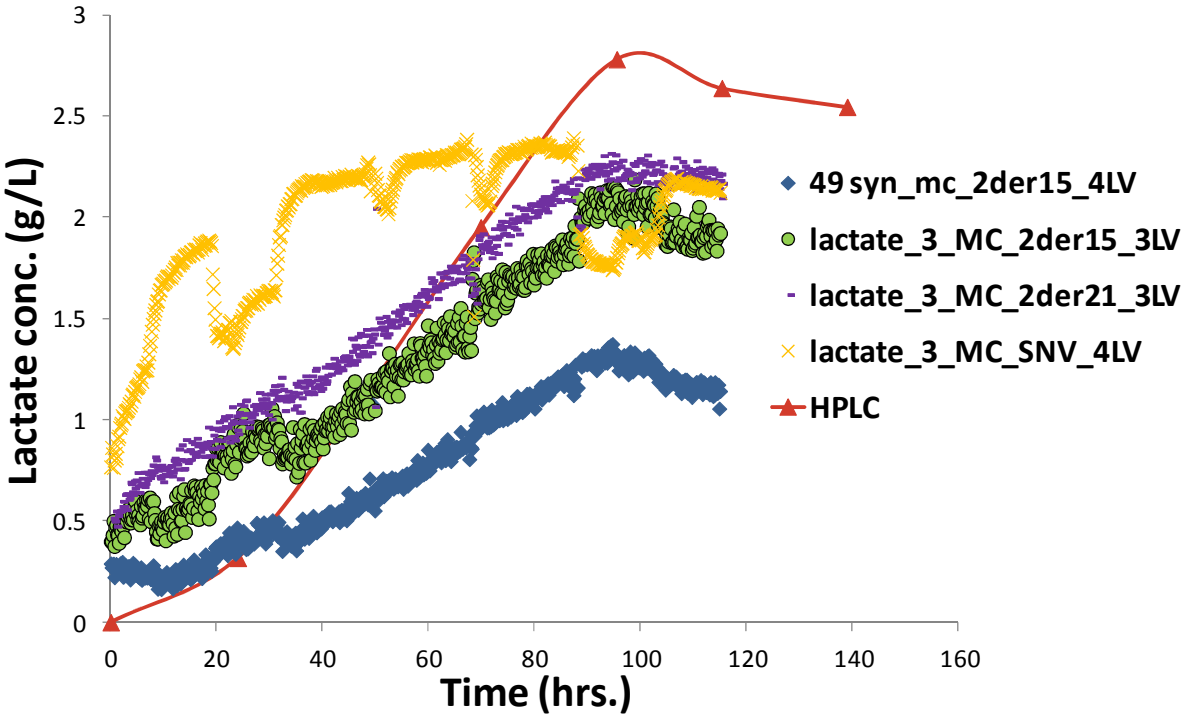


Fig. 4.4 (b) Model 3; Culture III

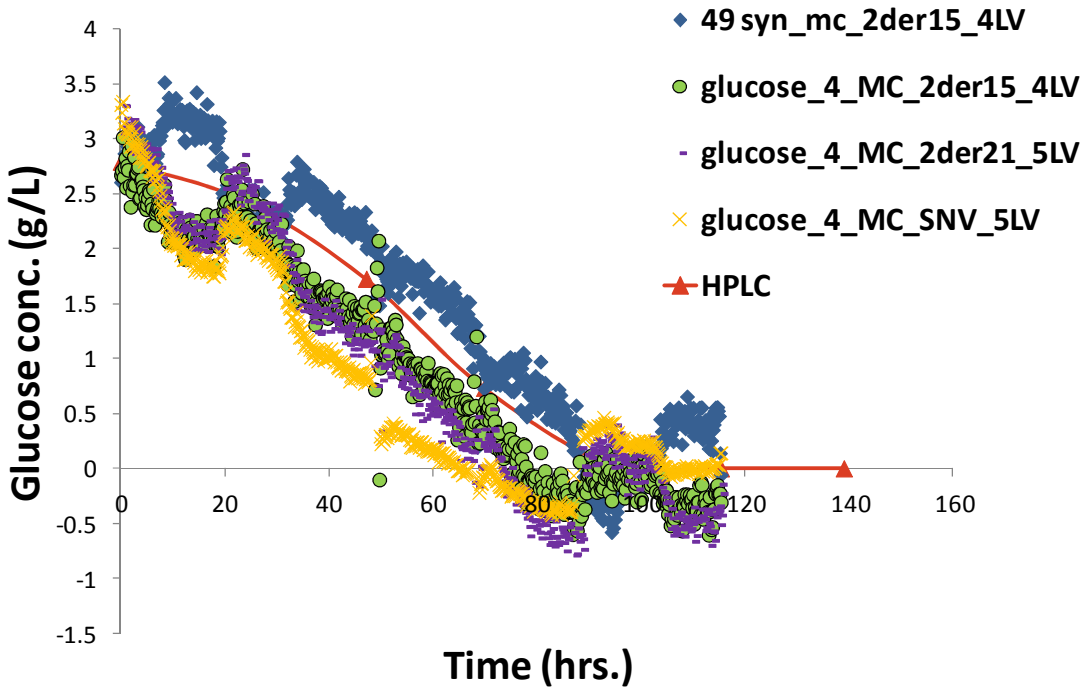


Fig. 4.4 (c) Model 4; Culture III

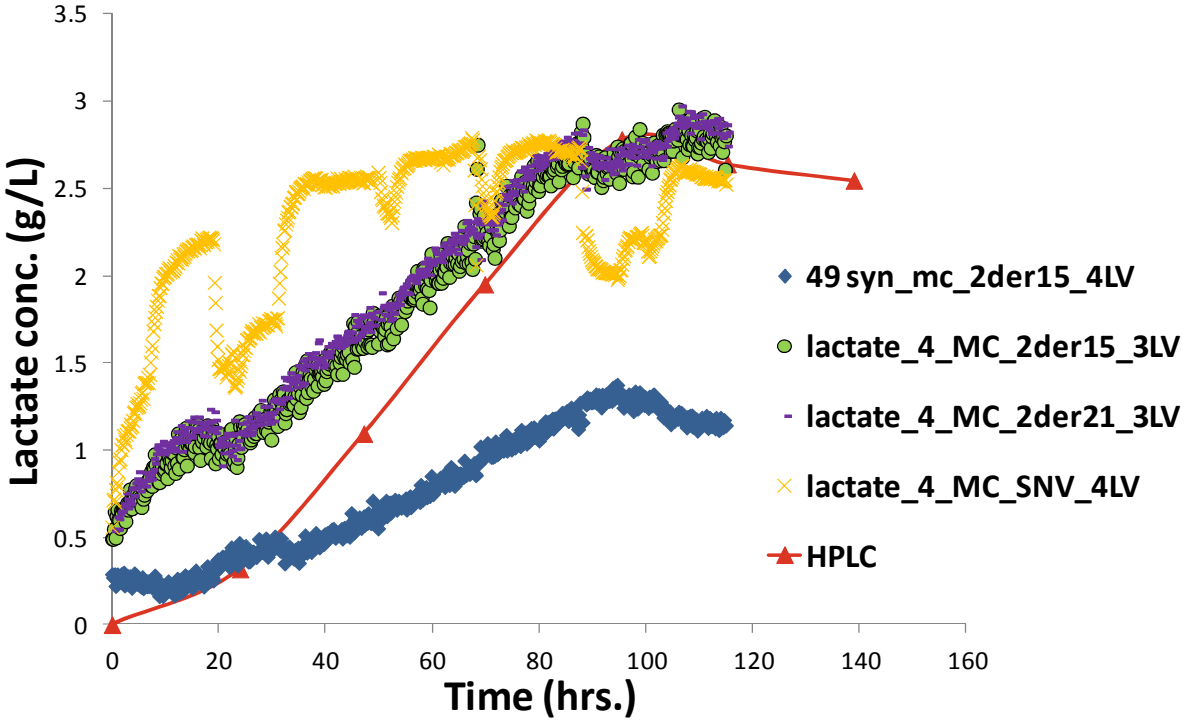


Fig. 4.4 (d) Model 4; Culture III

Fig. 4.4 (a-d) Hybrid models type "3 & 4" applied to culture III

Table 4.4 Glucose hybrid model results

Model	RMSEP (g/L)	Test set
49 syn_mc_2der15_4LV	0.73	in-situ spectra, cultures 1,2&3
glucose_**2a**_MC_2der15_4LV	0.65	in-situ spectra, cultures 1,2&3
glucose_**2a**_MC_2der21_5LV	0.77	in-situ spectra, cultures 1,2&3
glucose_**2a**_MC_SNV_5LV	1.17	in-situ spectra, cultures 1,2&3
49 syn_mc_2der15_4LV	0.29	off-line spectra, cultures 1,2&3
glucose_**2b**_MC_2der15_4LV	0.18	off-line spectra, cultures 1,2&3
glucose_**2b**_MC_2der21_6LV	0.21	off-line spectra, cultures 1,2&3
glucose_**2b**_MC_SNV_4LV	0.44	off-line spectra, cultures 1,2&3
49 syn_mc_2der15_4LV	0.14	in-situ spectra, culture 3
glucose_**3**_MC_2der15_4LV	0.34	in-situ spectra, culture 3
glucose_**3**_MC_2der21_4LV	0.72	in-situ spectra, culture 3
glucose_**3**_MC_SNV_6LV	0.63	in-situ spectra, culture 3
49 syn_mc_2der15_4LV	0.14	in-situ spectra, culture 3
glucose_**4**_MC_2der15_4LV	0.31	in-situ spectra, culture 3
glucose_**4**_MC_2der21_5LV	0.39	in-situ spectra, culture 3
glucose_**4**_MC_SNV_5LV	0.57	in-situ spectra, culture 3

Table 4.5 Lactate hybrid model results

Model	RMSEP (g/L)	Test set
49 syn_mc_2der15_4LV	1.21	in-situ spectra, cultures 1,2&3
lactate_**2a**_mc_2der15_4LV	0.46	in-situ spectra, cultures 1,2&3
lactate_**2a**_MC_2der21_3LV	1.02	in-situ spectra, cultures 1,2&3
lactate_**2a**_MC_SNV_4LV	0.69	in-situ spectra, cultures 1,2&3
49 syn_mc_2der15_4LV	1.05	off-line spectra, cultures 1,2&3
lactate_**2b**_MC_2der15_3LV	0.38	off-line spectra, cultures 1,2&3
lactate_**2b**_MC_2der21_3LV	0.36	off-line spectra, cultures 1,2&3
lactate_**2b**_MC_SNV_4LV	0.93	off-line spectra, cultures 1,2&3
49 syn_mc_2der15_4LV	0.96	in-situ spectra, culture 3
lactate_**3**_MC_2der15_3LV	0.49	in-situ spectra, culture 3
lactate_**3**_MC_2der21_3LV	0.46	in-situ spectra, culture 3
lactate_**3**_MC_SNV_4LV	0.87	in-situ spectra, culture 3
49 syn_mc_2der15_4LV	0.96	in-situ spectra, culture 3
lactate_**4**_MC_2der15_3LV	0.96	in-situ spectra, culture 3
lactate_**4**_MC_2der21_3LV	0.50	in-situ spectra, culture 3
lactate_**4**_MC_SNV_4LV	0.87	in-situ spectra, culture 3

The original glucose and lactate models were developed based on the identification of 8 common cell culture components (glucose, lactate, ammonia, glutamine, glutamate, sodium bicarbonate, phosphate and HEPES). Using these 8 components a cell culture environment was simulated and the calibration models developed. To optimise these models the effects of both unknown media components and products produced during the cell culture were accounted for in the model development. **Model type 2a** (Table 4.2), which consisted of synthetic samples and culture samples collected offline, was applied to the *in-situ* spectra of all cultures. Glucose results indicated an improvement on the RMSEP of 10.96% while the lactate error fell from 1.21g/L to 0.46g/L, resulting in a 61.98% improvement. Optimum models for predictions of in-situ glucose and lactate spectra were hybrid models utilising mean centering and Savitzky-Golay second derivative (filter width of 15) pre-treatments. The addition of actual culture samples to the calibration sets fortified the models and improved the accuracy of the model when applied to spectra collected in-situ in the reactor.

Model type 2b was applied to the offline spectra not included in the calibration set in order to determine if the *offline/at-line* predictions could be improved. It has already been seen that the original models preformed better when applied to offline/at-line culture spectra. Again, hybrid 2b models performed better than the original models. The error of prediction for glucose fell from 0.29g/L to 0.18g/L, a 37.93% improvement and for lactate fell from 1.05g/L to 0.36g/L, an improvement of 60.95%. In the case of glucose, a narrower filter width of 15 points in the second derivative pre-treatment performed best, while lactate results favoured a larger filter width of 21 points. These results highlight the importance of pre-treatment selection and

application and show different components with differing spectra may give better predictions with different spectral pre-treatments.

