Combining Spectroscopy with Automation in Bioprocess Development

Towards a new tool for Industry 4.0

Joana Guerreiro Murtinheira Faustino

Strathclyde Institute of Pharmacy & Biomedical Sciences (SIPBS)

A thesis submitted to the University of Strathclyde in part fulfilment of the regulations for the degree of Doctor of Philosophy.

Glasgow, Scotland

2021

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Contents

List of Figures

List of Figures

List of Tables

Abbreviations

ATR - Attenuated Total Reflectance

CQA - Critical Quality Attribute

CS - Culture Station

DO - Dissolved Oxygen

DoE - Design of Experiments

EDTA - Ethylenediaminetetraacetic Acid Disodium Salt Dihydrate

FT - Fourier Transformation

HCDC - High Cell Density Culture

HMI - Human-Machine Interface

HPLC - High Performance Liquid Chromatography

ICH - International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

MCM - Multiplex Calibration Model

MIR - Mid-Infrared

MSC - Multiplicative Scatter Correction

MVDA - Multivariate Data Analysis

NIR - Near Infrared

NIR - Near Infrared Spectroscopy

OSC - Ordinary Least Squares

PC - Principal Component

PCA - Principal Component Analysis

PDS - Piecewise Direct Standardisation

PLS - Partial Least Squares

QbD - Quality by Design

RE - Relative Error

RMSE - Root Mean Square Error

RMSEcv - Root Mean Square Error of Cross Validation

RMSEE - Root Mean Square Error of Estimation

RMSEP - Root Mean Square Error of Prediction

SNV - Standard Normal Variate

YSI - Yellow Springs Instruments

Acknowledgements

I would first like to thank my supervisors, Professor Brian McNeil and Dr. Adrian Stacey, for their knowledge, encouragement, and guidance over the years. I would also like to thank Dr. Linda Harvey and Dr. Alison Nordon for their support and Professor José Menezes for introducing me to the PAT world.

I am grateful to Sartorius Stedim and in particular to my industrial supervisor, Dr. Adrian Stacey, for all support throughout my internship. It was a invaluable experience. I am also thankful to Dr. Mariana Fazenda for her friendship and support in Cambridge.

I also thank the EPSRC Centre and the University of Strathclyde for awarding me with this Industrial CASE studentship.

On a more personal note, I had an amazing time in Glasgow and for that, I thank the Fermentation Group and my other friends in Glasgow, in particular Cinta Diez, for always being there. To my friends and family back home but particularly to *mãe*, *pai* and especially my *querida avó*: I'm sorry for all the *saudades* and moments I have missed. To my new Dutch family, thank you for making me part of the *clan* and keep cheering me on.

Lastly, the most wholeheartedly special thanks to my loving Loe, for the infinite support, discussions, walks, and patience, over this overly extended writing period. I would definitely not have done it without you and to be writing these acknowledgements still feels unreal but, thank you! I am truly looking forward to the next chapter together.

Abstract

The biopharmaceutical industry has long relied on the work of microbiologists to perform labour intensive experiments in bioprocess development. Real time information on the specific dynamics of each experiment would represent a major breakthrough in understanding bioprocesses. This would lead to higher productivity, shortened development cycles, ultimately resulting in cheaper drugs reaching the clinic faster with potentially improved safety and efficacy.

The purpose of this thesis is to expedite the biopharmaceutical development by combining automation and near infrared spectroscopy in the early stages of bioprocess development. It develops the first automated spectroscopic system for automated small-volume bioreactors. After describing the hardware, software, optimization and testing of this new tool, various advantages and challenges that come with such system are discussed.

The advantages of automation when put to use into spectroscopy are quickly apparent. The automated approach can collect a - highly replicable - calibration dataset six times faster than the traditional approach, eliminating human error. Multivariate data analysis, including PCA and PLS, proved that valid real time measurements of analytes of interest in the fermentation media can be obtained.

The use of the developed models for extrapolation showed limited success. Better results were obtained when an endogenous calibration dataset was prepared through means of automation. This exercise allowed the identification of critical success areas for developing a spectroscopy-based measuring system for small automated bioreactors.

In the end, a roadmap for future implementation of an automated NIR system is provided that summarises the key lessons drawn from this thesis.