Both model types 2a and 2b indicate that the inclusion of real culture sample spectra, which contain unknown components due to the proprietary formulation of many cell culture media, create more robust and reliable models for both online and at-line purposes.

Model type 3 was developed in order to investigate if the addition of spectra collected in-situ could further enhance results. As with models 2a and 2b, the additional spectra used in this calibration set were obtained from actual cell cultures, therefore allowing for the effect of unknowns to be built into the model. However as the spectra used to augment the calibration set were collected in-situ, other factors, not accounted for in the previous hybrid models were incorporated into the model, such as environmental and external influences.

Table 4.4 shows the errors obtained for all models of type 3, for glucose. None of the hybrid models resulted in an improvement on the original model when applied to the selected validation set. Although data collected after glucose had reached its LOD was omitted, the trends exhibited in Fig. 4.1 showed that even at an early stage glucose spectra collected online were not reliable, and the predictive error associated with the online measurements was quite large (0.73g/L).

Therefore the addition of such unreliable spectra to the calibration set only served to increase noise in the models and the additional chemical information that improved the performance of Model 2a was made ineffective.

Fig. 4.1 indicates that although lactate concentrations fell short of the reference values consistently, the resulting online trends were reliable. Smooth, steady trending was observed, and the slight decrease in the lactate concentration on day 4 was evident in the predictions of in-situ spectra for all 3 cultures. Addition of in-situ spectra created more robust models which, unlike glucose, do not appear to be as affected by the environmental factors in the reactor possibly due to the fact that the main absorbance peaks for lactate are in a different region to those for glucose. All model type 3 hybrid models performed better than the original model, with the optimum model resulting in a 52.08% improvement.

Model type 4 was in essence, an extension of model type 3. The purpose of this model was to investigate if a sheer volume of calibration samples could improve already existing models. The training set used for model type 4 contained spectra collected from all possible scenarios – synthetic spectra, online spectra and offline spectra, thereby including all possible factors. Glucose and lactate results for model 4 exhibited the same trend as for model 3, i.e. glucose results were not improved while those for lactate were. This was as expected as the addition of the in-situ spectra to the calibration set had a large influence on the efficacy of the models.

The optimisation of the models highlighted areas where unreliable spectral data was obtained, but also where reliable data could be further worked upon to create improved and robust models (up to 62% improvement).

4.4 Conclusion

This study evaluated the application of mid-infrared spectroscopy to the upstream processing of a CHO DP12 cell line, for the purpose of monitoring glucose and lactate concentrations during the cultivation. Both in-situ and at-line methods were assessed using an FTIR immersion probe. Predictions from a calibration model developed with only synthetic mixtures of typical cell culture media components and products of metabolism, exhibited the expected trends of glucose consumption and lactate production, however the prediction errors were considerable for in-situ application, and when glucose fell below its LOD, results became unreliable. Therefore in-situ, real-time application of this system may only be considered practicable where the concentration of these components remains above the LOD, such as in a fed-batch application or where simple trending of substrates and metabolites is required.

Using the same calibration models as applied in-situ, an at-line study was carried out. This application proved more reliable as prediction errors for both glucose and lactate fell, indicating that at-line application may be a viable means of monitoring these components of interest. Despite the improvement in results for at-line application, this method also poses a number of problems if it is to be incorporated into a PAT system. It requires the removal of the sample from the reactor and while it is faster than typical laboratory analysis, it still requires a sterile sampling system, one of the major advantages of online monitoring. An at-line monitoring system also cannot be integrated into a control system to the same level as an online monitoring system.

An investigation was carried out to determine if the glucose and lactate calibration models used for in-situ and at-line analysis could be optimised. This optimisation

study resulted in a number of hybrid models. In all cases where hybrid models were developed using spectra collected offline/at-line, the optimum model was always a hybrid model. This highlights the importance of including spectra collected from actual culture samples in the model training set.

When in-situ spectra were used to augment the calibration set, for glucose models, predictions did not improve. This result emphasises an important point, which is that a model is only as good as the spectra used to create it i.e. the glucose models containing spectra collected in-situ, did not perform as well as those without. The addition of the in-situ samples increased noise rather than true chemical information. In the case of lactate, all hybrid models were an improvement on the original model.

In summary, this method can be applied in-situ for monitoring and potentially control purposes and accuracy of the predictions can be improved by the addition of actual culture samples to the training set and also by varying pre-treatments used.

4.5 References

1. M. J. T. Carrondo, "How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology?" Biotechnol. J. 2012. **7**(12): 1522-1529

2."Process Analytical Technology; Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries", In: K. Bakeev, editor. Chichester, UK: John Wiley and Sons, 2010.

3. P. A. David, R. Roginski, S. Doherty, and J. Moe, "The impact of process analytical technology in pharmaceutical chemical process development," J. Process Anal. Chem. 2005. **9(1)**: 1-5

4. U.S. Department of Health and Human Services: Food and Drug Administration, "Pharmaceutical cGMPS for the 21st Century — A Risk-Based Approach - Final Report" (2004).

5. U.S. Department of Health and Human Services: Food and Drug Administration, "Guidance for Industry - Q8 Pharmaceutical Development" (2006)

6. D. Landgrebe, C. Haake, T. Höpfner, S. Beutel, B. Hitzmann, T. Scheper, M. Rhiel, and K. Reardon, "On-line infrared spectroscopy for bioprocess monitoring," Appl. Microbiol. Biotechnol. 2010. **88**(1): 11-22

7. M. Dabros, M. Amrhein, P. Gujral, and U. von Stockar, "On-line Recalibration of Spectral Measurements using Metabolite Injections and Dynamic Orthogonal Projection," Appl. Spectrosc. 2007. **61**(5): 507-513

8. H. Kornmann, S. Valentinotti, I. Marison, and U. von Stockar, "Real-time update of calibration model for better monitoring of batch processes using spectroscopy," Biotechnol. Bioeng. 2004. **87**(5): 593-601

9. S. A. Arnold, J. Crowley, N. Woods, L. M. Harvey, and B. McNeill, "In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation," Biotechnol. Bioeng. 2003. **84**(1): 13-19

10. C. B. Lewis, R. J. McNichols, A. Gowda, and G. L. Coté, "Investigation of Near-Infrared Spectroscopy for Periodic Determination of Glucose in Cell Culture Media in Situ," Appl. Spectrosc. 2000. **54**(10): 1453-1457

11. E. Tamburini, G. Vaccari, S. Tosi, and A. Trilli, "Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fibre probe," Appl. Spectrosc. 2003. **57**(2): 132-138

12. J. Henriques, S. Buziol, E. Stocker, A. Voogd, and J. Menezes, "Monitoring Mammalian Cell Cultivations for Monoclonal Antibody Production Using Near-Infrared Spectroscopy," In: Govind Rao, editor. Optical Sensor Systems in Biotechnology. Berlin Heidelberg: Springer, 2010. **116**, 73-97

13. P. Roychoudhury, R. O'Kennedy, B. McNeil, and L. M. Harvey, "Multiplexing fibre optic near infrared (NIR) spectroscopy as an emerging technology to monitor industrial bioprocesses," Anal. Chim. Acta**.** 2007. **590**(1): 110-117

14. N. Petersen, P. Ödman, A. E. C. Padrell, S. Stocks, A. E. Lantz, and K. V. Gernaey, "In situ near infrared spectroscopy for analyte-specific monitoring of glucose

and ammonium in streptomyces coelicolor fermentations," Biotechnol. Prog. 2010. **26**(1): 263-271,

15. M. Rhiel, P. Ducommun, I. Bolzonella, I. Marison, and U. von Stockar, "Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements," Biotechnol. Bioeng. **77**, 174 (2002). 2002. **77**(2): 174-185

16. M. H. Rhiel, M. Amrhein, I. W. Marison, and U. von Stockar, "The influence of correlated calibration samples on the prediction performance of multivariate models based on mid- infrared spectra of animal cell cultures," Anal. Chem. 2002. **74**(20): 5227-5236

17. D. L. Doak and J. A. Phillips, "In Situ Monitoring of an *Escherichia coli* Fermentation using a Diamond Composition ATR Probe and Mid-infrared Spectroscopy," Biotechnol. Prog. 1999. **15**(3): 529-539

18. J. Schenk, C. Viscasillas, I. W. Marison, and U. von Stockar, "On-line monitoring of nine different batch cultures of E. coli by mid-infrared spectroscopy, using a single spectra library for calibration," J. Biotechnol. 2008. **134**(1-2): 93-102

19. J. Schenk, I. W. Marison, and U. von Stockar, "A simple method to monitor and control methanol feeding of Pichia pastoris fermentations using mid-IR spectroscopy," J. Biotechnol. 2007. **128**(2): 344-353

20. J. Schenk, I. W. Marison, and U. von Stockar, "Simplified Fourier-transform mid-infrared spectroscopy calibration based on a spectra library for the on-line monitoring of bioprocesses," Anal. Chim. Acta. 2007. **591**(1): 132-140

21. A. E. Cervera, N. Petersen, A. E. Lantz, A. Larsen, and K. V. Gernaey, "Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation," Biotechnol. Prog. 2009. **25**(6): 561-1581

22. M. R. Riley, H. M. Crider, M. E. Nite, R. A. Garcia, J. Woo, and R. M. Wegge, "Simultaneous Measurement of 19 Components in Serum-Containing Animal Cell Culture Media by Fourier Transform Near-Infrared Spectroscopy," Biotechnol. Prog. 2001. **17**(2): 376-378

23. M. J. McShane and G. L. Cote, "Near-Infrared Spectroscopy for Determination of Glucose, Lactate, and Ammonia in Cell Culture Media," Appl. Spectrosc. 1998. **52**(8): 1073-1078

24. B. Finn, L. M. Harvey, and B. McNeil, "Near-infrared spectroscopic monitoring of biomass, glucose, ethanol and protein content in a high cell density baker's yeast fed-batch bioprocess," Yeast. 2006. **23**(7): 507-517

25. R. Foley, S. Hennessy, and I. W. Marison, "Potential of Mid-Infrared Spectroscopy for On-Line Monitoring of Mammalian Cell Culture Medium Components," Appl. Spectrosc.2012. **66**(1): 33-39

26. J. A. Munoz and R. G. Brereton, "Partial factorial designs for multivariate calibration: extension to seven levels and comparison of strategy," Chemometrics Intell. Lab. Syst. 1998. **43**(1-2): 89-105

27. A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," Anal. Chem. **36**(8): 1627-1639

28. T. Naes, T. Isaksson, T. Fearn, and T. Davies, "*A User-Friendly Guide to Multivariate Calibration and Classification*" Chichester, UK, NIR Publications, 2002. p. 344.

29. R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra," Appl. Spectrosc.1989. **43**(5): 772-777

30. S. Sivakesava, J. Irudayaraj, and D. Ali, "Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques," Process Biochem. 2001. **37**(4): 371-378

31. G. Mazarevica, J. Diewok, J. R. Baena, E. Rosenberg, and B. Lendl, "On-line fermentation monitoring by mid-infrared spectroscopy," Appl. Spectrosc. 2004. **58**(7): 804-810

32. P. Roychoudhury, L. M. Harvey, and B. McNeil, "At-line monitoring of ammonium, glucose, methyl oleate and biomass in a complex antibiotic fermentation process using attenuated total reflectance-mid-infrared (ATR-MIR) spectroscopy," Anal. Chim. Acta**.** 2006. **561**(1-2): 218-224

33. A. M. C. Davies and T. Fearn, "Back to basics: calibration statistics," Spectrosc. Eur. 2006. **18**(2): 31-32

34. A. C. Olivieri, N. M. Faber, J. Ferré, R. Boqué, J. H. Kalivas, and H. Mark, "Uncertainty estimation and figures of merit for multivariate calibration," Pure Appl. Chem. 2006. 78: 633–661

Chapter 5

Study 3: The use of Mid-infrared spectroscopy as an on-line PAT tool in total and recombinant protein monitoring

Abstract:

The aim of this study was to establish the feasibility in using mid-infrared (MIR) spectroscopy as on online monitoring tool for protein production at various stages in a bioprocess. This exploratory work involved a number of steps to determine firstly, the possibility of using MIR for total and recombinant protein detection and secondly, the efficacy of this technique for online use. An initial principal component analysis (PCA) was performed using 5 common proteins and the recombinant protein of interest. Results indicated that the MIR immersion probe used was capable of distinguishing between the differing proteins, highlighting the potential of MIR as a qualitative process analytical technology (PAT) tool for protein detection and characterisation. Using process samples provided by an industrial collaborator several partial least squares regression (PLS) models were developed to establish the possibility of using this technique as a quantitative PAT tool in online protein monitoring. Models capable of evaluating total and recombinant protein were constructed. Results indicated a greater accuracy in the prediction of the total protein with a minimum percentage error of prediction (PEP) of 2.39%. The smallest PEP for the recombinant protein was found to be 6.66%. These results highlight the existing potential for the application of this technique to in-situ protein monitoring. Due to the limited available size of the calibration set, this study also raised questions as to the

best means of evaluation of results. This work outlined 2 possible methods for model evaluation; standard error of cross validation and standard error of prediction. Each technique was examined and results based on these methods were presented for comparative purposes. Finally, an investigation into the impact of 2 detergents on model performance was completed. As detergents are frequently used in the course of the bioprocess for e.g. protein solubilisation or virus inactivation, a PCA was completed to determine if the presence of detergent would be likely to affect protein predictions. This analysis indicated that detergents were likely to impact results, particularly at concentrations at the higher end of the typically used range.

5.1 Introduction

Traditional protein quantification techniques such as SDS-Page are laborious, requiring considerable sample preparation and processing time. Due to the amount of interaction with the process sample, these methods are also open to error at any one of the many analysis steps. Current, conventional methods also require the removal of a sample from the bioreactor and result in sample destruction. Mid-infrared spectroscopy offers an alternative to protein quantification.

MIR is a well established technique in the determination of protein structure.[1-3] The amide I band located at 1700 – 1600 cm^{-1} provides a significant amount of information.[4,5] The C=O stretch at the amide I band is affected by the strength of the hydrogen bonds between the C=O bond and the N-H group.[2] The resulting variation in the absorption of proteins in this region has been utilised to determine secondary protein structures.[5] Gross-Selbeck et al have further shown that the intensity of the absorption at the amide I band corresponds to the quantity of protein in the sample.[6]

Based on this knowledge, MIR has huge potential in protein analysis presenting a serious case for its application as a process analytical technology tool (PAT).

Infrared spectroscopy is a powerful tool in bioprocess monitoring.[7-9] It is capable of simultaneously monitoring several of the components present in a cell culture medium. Both near infrared (NIR) and MIR have been used for analyte and metabolite monitoring,[10-15] with NIR also being applied to cultures as a biomass monitor.[16] Previous studies in the use of MIR in protein detection and classification have focused on this technique as a rapid off-line method,[2,17,18] with few examining its in-situ capabilities.[19]

Infrared spectroscopy can be used in-situ. In such applications it is non-invasive and eliminates the need for sample removal. In bioprocessing, where sterility is of the utmost importance, obtaining samples without compromising the sterility of the system is always a major concern. Non-invasive methods, like in-situ MIR, reduce the risk of culture contamination.[20] In addition, sample preparation is not required and data is obtained instantaneously. These features meet much of the criteria used to define a Process Analytical Technology (PAT) tool, as outlined by the FDA in their "PAT Guidance for Industry" framework.[21]

The purpose of this study was to examine any existing potential for protein detection using an in-situ MIR instrument. The investigative work completed used in-process industrial samples to develop PCA and PLS models to examine the feasibility of using online MIR, to both qualify and quantify total and recombinant protein. This study is a precursor to further work on model development and optimisation. It does not present

a detailed evaluation of optimum chemometric models but rather establishes a methodology for preliminary, exploratory work in determining the efficacy of online MIR as a PAT tool for protein detection and monitoring. In addition, the detergent study highlighted the possible difficulties the technique may encounter during the processing of recombinant protein. The impact of 2 commonly used detergents on spectral data was investigated to identify any potential influences on the PLS model predictions. Finally, results were evaluated using both root mean square error of cross-validation (RMSECV) and root mean square error of prediction (RMSEP) in an attempt to answer questions as to the best evaluation method given the size of the training and validation sets.[22]

To the author's knowledge, no other study exits in which recombinant and total protein have been quantified in industrial samples using an MIR immersion probe capable of online monitoring.

5.2 Materials & Methods

5.2.1 Preliminary analysis

Five commonly used proteins, bovine serum albumin (BSA), human serum albumin (HSA), amylase, pepsin and lipase, (Sigma Aldrich, Ireland, Ltd.) in addition to the recombinant protein of interest, were selected, on which to perform a preliminary analysis. Standard solutions of each protein, at concentrations of 10, 5 and 2.5 g/L, were made. These solutions were then scanned in triplicate using a Fourier transform mid-infrared ReactIR iC10 instrument with MCT detector (Mettler Toledo AutoChem, Inc., Columbia, US) against a background of deionised water. A K6 conduit 16mm immersion probe was used with a fixed path length of approx. 1-2 μm and a diamond

ATR crystal with 6 internal reflections. The detector and probe were purged with nitrogen gas continuously. The mean spectrum of each triplicate scan was calculated and imported into MatLab (v7.9.0.529 (R2009b), The MathWorks Inc., Cambridge, UK). A scan of each protein at 10 g/L was taken and spectra were plotted for preliminary examination. This allowed for spectral similarities and differences to be identified while also establishing the usable wavenumber range for further work. The PLS toolbox for MatLab (V6.2 Eigenvector Research Inc, Wenatchee, WA, US) was used to complete a principle component analysis (PCA) to determine if it was possible to distinguish between the six different proteins. Pre-treatments employed for PCA investigation were mean centering combined with normalisation and mean centering combined with standard normal variate (SNV). Both these pre-treatments are effective at scaling and normalising spectra.

5.2.2 Samples and reference analysis

Recovery process samples were obtained from an industrial collaborator. These samples contained the recombinant protein of interest, in addition to a large number of unknown host cell proteins. Also present in these untreated samples were cell particulates, as samples were previously lysed with acid, and sodium hydroxide, used for pH adjustment. Using these samples, 3 sample types were generated: untreated samples, consisting of the samples in their original form; supernatant samples, resulting from untreated sample centrifugation and containing aqueous based fermentation broth, sodium hydroxide, proteins and other components soluble in water; re-dissolved pellet samples generated by dissolution of the remaining pellet in a Tris (Sigma Aldrich, Ireland, Ltd.)/EDTA (VWR International, West Chester, Pennsylvania

(USA)) extraction buffer and containing, proteins which were not previously dissolved and insoluble cell particulates.

An estimation of the concentration of NaOH present in the untreated and supernatant samples was calculated from process data available (as samples were supplied by an industrial partner the exact composition of other components was not provided and was calculated based on limited data released by the company) and both samples types were scanned in triplicate against a background of deionised water and NaOH, at this concentration. The re-dissolved pellet samples were scanned against a background of the Tris/EDTA extraction buffer used to dissolve the pellet. All samples were vortexed prior to scanning to ensure homogeneity.

For reference analysis a Bradford assay (Sigma Aldrich Ireland Ltd.) was used to quantify the total protein present in the samples. Assays were performed on all the supernatant and re-dissolved pellet samples and the total protein within the untreated samples was quantified by summing the results of these. The recombinant protein was determined via SDS-PAGE using precast gels, NuPAGE Novex Bis-Tris Gel 4-12% (Invitrogen, Carlsbad, CA, USA). Again, supernatant and re-dissolved pellet samples were used to quantify the recombinant protein present in the untreated samples.

5.2.3 Quantitative model development

Partial least squares regression (PLS) models were developed to quantify both the total protein and the recombinant protein present in the samples. Untreated in-process samples were available from 12 batches and from these 12 samples, the 3 samples types were generated as described above. Samples from 3 of these batches were

randomly selected for model validation and the samples from the 9 remaining batches were used as the calibration set for the various models. Four model types were generated: the first model type was developed using all forms of samples, i.e. untreated, supernatant and re-dissolved pellet samples; the second, untreated samples only; the third, supernatant samples only and the final model type was developed using re-dissolved pellet samples only. Model type 1 therefore had 27 samples for development and 9 for validation. The remaining 3 model types were created using 9 samples and validated with 3 samples. Figure 5.1 outlines the experimental design employed for total protein model development. Recombinant protein model development followed the same procedure however derivative order, filter width and number of latent variables differed.

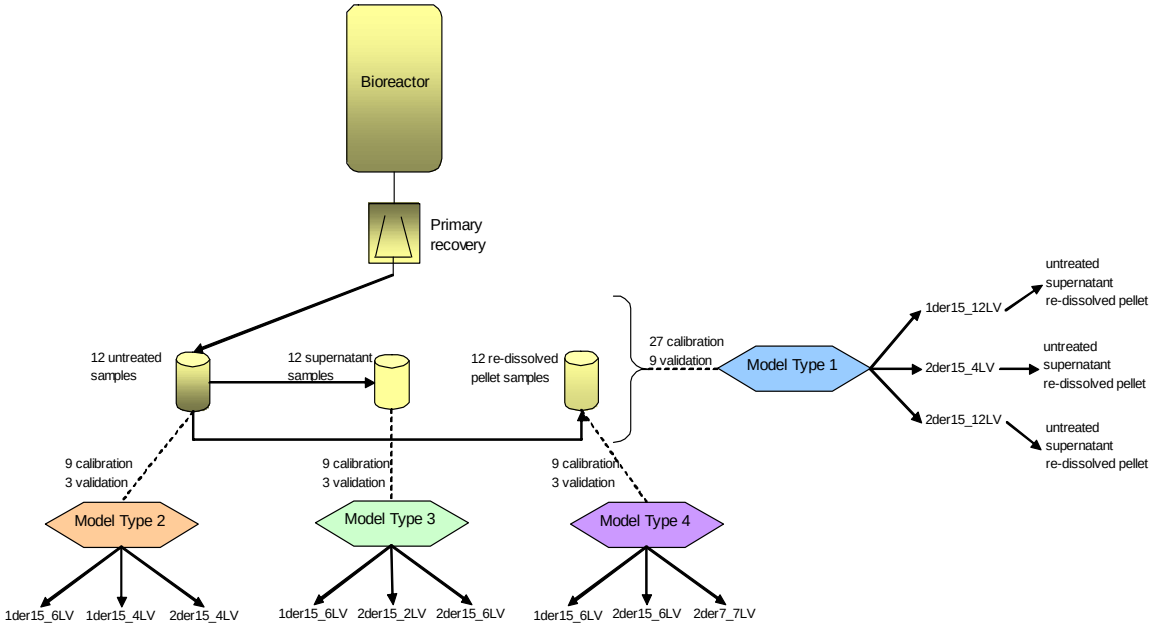


Fig. 5.1 Schematic of experimental design for total protein model development

Following the preliminary analysis the usable region of the spectrum was identified to be 1700-900cm^{-1}. Only spectral data lying within this region was selected for model development. All other extraneous data was removed. This ensured that all models were developed based on absorbance data relevant to the protein absorbance area of the spectrum and eliminated all other regions which exhibited considerable spectral noise and would negatively impact the models. All spectra were mean centred. First and second derivative pre-treatments, of filter widths 7, 15 and 21 were then applied, calculated using Savitzky-Golay filters.[23] These pre-treatments were selected as they can eliminate the effect of linear baseline off-sets and smooth noisy spectra.[24] Results tables indicate whether first or second order derivatives were used, in addition to specifying the filter width. These are denoted by XderY, where X is the Savitzky-Golay Xth derivative pre-treatment and Y, the filter width. For each of the 4 model types, 3 models were developed for both the total protein and recombinant protein predictions. These models varied in the number of latent variables (LV) used and the order of the derivative and filter widths of the Savitzky-Golay pre-treatment.