Chapter 0. Abstract

Chapter 1

Introduction

The global biopharmaceutical market continues to increase, and its value might reach more than $500 billion by the year 2025 (Deloitte, 2018). Biopharmaceuticals are revolutionary therapies, but they are also the most complex to develop and manufacture. They are obtained through bioprocesses, i.e. the culture of biological systems such as bacteria, yeasts or mammalian cell lines. Due to the highly sensitive nature of the cells, their physiology varies with culture conditions in the bioreactor.

The industry has long depended on the work of microbiologists to perform exhaustive experiments involved in bioprocess development. During its development, and to reach a feasible process, microbiologists must run a large number of experiments. Due to their dynamic nature, bioprocesses need to be closely monitored. However, to date, only process variables, such as pH and oxygen levels, are routinely measured online, i.e. in real time. The availability of real-time information could represent a significant step in the bioprocessing industry, as a deeper knowledge of the bioprocess would result in cheaper drugs reaching the clinic faster.

Industry is in continuous demand for new and improved technologies to reduce costs and increase efficiency. Process Analytical technology (PAT), firstly conceptualised by the FDA in 2004 in a Guidance for Industry, was one of these new tools. In this document, the FDA considers PAT as "a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality" (FDA, 2004, p4). It also encouraged the pharmaceutical industry to pursue the implementation of this type of tools as they should allow for monitoring and control of the process, while maintaining high product quality during manufacturing operations. The ongoing trend of "more automation, monitoring and process control" (BioPlan, 2017) is more relevant than ever and provides fertile ground for PAT to develop further.

The growing interest in automated single-use systems in Biopharma is one example of this. Likewise, continuous manufacturing might enter the world of bioprocessing and Process Analytical Technology (PAT) solutions might see further development and investment in its implementation because they are essential in these developing

technologies.

In an age of ever-increasing automation, "Industry 4.0" is now on the horizon, manual errors and variability can be greatly reduced in Pharmaceutical Industry. Tests in laboratory environments have been responsible for an increase in productivity between 30 to 40 percent, while a full range of improvements could lead to reduction of overall quality-costs in more than 50 percent (Makarova et al., 2019). In combination with the digitalisation required in Industry 4.0, prevention of major compliance issues can worth millions in cost savings while reducing the quality control lead times by 60 to 70 percent to eventually, real time releases. Tools such as the Ambr® systems from Sartorius Stedim have been in increasing demand in the biotechnology industry. These automated single-use bioreactor systems deal with volumes in a highly accurate, reproducible, and sterile manner (Xu et al., 2017). Design of Experiments (DoE) methodologies, as advised on ICH guideline Q11 FDA (2012), are also efficiently followed.

The routine operation of spectroscopy (such as NIR) could also benefit from this wave of automation. The concretisation of spectroscopy in automated single-use miniaturized bioreactors and PAT would represent a major breakthrough in bioprocess development. The information generated by PAT will improve safety and efficiency, while the use of automation will drastically reduce the 'time-to-market', ultimately resulting in cheaper drugs. In addition, the fusion of PAT, automation and bioreactors (e.g., Ambr®), can allow the end-users of Ambr® systems (R&D scientists) to obtain real-time information about the metabolites in the medium, without knowledge on chemometrics. This would critically depend on the development of new data analysis methodologies.

The purpose of this thesis is to expedite the biopharmaceutical development using spectroscopic methods in early stages of development. The abilities of spectroscopy such as Near Infrared (NIR) on monitoring valuable components of a bioprocess have long been proven. However, there are still hurdles for its adoption and particularly its implementation into early stages of bioprocess development.

Through the use of automation, novel methodologies for the routine use of NIR in such bioreactors can be investigated and subsequently employed. The challenges of

such implementation are threefold: calibration, integration and cost.

Calibration is the first step involved in the application of NIR measurements. For the specific case of bioprocesses, known data is used to determine the relationship between the observed spectral variation and the corresponding levels of relevant components, i.e. analytes, in the media inside the bioreactor. Once the bioprocess is fully developed, it is then largely consistent and reproducible and thus calibration methodology is straightforward, and it varies around a central point. On the contrary, at early stages of bioprocess development many experiments need to be conducted to define the best culture conditions or to select the best clone. Hence a response to this challenge is to develop approaches that do not rely on processes already established, but instead solutions that function across a range of conditions without extensive calibration methods.