Model efficacy was evaluated in two ways. The root mean squared error of prediction, (RMSEP), was used to assess the models prediction ability using 3 sample batches not used in the development of the models.[25] Percentage error of prediction (PEP) values were calculated in order to put the RMSEP values in perspective. The PEP was calculated by dividing the RMSEP by the average total or recombinant protein concentration in the 3 test samples. Using a calibration or training set to develop a model, followed by the use of a completely independent validation set to test the model is a procedure typically used and widely applied in the field of chemometrics. However this methodology comes with a caveat that should be duly noted, that is, the

impact of the size of the calibration and validation sets on the overall model results. Where possible, calibration sets should be large enough to contain all possible forms of variation within the samples to ensure the model accounts for all scenarios, both expected and unexpected. The validation set should be capable of unearthing any model weaknesses, and should prove a difficult test of the model. Due to limitations in sample availability this is not always feasible. Kramer acknowledges this and suggests that in such circumstances other tests, such as cross validation, be employed as a means of evaluating model accuracy.[22] As this study was limited by the number of samples available for training and validation set compilation, all results were also evaluated based on the root mean squared error of cross-validation (RMSECV) and for model types 2, 3 and 4, where it was possible to associate a percentage value, the percentage error of cross validation (PECV) was also calculated. 'Leave-one-out' cross validation was the selected cross validation technique and was used in the development of all models generated. This method is widely used where small sample sets are in use. This procedure removed one sample from the calibration set, developed a PLS model from the remaining samples, and then applied this model to the removed sample to predict the concentration. Each sample in turn was removed and its concentration predicted. A prediction error for each of the samples was calculated and the RMSECV was determined by combining these errors to generate a standard error.

5.2.4 Detergent investigation

A further, exploratory study was carried out to investigate the impact of the addition of 2 standard, industrially used detergents, detergent A and detergent B, on untreated

samples. As detergent is likely to be introduced at the raw, untreated stage, it was to the untreated samples that detergent was added.

The following samples were scanned and spectra plotted for preliminary examination:

- Detergent A
- Detergent B
- Untreated sample, prior to the addition of any detergent
- Untreated sample with 3% detergent A
- Untreated sample with 3% detergent B

The area of interest was again localised to 1,700 cm^{-1} – 900 cm^{-1}.

To 3 aliquots of an untreated sample, concentrations of 0.006%, 1.5% and 3% of detergent A were added respectively. Similarly, concentrations of 0.3%, 1.5% and 3% of detergent B were added to another 3 aliquots of the same untreated sample. The selected concentrations were based on recommended range of use for each of the detergents.[26] All 6 samples were scanned in triplicate and the mean of each spectrum obtained. A qualitative analysis was carried out by performing a PCA. In addition to these 6 spectra collected, 2 spectra, one of each detergent, and six spectra, of untreated samples without any detergent present, were added to the data set. Pre-treatments used in advance of the PCA were mean centering with SNV and mean centering with normalisation.

5.3 Results and discussion

5.3.1 Preliminary analysis

The MIR region is considered to lie between 4000-400 cm^{-1}. The ATR crystal of the probe used in this study absorbs over the range 2250-1950 cm^{-1}, therefore creating an instrument 'blind spot'. At regions of the spectrum above 2250 cm^{-1} and below 900

cm^{-1} spectra exhibited considerable noise and were deemed unusable. Spectra immediately adjacent to the lower end of the 'blind spot' were also noisy. This effectively reduced the usable wavenumber range to 1700-900 cm^{-1}. Figure 5.2 shows the absorbance, of all 6 proteins investigated, over this region.

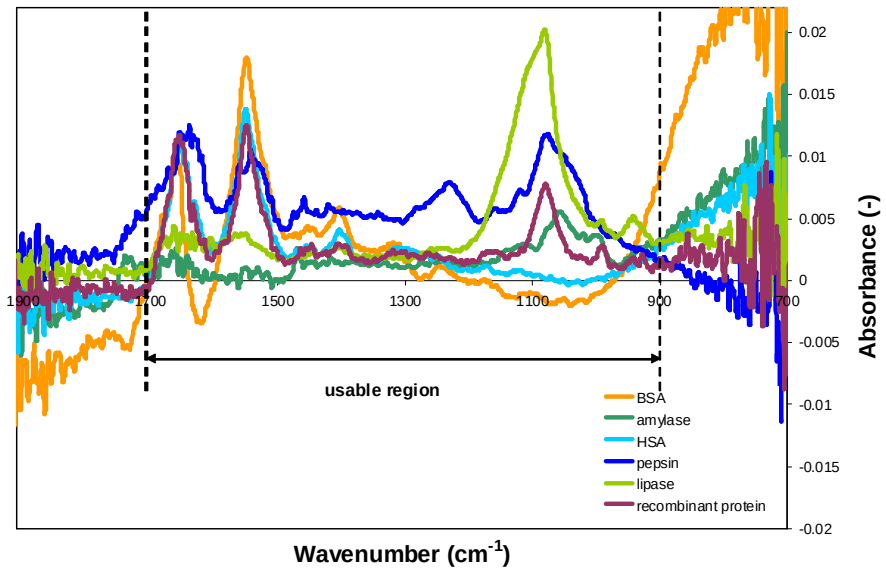


Fig. 5.2 Usable MIR spectral region for protein absorbance

The infrared spectrum of protein is characterised by a set of absorption regions known as the amide modes.[3] In the case of the reduced MIR spectral region used here, absorption at the amide I and amide II modes was clearly visible at ~1650 cm^{-1} and ~1550 cm^{-1} respectively. Six overlapping peaks, representing each of the 6 proteins, were evident at these locations on the spectrum, highlighting the difficulties encountered when attempting to characterise proteins using MIR. Although each protein was present at the same concentration, the intensity of the vibration varied. This occurrence can be exploited to determine varying protein structures.[1] At the lower end of the spectrum shown in figure 5.2, another series of peaks was observed,

however these peaks did not entirely overlap at the same point. Pepsin and the recombinant protein reached a maximum at 1076 cm^{-1}, amylase at 1053 cm^{-1} and lipase at 1080 cm^{-1}, while BSA and HSA did not exhibit any peaks at all. This observation supported the notion that it was possible to adequately distinguish between all six proteins and a PCA was carried out to determine this. Figure 5.3 shows the optimum PCA, where separation of differing proteins, and clustering of the same protein, but varying concentrations, occurred.

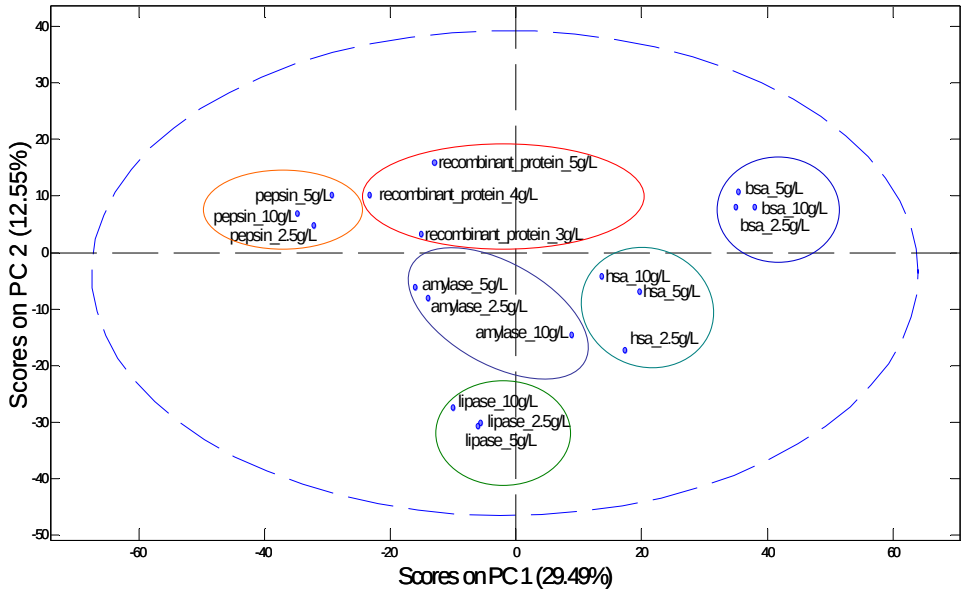


Fig. 5.3 Scores plot of 5 common proteins and the recombinant protein of interest

The spectral data used for this PCA underwent mean centering and SNV pre-treatments before completing the analysis. Other pre-treatments were also investigated such as first and second derivatives however mean centering combined with SNV was found to be the most effective. This preliminary study enabled identification of the usable region of the MIR spectrum for protein analysis. It also highlighted the ability

of MIR to distinguish between differing proteins, supporting the development of PLS calibration models for total and recombinant protein quantification.

5.3.2 Quantitative model development

For each model developed for model type 1, the RMSECV values for untreated, supernatant and re-dissolved pellet samples were the same. This is because the RMSECV is based on the cross validation model error and not a predictive error. Therefore it was the same 3 models that were developed and applied to all 3 sample types. Also, it was not possible to calculate a PECV for model type 1 as the actual differences in the concentration ranges of all 3 sample types were not comparable and so would not provide rational PECV results. In the case of predictive errors, PEP was calculated. Here a validation set containing all 3 sample types was used and the models applied. Each prediction was therefore associated with a sample type and hence a specific RMSEP for each sample type could be calculated and in addition, a PEP. This procedure was applied to both total and recombinant protein evaluations.

Models 2, 3 and 4 were each developed for a specific sample type and as such should have a greater degree of accuracy than model type 1 when applied to their associated sample types. This is evident by comparison between Table 5.1 and Table 5.2, which outline the total protein cross validation and prediction errors. Lowest predictive errors for model type 1 of the untreated and supernatant models were 10.4% and 40.23% respectively which remained above the highest predictive errors for the specific untreated (model type 2) and supernatant (model type 3) models of 6.97% and 8.7%. This trend was also observed for RMSECV values where the lowest RMSECV for model type 1 of 3.2961 g/L was greater than the highest RMSECV values for

model types 2 and 3, of 2.7282 g/L and 1.1091 g/L respectively. A direct comparison between re-dissolved pellet RMSECV and RMSEP values shows that model type 4, which was specific to the re-dissolved pellet samples, did not always perform better than model type 1, however the lowest RMSECV and RMSEP values in all cases resulted from models generated from specific sample types i.e. model types 2, 3 and 4.

TABLE 5.1 Total protein errors for Model Type 1

Sample Type	Model	RMSECV (g/L)	RMSEP (g/L)	PEP (%)
Untreated	1der15_12LV	3.2961	2.8855	18.80
	2der15_4LV	4.3027	2.0649	13.46
	2der15_12LV	4.6434	1.5955	10.40
Supernatant	1der15_12LV	3.2961	1.1224	46.31
	2der15_4LV	4.3027	0.9751	40.23
	2der15_12LV	4.6434	1.3059	53.89
Re-dissolved Pellet	1der15_12LV	3.2961	0.6368	5.29
	2der15_4LV	4.3027	1.6662	13.85
	2der15_12LV	4.6434	0.4591	3.81

TABLE 5.2 Total protein errors for Model Types 2, 3 & 4

Model Type	Sample Type	Model	RMSECV (g/L)	PECV (%)	RMSEP (g/L)	PEP (%)
Model Type 2	**Untreated**	1der15_6LV	2.0902	13.81	1.0701	6.97
		1der15_4LV	2.0922	13.83	1.0391	6.77
		2der15_4LV	2.7282	18.03	0.7633	4.97
Model Type 3	**Supernatant**	1der15_6LV	1.1091	35.96	0.1127	3.69
		2der15_2LV	0.5747	18.63	0.2213	7.24
		2der15_6LV	0.7507	24.34	0.2651	8.67
Model Type 4	**Re-dissolved Pellet**	1der15_6LV	2.9132	25.56	0.2795	2.39
		2der15_6LV	3.8104	33.44	0.4757	4.07
		2der7_7LV	4.0431	35.48	0.6915	5.91

Direct comparison of Table 5.1 and Table 5.2 shows that despite the evaluation method applied; RMSECV or RMSEP, model type 1 did not attain the same level of accuracy as model types 2, 3 and 4. Therefore, although the training sets for model types 2, 3 and 4 were one third the size of that used for model type one, the specificity of the samples used created more accurate models. However this does not infer that increased accuracy results in more robust models, and it is imperative that variation be

built into all models to protect against the occurrence of outliers which could cause the model to fail.[22] Although not possible in this feasibility study due to limitations in sample availability, the training set should be large enough to represent all components present at a number of concentration levels, ranging from the minimum to the maximum concentrations likely to be encountered, if the calibration is to be applied to, and trusted in, the monitoring of a bioprocess.

Recombinant protein errors, shown in Table 5.3 and Table 5.4, followed a similar trend to those of the total protein, in that the highest errors for the specific model types (model types 2, 3 and 4) were all considerably lower than the lowest errors of the general model type 1. This was true of both the RMSECV and RESEP values and in this instance, there were no exceptions.

TABLE 5.3 Recombinant Protein errors for Model Type 1

Sample Type	Model	RMSECV (g/L)	RMSEP (g/L)	PEP (%)
Untreated	1der15_7LV	0.1975	0.1035	17.30
	2der15_4LV	0.2796	0.1167	19.50
	2der7_7LV	0.2400	0.1724	28.80
Supernatant	1der15_7LV	0.1975	0.1943	68.60
	2der15_4LV	0.2796	0.1809	63.89
	2der7_7LV	0.2400	0.1098	38.77
Re-dissolved Pellet	1der15_7LV	0.1975	0.0905	26.23
	2der15_4LV	0.2796	0.0979	28.38
	2der7_7LV	0.2400	0.0551	15.98

TABLE 5.4 Recombinant Protein errors for Model Types 2, 3 & 4

Model Type	Sample Type	Model	RESECV (g/L)	PECV (%)	RMSEP (g/L)	PEP (%)
Model Type 2	**Untreated**	1der15_4LV	0.0967	14.96	0.0754	12.59
		2der15_4LV	0.0890	13.76	0.0848	14.17
		2der7_4LV	0.1210	18.72	0.0908	15.18
Model Type 3	**Supernatant**	1der15_4LV	0.1205	42.41	0.0330	11.66
		2der15_5LV	0.0874	30.75	0.0189	6.66
		2der21_4LV	0.1082	38.09	0.0253	8.94
Model Type 4	**Re-dissolved Pellet**	1der15_7LV	0.0898	23.43	0.0519	15.06
		1der15_3LV	0.0673	17.57	0.0489	14.19
		2der15_7LV	0.0586	15.29	0.0424	12.30

Considering both the total and recombinant protein predictive errors for model types 2, 3 and 4, in Table 5.2 and Table 5.4, it was the untreated model type 2 that resulted in the greatest errors in both cases. This was expected as the untreated samples contained a large degree of particulates, including un-dissolved proteins, thus making the measurement of these samples and accuracy of the model predictions, more difficult and unreliable. However it should be noted that the cross-validation errors did not follow this trend.

Cross-validation errors of total and recombinant protein were comparable and results did not favour greater accuracy in the prediction of total protein over recombinant

protein, or vice versa. Untreated and supernatant prediction errors however, were greater for recombinant protein than those of total protein by a factor of 2, and re-dissolved pellet errors increased by a factor of 6 for recombinant protein. The concentration of recombinant protein present in each sample was an order of magnitude lower than the total protein concentration and therefore more difficult to accurately detect. Also, off-line results for recombinant protein, quantified via SDS-Page analysis, exhibited a higher variance (6.1-11.3%) than those for total protein (3.8%) quantified via Bradford assay. Increased variance present in the y-data of the training set most likely further impacted the accuracy of the recombinant protein models.

The trends exhibited by the predictive errors were typical given the constitution of each sample type and the concentration of total and recombinant proteins present in the samples. These trends were not exhibited in the RMSECV values however and this discrepancy highlights the issues encountered when choosing the optimum technique to evaluate all models. Correct selection involves striking a balance between independent validation samples and sufficiently large sample number so as not to excessively skew results.

5.3.3 Techniques for model evaluation

The lowest RMSECV and lowest RMSEP for all model types did not always result from the same model i.e. the optimum model for each model type varied depending on the method of evaluation. This result again highlighted the importance in accurately assessing the predictive capabilities of the model. Ideally a validation set used to test a model should have as many samples, if not more samples, than the training set used to

create it.[22] However this is seldom the case and an astute assessment of the given scenario is necessary. This study outlined two possible ways of evaluating the predictive ability of a model. As previously mentioned, the preferred technique is the application of an independent validation set to test the models accuracy and robustness. This is considered a more rigorous test of the models capabilities and results in the calculation of an RMSEP value. However, where calibration and validation sets are limited in sample number this is not necessarily the best method. In such sample sets, one sample can exert undue weight, either favourably or not, leading to a conclusion on the predictive capacity of a model that is not wholly accurate. The second method used to evaluate the models capabilities was examining the RMSECV. This test is considered to be less demanding on the model as each of the samples used to cross validate the model were used in the development of the model and so this technique is not entirely independent.

5.3.4 Detergent investigation

Figure 5.4 outlines the wavenumbers at which detergents A and B and also the untreated sample absorbed. Both detergents exhibited clear and distinct peaks. It was not possible to identify the spectrum of the untreated sample as spectra of samples containing untreated sample and detergent masked its presence. These initial observations indicated that models used for the prediction of protein from untreated culture broth were likely to be impacted by the addition of detergent to the culture.

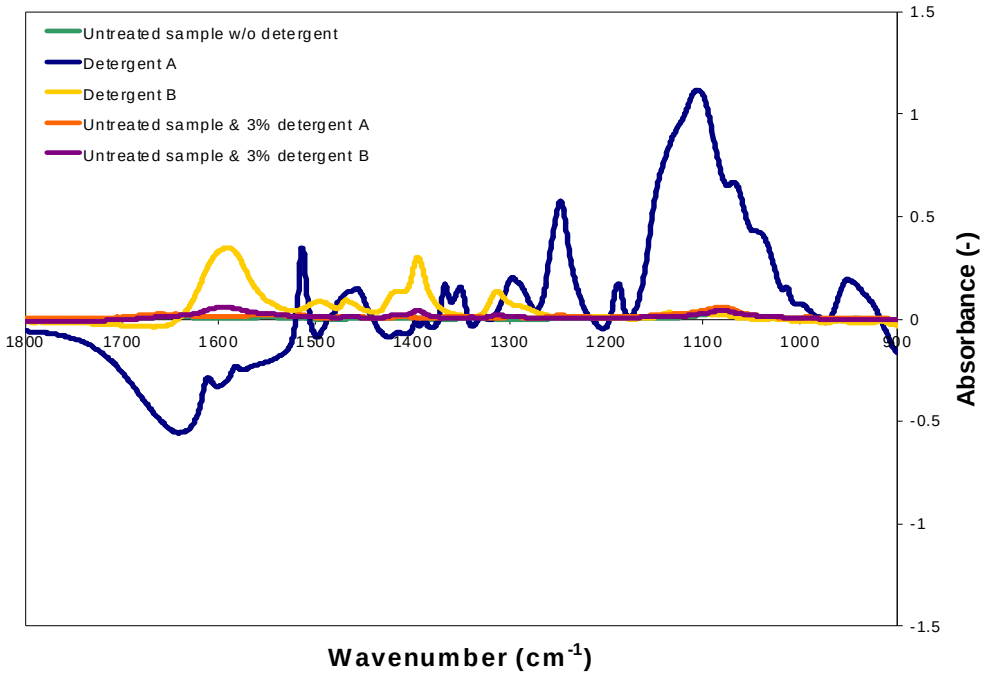


Fig. 5.4 Spectra of untreated samples, detergents A & B and untreated samples spiked with detergents A & B

The scores plot shown in figure 5.5 was generated by performing a PCA using mean centering and SNV as data pre-treatment methods. The PCA indicated that both detergents A and B in their pure form, were clearly distinguishable from untreated samples. Samples containing 3 differing concentrations of detergent A separated from samples which were not spiked with any detergent, however only the sample with the highest concentration of detergent B separated from non-spiked samples, with the lower 2 concentrations clustering with these samples.

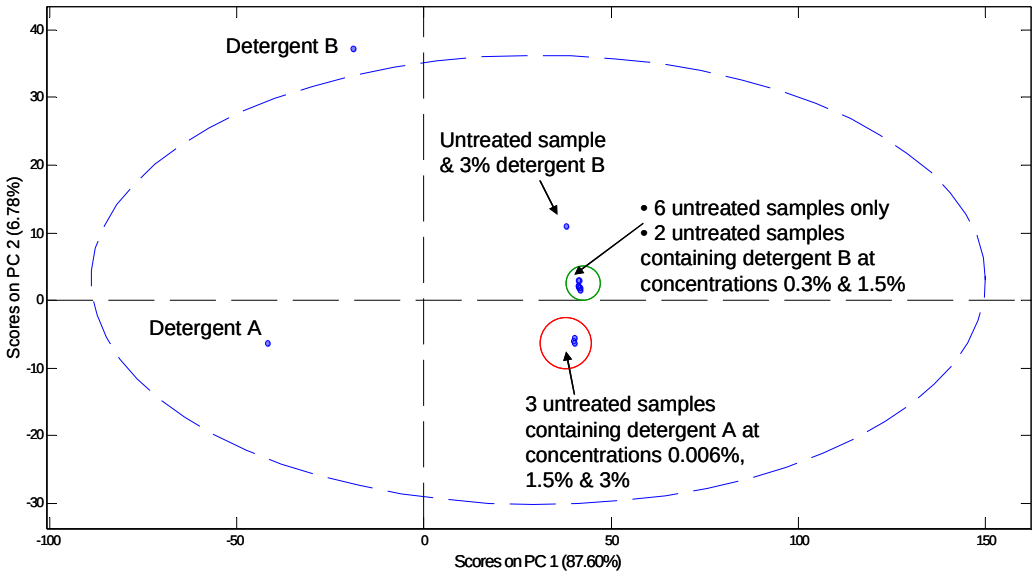


Fig. 5.5 Scores plot of untreated samples; detergents A & B; untreated samples spiked with detergents A & B

This preliminary investigation highlighted the potential impact the addition of detergent may have on the quantitative protein prediction models. Separate models should be developed for protein prediction depending on whether detergent is present or not. Where detergent is present, the range of its concentration should be incorporated into the model to include sample variation. The protein concentration range of each model should also be considered. Reference samples analysed showed that after the addition of detergent A, the concentration of the protein remained the same as before. Addition of detergent B to the untreated samples resulted in solubilisation of the protein and an increase of up to 4 times the original protein concentration in the supernatant samples. Therefore models constructed for use in the presence of detergents that facilitate solubilisation should include a higher upper limit concentration to account for increased quantities of soluble protein in the sample.

5.4 Conclusion

This exploratory study indicated that it was possible to qualitatively and quantitatively measure total and recombinant protein present in in-process recovery samples using an in-situ MIR immersion probe. This study highlighted the potential of MIR as an online monitoring technique for the production of a recombinant protein in a bioprocess.

A preliminary PCA indicated that despite the clear overlapping peaks at particular wavenumbers, it was possible to distinguish between the 6 proteins selected for this study. It has previously been reported that MIR can be used in the characterisation of protein structures however little work has been undertaken in this area using in-situ MIR immersion probes. The specificity of the instrument and its application in a particular environment have been shown to have a large bearing on whether or not a technique may be deemed acceptable. Initial analysis in this study indicated MIR has potential as a qualitative analysis tool; when applied to protein solutions the technique was capable of distinguishing between varying clusters of proteins.

Four model types were identified based on the samples available. PLS models were used to predict the total and recombinant protein present in 9 validation samples in the case of model type 1 and 3 validation samples for model types 2, 3 and 4. Model type 1 was the least accurate for both total and recombinant protein prediction with highest predictive errors in the supernatant samples at 53.89% and 68.60% respectively. The calibration set used for model type 1 was a combined set consisting of all sample types. Although it is recommended that variation be introduced into all models to account for atypical scenarios, the samples used to generate model type 1 differed

considerably and failed to enhance the robustness of the models. Results for specific models 2, 3 and 4 were consistently more accurate for both total and recombinant protein prediction. Given the limited number of available samples it was not possible to develop a thorough model capable of precise predictions while being easily able to identify outliers. Kramer's "rule of 3" was employed here,[22] which he states should only be used when completing "preliminary or exploratory work". However the methodology applied did provide a proof of concept, highlighting the potential of an MIR immersion probe in online protein monitoring.

Due to the limited calibration set number, this study also resulted in a comparison of model evaluation techniques. The cross-validation and prediction errors were calculated, with cross-validation errors indicating a higher degree of model inaccuracy. However, major trends outlined by one method were also reflected in the other; in the case of total and recombinant protein prediction, both cross-validation and application of an independent validation set indicated that models predicting total protein were more accurate, and when comparing the accuracy of model type 1 compared to that of the specific models, both techniques favoured the specific sample type models despite the lower sample number used in the calibration set.

The detergent investigation study provided insight into the necessary requirements for further model development, when detergent is present in the process. Results for both detergents tested indicated that at typically used concentrations, the protein predictions would be impacted by their presence and so conclude that further model development should account for detergent used in the process.

5.5 References

1. A. Barth, "Infrared spectroscopy of proteins," Biochim. Biophys. Acta-Bioenerg. **1767**, 1073 (2007).

2. C. A. Sellick, R. Hansen, R. M. Jarvis, A. R. Maqsood, G. M. Stephens, A. J. Dickson, and R. Goodacre, "Rapid monitoring of recombinant antibody production by mammalian cell cultures using fourier transform infrared spectroscopy and chemometrics," Biotechnol. Bioeng. **106**, 432 (2010).

3. P. I. Haris and F. Severcan, "FTIR spectroscopic characterization of protein structure in aqueous and non-aqueous media," J Molec Catal B **7**, 207 (1999).

4. A. Dominguez-Vidal, M. P. Saenz-Navajas, M. J. Ayora-Canada, and B. Lendl, "Detection of Albumin Unfolding Preceding Proteolysis Using Fourier Transform Infrared Spectroscopy and Chemometric Data Analysis," - Anal. Chem. **78**, 3257 (2006).

5. J. A. Hering, P. R. Innocent, and P. I. Haris, "Automatic amide I frequency selection for rapid quantification of protein secondary structure from Fourier transform infrared spectra of proteins," Proteomics **2**, 839 (2002).

6. S. Gross-Selbeck, G. Margreiter, C. Obinger, and K. Bayer, "Fast Quantification of Recombinant Protein Inclusion Bodies within Intact Cells by FT-IR Spectroscopy," Biotechnol. Prog. **23**, 762 (2007).

7. D. Landgrebe, C. Haake, T. Höpfner, S. Beutel, B. Hitzmann, T. Scheper, M. Rhiel, and K. Reardon, "On-line infrared spectroscopy for bioprocess monitoring," Appl. Microbiol. Biotechnol. **88**, 11 (2010).

8. M. Dabros, M. Amrhein, P. Gujral, and U. von Stockar, "On-line Recalibration of Spectral Measurements using Metabolite Injections and Dynamic Orthogonal Projection," Appl. Spectrosc. **61**, 507 (2007).

9. H. Kornmann, S. Valentinotti, I. Marison, and U. von Stockar, "Real-time update of calibration model for better monitoring of batch processes using spectroscopy," Biotechnol. Bioeng. **87**, 593 (2004).

10. S. A. Arnold, J. Crowley, N. Woods, L. M. Harvey, and B. McNeill, "In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation," Biotechnol. Bioeng. **84**, 13 (2003).

11. C. B. Lewis, R. J. McNichols, A. Gowda, and G. L. Coté, "Investigation of Near-Infrared Spectroscopy for Periodic Determination of Glucose in Cell Culture Media in Situ," Appl. Spectrosc. **54**, 1453 (2000).

12. E. Tamburini, G. Vaccari, S. Tosi, and A. Trilli, "Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fibre probe," Appl. Spectrosc. **57**, 132 (2003).

13. J. Henriques, S. Buziol, E. Stocker, A. Voogd, and J. Menezes, "Monitoring Mammalian Cell Cultivations for Monoclonal Antibody Production using Near-Infrared Spectroscopy," in *Optical Sensor Systems in Biotechnology*, Govind Rao, eds. (Springer Berlin / Heidelberg, 2010) p. 73.

14. P. Roychoudhury, R. O'Kennedy, B. McNeil, and L. M. Harvey, "Multiplexing fibre optic near infrared (NIR) spectroscopy as an emerging technology to monitor industrial bioprocesses," Anal. Chim. Acta **590**, 110 (2007).

15. N. Petersen, P. Ödman, A. E. C. Padrell, S. Stocks, A. E. Lantz, and K. V. Gernaey, "In situ near infrared spectroscopy for analyte-specific monitoring of glucose and ammonium in streptomyces coelicolor fermentations," Biotechnol. Prog. **26**, 263 (2010).

16. S. A. Arnold, R. Gaensakoo, L. M. Harvey, and B. McNeil, "Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch Escherichia coli process," Biotechnol. Bioeng. **80**, 405 (2002).

17. A. C. McGovern, R. Ernill, B. V. Kara, D. B. Kell, and R. Goodacre, "Rapid analysis of the expression of heterologous proteins in Escherichia coli using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: application to α2-interferon production," J. Biotechnol. **72**, 157 (1999).

18. C. L. Winder, E. Carr, R. Goodacre, and R. Seviour, "The rapid identification of Acinetobacter species using Fourier transform infrared spectroscopy," J. Appl. Microbiol. **96**, 328 (2004).

19. P. Rigler, W. Ulrich, P. Hoffmann, M. Mayer, and H. Vogel, "Reversible Immobilization of Peptides: Surface Modification and In Situ Detection by Attenuated Total Reflection FTIR Spectroscopy," ChemPhysChem **4**, 268 (2003).

20. D. J. Pollard, R. Buccino, N. C. Connors, T. F. Kirschner, R. C. Olewinski, K. Saini, and P. M. Salmon, "Real-time analyte monitoring of a fungal fermentation, at pilot scale, using in situ mid-infrared spectroscopy," Bioprocess. Biosyst. Eng. **24**, 13 (2001).

21. U.S. Department of Health and Human Services: Food and Drug Administration, *Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*, 2004).

22. R. Kramer, *Chemometric Techniques for Quantitative Analysis* (Marcel Dekker, Inc., New York, 1998) p. 203.

23. A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," Anal. Chem. **36**, 1627 (1964).

24. T. Naes, T. Isaksson, T. Fearn, and T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification* (NIR Publications, Chichester, UK, 2002) p. 344.

25. A. M. C. Davies and T. Fearn, "Back to basics: calibration statistics," Spectroscopy Europe **18**, 31 (2006).

26. H. Tao, W. Liu, B. N. Simmons, H. K. Harris, T. C. Cox, and M. A. Massiah, "Purifying natively folded proteins from inclusion bodies using sarkosyl, Triton X-100, and CHAPS," BioTechniques **48**, 61 (2010).

Chapter 6: Conclusions and Recommendations

This thesis investigated the use of Fourier Transform Infrared (FT-IR) spectroscopy, in the form of an in-situ MIR ATR probe, to determine the potential of this method as a PAT tool. The 3 studies presented provided a detailed account of the capabilities of the technique, outlining both the strengths and limitations of this application. The entire work evaluated the use of this technique at varying process steps. This was achieved in 3 studies, which examined its applicability and performance in upstream cell cultivation and also, downstream primary recovery. Initial studies focused on the use of in-situ MIR for monitoring of analytes and metabolites present in the bioprocess, while the latter study investigated the possibility of using the technique for protein quantification. Results indicated that MIR is deserving of its place in the PAT tool kit. Its capabilities lie both in qualitative and quantitative analysis.

6.1 Study 1: Potential of Mid-infrared spectroscopy for on-line monitoring of mammalian cell culture medium components

This initial study developed a methodology for evaluating the potential of MIR, for monitoring cell cultures medium components. The comparative probe investigation allowed for the identification of the optimum sampling accessory for this given application, which was then used in further studies. The fixed conduit immersion probe was found to be more reliable, with fibre-optic probe issues owing predominately to the ‘sample to crystal’ interface area. The design of the interface area resulted in frequent bubble entrapment, which consequently impacted results.

The spectral characterisation study completed identified the main areas of absorbance for each of the 8 components under investigation, and provided an indication of the areas of interest on the spectrum. The limit of detection (LOD) analysis highlighted at

an early stage, components that were not likely to be detected when the instrument was applied in-situ in a bioreactor. The concentrations of some components, e.g. glutamine and glutamate, in a cell culture medium, fell below the LOD. At this point therefore, the preliminary analysis facilitated the identification of components likely to be detected on-line and those that were not likely to be detected. Also, the LOD for a given component in a multicomponent mixture was found to be significantly higher that of the same component in a single component solution. It can be concluded that the multivariate LOD analysis presented in this study outlines a novel methodology for initial evaluation of the technique for a given purpose.

The calibration models developed in this study indicated the applicability of the technique in monitoring certain components such as glucose, ammonia and lactate, while also highlighting the limitations encountered when the technique is applied to media components that are at much lower concentration levels. It should be noted that the accuracy of the prediction ability of a model is very much dependent on the sample set size and the concentration level of each of the components present. This was further investigated in study 2.

Finally, the external influence investigation performed provided a novel method for pre-empting possible interferences the instrument, and hence the models, would encounter when applied on-line. This allows for understanding and mitigation of possible influences, but cannot completely eliminate these prior to on-line application.

6.2 Study 2: Application and optimisation of in-situ MIR calibration models for the prediction of glucose and lactate in mammalian cell cultures

This second study evaluated the application of MIR spectroscopy to the upstream processing of a CHO DP12 cell line, for the purpose of monitoring glucose and lactate concentrations during the cultivation. The fixed conduit FTIR immersion probe identified in study 1, as the optimum probe for bioprocess monitoring was selected and calibration models for glucose and lactate, developed using synthetic samples, were applied on-line and at-line. Both on-line and at-line trends reflected the expected trends, and those of the reference analysis, that being the depletion of glucose and the increase in lactate. However, as observed in study 1, when concentration values fell below the LOD, results became unreliable. This was evident in the glucose results of all 3 cultures, from day 4, when the glucose concentration fell below the LOD. Therefore application of this system may only be considered practicable where the concentration of these components remains above the LOD. Despite the observed trending, the prediction errors were considerably large for the in-situ application. At-line detection proved more accurate, and presented a viable alternative to on-line detection, however, at-line detection is not without its complications and further optimisation of the models was investigated.

The optimisation study resulted in the development of a number of hybrid models. These models were generated using training sets of spectroscopic data collected from synthetic samples, at-line samples and on-line samples. Where at-line spectroscopic data was used in the calibration set, models consistently performed better. It can therefore be concluded that although synthetic samples simulate the composition of the culture medium, the addition of spectra collected from “real” culture samples fortifies

the models as it accounts for minor unknowns that could impact the spectra, but are not accounted for in synthetic samples.

The addition of spectra collected in-situ to the calibration set provided interesting results. In the case of glucose, predictions did not improve. As previously stated, in-situ glucose results were not entirely reliable. When these spectra were used to increase the calibration set, they only served to enhance noise and inaccuracies, and did not increase the robustness of the models. This is an important result. It highlights that it is not the quantity of the spectra in a calibration set that will always strengthen the model, but also the quality of the spectra. Should spectra of an inferior quality be used in the creation of a model, the accuracy of the model may be compromised. In the case of lactate, all hybrid models were an improvement on the original model, however unlike glucose, lactate did not exhibit erratic predictions.

From this study it can be concluded that this technique can be applied in-situ for monitoring and potentially control purposes. Accuracy of calibration models can be improved by addition of spectra of actual culture samples; however it is imperative that care is taken when choosing spectra to ensure they are of good and reliable quality.

6.3 Study 3: The use of Mid-infrared spectroscopy as an on-line PAT tool in total and recombinant protein monitoring

This final study investigated the potential of MIR spectroscopy in detecting and quantifying total and recombinant protein during primary recovery. A spectral characterisation of 5 standard proteins and the recombinant protein of interest was

initially performed. This allowed the region of interest to be identified and clearly highlighted the amide I and amide II bands, located at ~1650 cm^{-1} and ~1550 cm^{-1} respectively, which are of great significance when MIR is used for protein characterisation. A PCA was performed on varying concentrations of the 6 proteins. This showed that the instrument in question was capable of distinguishing between differing proteins, and as a preliminary step, supported the development of PLS models suggesting potential for quantifying the recombinant protein of interest.

Using in-process samples, 3 sample types were available; untreated, supernatant and re-dissolved pellet samples. Based on these, four model types were identified; the first, using a training set consisting of spectra from all 3 sample types, while each of the remaining 3 models types used training sets corresponding to one of the 3 sample types.

Model type 1 was the least accurate for both total and recombinant protein prediction with highest predictive errors in the supernatant samples at 53.89% and 68.60% respectively. Specific model types 2, 3 and 4 were consistently more accurate for both total and recombinant protein prediction. It may therefore be concluded that varying sample type did not enhance model robustness, but actually served to destabilise the model.

This study also showed that although both total and recombinant protein could be detected and quantified, models developed for total protein prediction exhibited greater accuracy than those for recombinant protein.

Due to the limited number of samples available and hence the calibration set size, this study outlined the potential of MIR when applied on-line for protein detection. Also as a result of the sample set size, this work presented 2 possible methods of evaluating the results, RMSECV and RMSEP. Both methods produced varying results; however the trends observed were identifiable using either method.

Finally, the detergent investigation study provided insight into the necessary requirements for further model development, when detergent is present in the process. It can be concluded that at typically used detergent concentrations, protein predictions would be affected, hence further model development should be used by their presence and so conclude that further model development should account for detergent used in the process.

6.4 The role of chemometrics and pre-treatments

This work investigated the potential of MIR spectroscopy as a PAT tool in the monitoring of bioprocesses. However in order realise its full potential and harness its capabilities, the importance of chemometrics and spectroscopic pre-treatments must be understood. It is only through the use of chemometric techniques such as PCA and PLS can the information obtained via MIR be translated into a usable form.

All 3 studies indicate the importance of careful selection of pre-treatments and number of latent variables or principal components. For PLS models developed in all 3 studies, the pre-treatments of choice were mean-centering followed by first or second derivative application. These pre-treatments eliminate the effect of linear baseline

off-sets and smooth noisy spectra. For this instrument and this application, these were found to be most effective.

For the PCA completed in study 3, the optimum pre-treatments were mean-centering with standard normal variate (SNV). SNV is a normalisation type of pre-treatment. The qualitative information that distinguishes one sample from another is retained but information that separates two samples of identical composition, but different concentration, is removed. Therefore this is more applicable for this use.

6.5 Overall conclusions

The 3 studies presented in this thesis outline the possible applications of MIR when used as a PAT tool. The technique has definite potential in an on-line capacity, where continuous real-time monitoring is required. However limitations do exist, one major one of which is the low concentrations of the components it is monitoring. Identification of the LOD of a component is imperative to determine if the technique is a viable option. Integration into a control system should only be considered when the typical concentrations of components to be monitored are significantly greater than the LOD.

Model stability should also be evaluated. Careful selection of an appropriate training set is paramount, along with identification of a validation set that rigorously tests the model. As outlined in study 2, combining synthetic and real culture samples reinforces the model, provided the spectra used are of good quality.

In terms of a PAT tool, on-line MIR has much to offer. It is rapid, non-invasive and relatively easy to calibrate. However this technique, when applied to bioprocessing is not quite yet “industry ready”. Industry requires a simple, easy to use, robust instrument that is in effect, ready to “plug and play”. Although MIR has far reaching capabilities, they require further development if the technique is to be applied on a large scale across the biotechnology sector. Nevertheless, a little development may go a very long way in securing its place as one of the analytical tools of choice going forward.

Appendix I – MatLab Codes

%% Routine for online monitoring of media components

```
%% This programme uses command line functions from Eigenvector PLS
toolbox

% Siobhán Hennessy & Róisín Foley, LiB, School of Biotechnology, DCU


%%
go = 0;
count = 0; % Global counter which increases by 1 on each iteration
tic
%% Import component quatification model
load comp_model_data


while go==0 %Starts a loop to begin programme

start_time = toc; % Loop start time
count = count + 1; % Update counter
time(count) = toc / 3600; % Time stamp
%Insert the directory where spectral files are being imported
dirlist=dir('C:\Documents and Settings\LiB\My
Documents\MATLAB\online\*.spc');
dirsize = length(dirlist);

if dirsize>0 %Statement to put spectral files in chronological order

    for i = 1 : dirsize
        filedates(i) = dirlist(i).datenum;
    end
%Ensure the most recent spectrum is used
lastfiledate = max(filedates);
lastfileindex = findindx(filedates,lastfiledate);%Finds the index of
the array element closest to chosen value

filename = ['C:\Documents and Settings\LiB\My
Documents\MATLAB\online\', dirlist(lastfileindex).name];  % This
should be the same directory as above

s=spcreadr(filename); %Calls plstoolbox function spcreadr to import
.spc spectral file
wavenumber=s.axisscale{2}'; %assigned wavenumber data to the variable
wavenumber
spectrum=s.data; %assigned spectral data to the variable spectrum

%%sanity check 1 %%
% Test to see if spectra are of good quality - if absorbance value at
this
% point is greater than 0.5 it is likely that the MIR instrument has
run
% out of liquid nitrogen
if spectrum(909)>0.5
```

```
    %% start again
    pause(300)
    continue
else


% Defines regions of the spectrum for each component
gluc=[950 1450]; glne=[950 1700];amm=[950 1500]; phos=[900
1320];glte=[900 1760]; lact=[1000 1620];hep=[1000 1250]; bicarb=[960
1750];
wavenumrange = struct('regions', {gluc,glne, amm, phos, glte, lact,
hep, bicarb});
j=length(wavenumrange);

for  k= 1 : j %%SH for i=1:number of components
      T(k).ex = [];
       %wnrange(i).regions = [950 1500]; % Override mechanism
        %% SH sets up an empty matrix for field S.calib
for r = length(wavenumrange(k).regions)/2 : -1 : 1
       wavenumrange(k).downlimit(r) =
max(lamsel(wavenumber,[wavenumber(1)  wavenumrange(k).regions(r*2-
1)],0));
       wavenumrange(k).uplimit(r) =
min(lamsel(wavenumber,[wavenumrange(k).regions(r*2)
wavenumber(length(wavenumber))],0));

       T(k).ex = [T(k).ex
spectrum(:,wavenumrange(k).uplimit(r):wavenumrange(k).downlimit(r))];
    end
T(k).ex;
end

x_glucose=T(1).ex;
x_lactate=T(2).ex;

% x_ammonia=T(3).ex;
% x_phosphate=T(4).ex;
% x_glutamate=T(5).ex;
% x_glutamine=T(6).ex;
% x_HEPES=T(7).ex;
% x_Bicarb=T(8).ex;

%load gluc_model_data

options.display = 'off'; options.plots = 'none'; %Turns off default
plot options for pls function
% Defines variable conc which is the concentration predicted by the
model
conc1=pls(x_glucose, mc_2der15_4lvs,options);
glucose_conc(count)=conc1.pred{1,2};
conc2=pls(x_lactate, mc_2der15_4lvs,options);
lactate_conc(count)=conc2.pred{1,2};
```

```
%Series of additional criteria to be met to ensure no spurious
predictions
%are recorded i.e. If concentration is greater than initial media
%concentration, if concentration is a minus number or if the change
in
%concentration is too large to be physically possibly then the
program
%waits for the next spectrum and does not send the concentration
value to
%csv file where it could be used as part of a control system.

if glucose_conc(count)>10
elseif glucose_conc(count)<0
elseif glucose_conc(count)-glucose_conc(count-1)>abs(0.75)
   continue
else
timenow=clock;
csvwrite('concentration.csv',glucose_conc(count)); %Writes glucose
concentration to a csv file

if lactate_conc(count)>10
elseif lactate_conc(count)<0
elseif lactate_conc(count)-lactate_conc(count-1)>abs(0.75)
   continue
else
timenow=clock;
csvwrite('concentration.csv',lactate_conc(count)); %Writes lactate
concentration to a csv file


elapsed_time = toc - start_time;
    remaining_time = 300 - elapsed_time; %This figure can be changed
according to the frequency of spectral collection.
    %It is currently set for collection every 5 mins
    disp(['Waiting for next measurement... in '
num2str(remaining_time) ' seconds']) %Displays message on screen
    disp(' ')
    %Records the predicted value for each spectrum in a text file in
    %specified directory.
    dlmwrite('C:\Documents and Settings\LiB\My
Documents\MATLAB\MATLAB\spectra_log.txt',[timenow,
glucose_conc(count),lactate_conc(count)],'-
append','delimiter','\t','newline','pc');
    pause(remaining_time)
    if dirsize==1000 %Programme finished when 1000 spectra have been
collected. Alternatively press Ctrl+c
        break;
end
end
end
end
end
```

```
%% Protein Quantification

filename=uigetfile('*.xls'); % gets user to select the excel file
containing spectral data


importfile1(filename); % function to import data from sheet 1 of
excel file

X=deletecolumn(data,1)'; %deletes the 1st column of the data i.e. the
column of wavenumbers leaving only the spectral data.
wavenumber=selectcolumn(data,1); %selected the 1st column of data
i.e. the wavenumbers.

protein=[1200 1910];
wavenumrange = struct('regions', {protein});
clear protein data data2 colheaders colheaders2 textdata textdata2
filename;


plot(wavenumber, X); %plots the spectra for visual examination



protein.ex = []; %% SH sets up an empty matrix for field S.calib
    for r = length(wavenumrange.regions)/2 : -1 : 1
        wavenumrange.downlimit(r) =
max(lamsel(wavenumber,[wavenumber(1) wavenumrange.regions(r*2-
1)],0));
        wavenumrange.uplimit(r) =
min(lamsel(wavenumber,[wavenumrange.regions(r*2)
wavenumber(length(wavenumber))],0));

protein.ex =
[protein.ex X(:,wavenumrange.uplimit(r):wavenumrange.downlimit(r))];
    end
x_protein=protein.ex;

figure;
plot(wavenumrange.uplimit(r):wavenumrange.downlimit(r), x_protein)

clear protein X and r wavenumber wavenumrange;
```

```
%% Import File Function

function importfile1(fileToRead1)

%IMPORTFILE(FILETOREAD1)
%  Imports data from the specified file
%  FILETOREAD1:  file to read


% Import the file
sheetName='Sheet1';
[numbers, strings] = xlsread(fileToRead1, sheetName);
if ~isempty(numbers)
    newData1.data =  numbers;
end
if ~isempty(strings)
    newData1.textdata =  strings;
    newData1.colheaders =  strings;
end


% Create new variables in the base workspace from those fields.
vars = fieldnames(newData1);
for i = 1:length(vars)
    assignin('base', vars{i}, newData1.(vars{i}));
end
```

%% Delete Column Function

```
function [spec] = deletecolumn(irdata,index)
irdata(:,index)=[];
spec=irdata;
%deletecol - deletes columns of matrices
% usage: [X]= deletecol(X1,index)
%The deleted columns are indicated by the vector index (numbers of
booleans)
```

%% Select Column Function

```
function [spec] = selectcolumn(irdata,index)
spec=irdata(:,index);
%selectcol - creates a new data matrix with the selected columns
% the resulting file corresponds to the selected columns
% index is a vector of indices (integer) or of booleans
```

Appendix II – 7-Level & 4-Level design matrices

7-Level Partial Factorial Design (calibration matrix development)

Difference vector: {5 3 0 1 4 2}

Cyclic generator: -3→ 1→ -1→ 2 → 3 → -2→ -3

Repeater level: 3

7-level design	Sample no.	Glucose	Glutamine	Ammonia	Phosphate	Glutamate	Lactate	HEPES	Bicarbonate	Buffer
0	1	3	3	3	3	3	3	3	3	11
0	2	**3**	0	1	2	2	5	4	5	13
-3	3	0	1	2	2	5	4	5	3	13
-2	4	1	2	2	5	4	5	3	4	9
-1	5	2	2	5	4	5	3	4	0	10
-1	6	2	5	4	5	3	4	0	5	7
2	7	5	4	5	3	4	0	5	5	4
1	8	4	5	3	4	0	5	5	6	3
2	9	5	3	4	0	5	5	6	2	5
0	10	**3**	4	0	5	5	6	2	6	4
1	11	4	0	5	5	6	2	6	3	4
-3	12	0	5	5	6	2	6	3	2	6
2	13	5	5	6	2	6	3	2	4	2
2	14	5	6	2	6	3	2	4	6	1
3	15	6	2	6	3	2	4	6	6	0
-1	16	2	6	3	2	4	6	6	1	5
3	17	6	3	2	4	6	6	1	5	2
0	18	**3**	2	4	6	6	1	5	1	7
-1	19	2	4	6	6	1	5	1	3	7
1	20	4	6	6	1	5	1	3	5	4
3	21	6	6	1	5	1	3	5	2	6
3	22	6	1	5	1	3	5	2	1	11
-2	23	1	5	1	3	5	2	1	1	16
2	24	5	1	3	5	2	1	1	0	17
-2	25	1	3	5	2	1	1	0	6	16
0	26	**3**	5	2	1	1	0	6	0	17
2	27	5	2	1	1	0	6	0	3	17
-1	28	2	1	1	0	6	0	3	6	16
-2	29	1	1	0	6	0	3	6	5	13
-2	30	1	0	6	0	3	6	5	0	14
-3	31	0	6	0	3	6	5	0	0	15
3	32	6	0	3	6	5	0	0	4	11
-3	33	0	3	6	5	0	0	4	1	16
0	34	**3**	6	5	0	0	4	1	4	12
3	35	6	5	0	0	4	1	4	3	12
2	36	5	0	0	4	1	4	3	1	17
-3	37	0	0	4	1	4	3	1	6	16
-3	38	0	4	1	4	3	1	6	4	12
1	39	4	1	4	3	1	6	4	4	8
-2	40	1	4	3	1	6	4	4	2	10
1	41	4	3	1	6	4	4	2	0	11
0	42	**3**	1	6	4	4	2	0	2	13
-2	43	1	6	4	4	2	0	2	3	13
3	44	6	4	4	2	0	2	3	0	14
1	45	4	4	2	0	2	3	0	1	19
1	46	4	2	0	2	3	0	1	2	21
-1	47	2	0	2	3	0	1	2	2	23
-3	48	0	2	3	0	1	2	2	5	20
-1	49	2	3	0	1	2	2	5	4	16

Sample no.	Glucose	Glutamine	Ammonium Sulphate	Potassium Phosphate	Glutamate	Lactate	HEPES	Bicarbonate
1	3.00	0.30	1.89	0.73	0.35	1.72	2.23	0.63
2	3.00	0.00	0.63	0.49	0.23	2.87	2.97	1.04
3	0.00	0.10	1.26	0.49	0.59	2.30	3.71	0.63
4	1.00	0.20	1.26	1.21	0.47	2.87	2.23	0.83
5	2.00	0.20	3.14	0.97	0.59	1.72	2.97	0.00
6	2.00	0.50	2.51	1.21	0.35	2.30	0.00	1.04
7	5.00	0.40	3.14	0.73	0.47	0.00	3.71	1.04
8	4.00	0.50	1.89	0.97	0.00	2.87	3.71	1.25
9	5.00	0.30	2.51	0.00	0.59	2.87	4.46	0.42
10	3.00	0.40	0.00	1.21	0.59	3.45	1.49	1.25
11	4.00	0.00	3.14	1.21	0.70	1.15	4.46	0.63
12	0.00	0.50	3.14	1.46	0.23	3.45	2.23	0.42
13	5.00	0.50	3.77	0.49	0.70	1.72	1.49	0.83
14	5.00	0.60	1.26	1.46	0.35	1.15	2.97	1.25
15	6.00	0.20	3.77	0.73	0.23	2.30	4.46	1.25
16	2.00	0.60	1.89	0.49	0.47	3.45	4.46	0.21
17	6.00	0.30	1.26	0.97	0.70	3.45	0.74	1.04
18	3.00	0.20	2.51	1.46	0.70	0.57	3.71	0.21
19	2.00	0.40	3.77	1.46	0.12	2.87	0.74	0.63
20	4.00	0.60	3.77	0.24	0.59	0.57	2.23	1.04
21	6.00	0.60	0.63	1.21	0.12	1.72	3.71	0.42
22	6.00	0.10	3.14	0.24	0.35	2.87	1.49	0.21
23	1.00	0.50	0.63	0.73	0.59	1.15	0.74	0.21
24	5.00	0.10	1.89	1.21	0.23	0.57	0.74	0.00
25	1.00	0.30	3.14	0.49	0.12	0.57	0.00	1.25
26	3.00	0.50	1.26	0.24	0.12	0.00	4.46	0.00
27	5.00	0.20	0.63	0.24	0.00	3.45	0.00	0.63
28	2.00	0.10	0.63	0.00	0.70	0.00	2.23	1.25
29	1.00	0.10	0.00	1.46	0.00	1.72	4.46	1.04
30	1.00	0.00	3.77	0.00	0.35	3.45	3.71	0.00
31	0.00	0.60	0.00	0.73	0.70	2.87	0.00	0.00
32	6.00	0.00	1.89	1.46	0.59	0.00	0.00	0.83
33	0.00	0.30	3.77	1.21	0.00	0.00	2.97	0.21
34	3.00	0.60	3.14	0.00	0.00	2.30	0.74	0.83
35	6.00	0.50	0.00	0.00	0.47	0.57	2.97	0.63
36	5.00	0.00	0.00	0.97	0.12	2.30	2.23	0.21
37	0.00	0.00	2.51	0.24	0.47	1.72	0.74	1.25
38	0.00	0.40	0.63	0.97	0.35	0.57	4.46	0.83
39	4.00	0.10	2.51	0.73	0.12	3.45	2.97	0.83
40	1.00	0.40	1.89	0.24	0.70	2.30	2.97	0.42
41	4.00	0.30	0.63	1.46	0.47	2.30	1.49	0.00
42	3.00	0.10	3.77	0.97	0.47	1.15	0.00	0.42
43	1.00	0.60	2.51	0.97	0.23	0.00	1.49	0.63
44	6.00	0.40	2.51	0.49	0.00	1.15	2.23	0.00
45	4.00	0.40	1.26	0.00	0.23	1.72	0.00	0.21
46	4.00	0.20	0.00	0.49	0.35	0.00	0.74	0.42
47	2.00	0.00	1.26	0.73	0.00	0.57	1.49	0.42
48	0.00	0.20	1.89	0.00	0.12	1.15	1.49	1.04
49	2.00	0.30	0.00	0.24	0.23	1.15	3.71	0.83

Sample no.	Glucose	Glutamine	Ammonia	Phosphate	Glutamate	Lactate	HEPES	Bicarbonate
1	3.00	0.30	0.52	0.51	0.35	0.50	2.23	0.62
2	3.00	0.00	0.17	0.34	0.23	0.83	2.97	1.04
3	0.00	0.10	0.34	0.34	0.59	0.66	3.71	0.62
4	1.00	0.20	0.34	0.85	0.47	0.83	2.23	0.83
5	2.00	0.20	0.86	0.68	0.59	0.50	2.97	0.00
6	2.00	0.50	0.69	0.85	0.35	0.66	0.00	1.04
7	5.00	0.40	0.86	0.51	0.47	0.00	3.71	1.04
8	4.00	0.50	0.52	0.68	0.00	0.83	3.71	1.25
9	5.00	0.30	0.69	0.00	0.59	0.83	4.46	0.42
10	3.00	0.40	0.00	0.85	0.59	1.00	1.49	1.25
11	4.00	0.00	0.86	0.85	0.70	0.33	4.46	0.62
12	0.00	0.50	0.86	1.02	0.23	1.00	2.23	0.42
13	5.00	0.50	1.03	0.34	0.70	0.50	1.49	0.83
14	5.00	0.60	0.34	1.02	0.35	0.33	2.97	1.25
15	6.00	0.20	1.03	0.51	0.23	0.66	4.46	1.25
16	2.00	0.60	0.52	0.34	0.47	1.00	4.46	0.21
17	6.00	0.30	0.34	0.68	0.70	1.00	0.74	1.04
18	3.00	0.20	0.69	1.02	0.70	0.17	3.71	0.21
19	2.00	0.40	1.03	1.02	0.12	0.83	0.74	0.62
20	4.00	0.60	1.03	0.17	0.59	0.17	2.23	1.04
21	6.00	0.60	0.17	0.85	0.12	0.50	3.71	0.42
22	6.00	0.10	0.86	0.17	0.35	0.83	1.49	0.21
23	1.00	0.50	0.17	0.51	0.59	0.33	0.74	0.21
24	5.00	0.10	0.52	0.85	0.23	0.17	0.74	0.00
25	1.00	0.30	0.86	0.34	0.12	0.17	0.00	1.25
26	3.00	0.50	0.34	0.17	0.12	0.00	4.46	0.00
27	5.00	0.20	0.17	0.17	0.00	1.00	0.00	0.62
28	2.00	0.10	0.17	0.00	0.70	0.00	2.23	1.25
29	1.00	0.10	0.00	1.02	0.00	0.50	4.46	1.04
30	1.00	0.00	1.03	0.00	0.35	1.00	3.71	0.00
31	0.00	0.60	0.00	0.51	0.70	0.83	0.00	0.00
32	6.00	0.00	0.52	1.02	0.59	0.00	0.00	0.83
33	0.00	0.30	1.03	0.85	0.00	0.00	2.97	0.21
34	3.00	0.60	0.86	0.00	0.00	0.66	0.74	0.83
35	6.00	0.50	0.00	0.00	0.47	0.17	2.97	0.62
36	5.00	0.00	0.00	0.68	0.12	0.66	2.23	0.21
37	0.00	0.00	0.69	0.17	0.47	0.50	0.74	1.25
38	0.00	0.40	0.17	0.68	0.35	0.17	4.46	0.83
39	4.00	0.10	0.69	0.51	0.12	1.00	2.97	0.83
40	1.00	0.40	0.52	0.17	0.70	0.66	2.97	0.42
41	4.00	0.30	0.17	1.02	0.47	0.66	1.49	0.00
42	3.00	0.10	1.03	0.68	0.47	0.33	0.00	0.42
43	1.00	0.60	0.69	0.68	0.23	0.00	1.49	0.62
44	6.00	0.40	0.69	0.34	0.00	0.33	2.23	0.00
45	4.00	0.40	0.34	0.00	0.23	0.50	0.00	0.21
46	4.00	0.20	0.00	0.34	0.35	0.00	0.74	0.42
47	2.00	0.00	0.34	0.51	0.00	0.17	1.49	0.42
48	0.00	0.20	0.52	0.00	0.12	0.33	1.49	1.04
49	2.00	0.30	0.00	0.17	0.23	0.33	3.71	0.83

4-Level Partial Factorial Design (validation matrix development)

Difference vector: {0 2 1}

Cyclic generator: -1→ 1→ 2 → 1

Repeater level: 2

Units added to each sample

4-level design	Sample No.	Glucose	Glutamine	Ammonia	Phosphate	Glutamate	Lactate	HEPES	Bicarbonate	Buffer
-2	1	0	0	0	0	0	0	0	0	25
-2	2	0	1	1	4	1	0	3	3	12
-1	3	1	1	4	1	0	3	3	1	11
-1	4	1	4	1	0	3	3	1	3	9
2	5	4	1	0	3	3	1	3	0	10
-1	6	1	0	3	3	1	3	0	4	10
-2	7	0	3	3	1	3	0	4	4	7
1	8	3	3	1	3	0	4	4	3	4
1	9	3	1	3	0	4	4	3	4	3
-1	10	1	3	0	4	4	3	4	0	6
1	11	3	0	4	4	3	4	0	1	6
-2	12	0	4	4	3	4	0	1	1	8
2	13	4	4	3	4	0	1	1	4	4
2	14	4	3	4	0	1	1	4	1	7
1	15	3	4	0	1	1	4	1	0	11
2	16	4	0	1	1	4	1	0	3	11

Concentration of each compound added to each sample

Sample No.	Glucose	Glutamine	Ammonium Sulphate	Potassium Phosphate	Glutamate	Calcuim Lactate	HEPES	Bicarbonate
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.15	0.92	1.46	0.16	0.00	3.39	0.94
3	1.52	0.15	3.68	0.36	0.00	2.62	3.39	0.31
4	1.52	0.60	0.92	0.00	0.49	2.62	1.13	0.94
5	6.08	0.15	0.00	1.09	0.49	0.87	3.39	0.00
6	1.52	0.00	2.76	1.09	0.16	2.62	0.00	1.25
7	0.00	0.45	2.76	0.36	0.49	0.00	4.52	1.25
8	4.56	0.45	0.92	1.09	0.00	3.49	4.52	0.94
9	4.56	0.15	2.76	0.00	0.66	3.49	3.39	1.25
10	1.52	0.45	0.00	1.46	0.66	2.62	4.52	0.00
11	4.56	0.00	3.68	1.46	0.49	3.49	0.00	0.31
12	0.00	0.60	3.68	1.09	0.66	0.00	1.13	0.31
13	6.08	0.60	2.76	1.46	0.00	0.87	1.13	1.25
14	6.08	0.45	3.68	0.00	0.16	0.87	4.52	0.31
15	4.56	0.60	0.00	0.36	0.16	3.49	1.13	0.00
16	6.08	0.00	0.92	0.36	0.66	0.87	0.00	0.94

Concentration of each component added to each sample

Sample No.	Glucose	Glutamine	Ammonia	Phosphate	Glutamate	Lactate	HEPES	Bicarbonate
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.15	0.25	1.00	0.16	0.00	3.39	0.94
3	1.52	0.15	1.00	0.25	0.00	0.75	3.39	0.31
4	1.52	0.60	0.25	0.00	0.49	0.75	1.13	0.94
5	6.08	0.15	0.00	0.75	0.49	0.25	3.39	0.00
6	1.52	0.00	0.75	0.75	0.16	0.75	0.00	1.25
7	0.00	0.45	0.75	0.25	0.49	0.00	4.52	1.25
8	4.56	0.45	0.25	0.75	0.00	1.00	4.52	0.94
9	4.56	0.15	0.75	0.00	0.66	1.00	3.39	1.25
10	1.52	0.45	0.00	1.00	0.66	0.75	4.52	0.00
11	4.56	0.00	1.00	1.00	0.49	1.00	0.00	0.31
12	0.00	0.60	1.00	0.75	0.66	0.00	1.13	0.31
13	6.08	0.60	0.75	1.00	0.00	0.25	1.13	1.25
14	6.08	0.45	1.00	0.00	0.16	0.25	4.52	0.31
15	4.56	0.60	0.00	0.25	0.16	1.00	1.13	0.00
16	6.08	0.00	0.25	0.25	0.66	0.25	0.00	0.94