For the integration of NIR in the automated bioreactors the challenges lie in the approach used for the type of configuration with the constraints of small-volume vessels and space available for instrumentation.

Lastly, the cost challenge is one of economies of scale. Currently, only certain pharmaceutical companies with sufficient output are able to employ spectroscopy profitably. To change this reality, less expensive instruments need to be produced to also be used in small volume. To apply this type of instruments, the challenge lies in analysing degraded data through improved chemometrics.

Commercial deployment of NIRS based solutions depends upon the capacity to be confident in the measurements derived from NIRS, and the applicability of NIRS across a broad analyte range. A much better understanding of how different sources of variability impact upon the spectra is required. This in turn will affect the ability of the constructed models to better predict analyte levels.

This thesis aims to propose the appropriate strategy for integration of NIR measurements in the early stages of bioprocess development into small, automated bioreactor systems, in a cost effective and reproducible manner. This would represent the first tool that unlocks the full potential of spectroscopy to the use of biopharma.

In doing so, this thesis identifies and aims to tackle various challenges, e.g. related

to dealing with various sources of variability, constructing valid models for analyte determination, choice of the appropriate instrument, the hardware development, methodology and consequent analysis. These challenges are apparent in the build-up of this thesis, which is as follows:

- *Chapter 2 Literature Review* discusses the related literature on the topics of bioprocessing, spectroscopic methods and chemometrics. It provides further background on how current technological developments in the pharmaceutical industry provide fertile ground for revolutionary methodologies such as the implementation of NIR in the early stages of bioprocess development.
- *Chapter 3 Characterization of the Fermentation Matrix by NIR Spectroscopy* gives a background on sources of variability in NIR measurements from fermentation matrices and uses data collected through the use of two different instruments.
- *Chapter 4 Symbiosis of Automation and NIR Spectroscopy* discusses sources of variability when the sample is delivered to a NIR instrument in an automated manner. The predictive capability of the developed prototype is tested through an extensive exploratory data analysis. It concludes on the synergistic benefits of combining an automated system with NIR measurements.
- *Chapter 5 Realisation of NIR in Automated Bioreactor Systems* tests the models developed in Chapter 4 in the context of samples collected from the Ambr automated bioreactors. This chapter culminates in a roadmap for the final implementation of NIR into these bioreactors.
- *Chapter 6 Conclusions and Latest Developments* reviews the lessons of the previous chapters and finishes with a look at the latest developments, as the prototype developed in Chapter 4 became a reality and it is now an available commercial tool.

In the end, this thesis aims to contribute to the further implementation of PAT in the biopharmaceutical industry.

Chapter 2

Literature Review

2.1 Introduction

Biopharmaceuticals are the class of drugs produced by biological organisms for treating or preventing diseases or for diagnostic purposes. In 2012, the biopharmaceuticals world market already represented more than $145 billion in sales (Coggins et al., 2012). By 2024, biopharmaceuticals should represent half of the pharmaceutical market amounting to US$178 billion (Reh, 2020).

The majority of the approved biopharmaceuticals for therapeutic applications are proteins that have been produced by different cell factories (expression systems) by means of recombinant DNA technology (Berlec and Strukelj, 2013; Walsh, 2010). In most cases, they are used to compensate a deficiency or lack of body proteins relevant for the normal functioning of the organism. There are different biopharmaceutical types: blood factors, thrombolytics and anticoagulants, hormones, enzymes, growth factors, interferons and interleukins, vaccines and monoclonal antibodies (Walsh, 2010). Due to these recombinant therapeutic proteins, innovative and effective therapies for numerous previously refractory illnesses are today being provided and treatments are currently available for diseases ranging from cancers to infertility (Jayapal et al., 2007).

Despite being the most potent type of drugs ever developed, biopharmaceuticals also constitute the most complex and expensive ones to manufacture. The cost of the product development stage alone can be estimated as approximately 30 to 35% of the total cost of developing a new drug (Suresh and Basu, 2008). Any time reduction in the development stage of biopharmaceuticals will therefore have a major impact on the drug economics (Walsh, 2014). For this to happen, the understanding of the protein expression system's metabolism throughout the upstream phase needs to be enhanced. Even though fermentation technology has made advances in recent years, there is considerable room to improve our ability to understand and control in real time the metabolism of cell culture systems (Henriques et al., 2009). This could have a major impact in the bioprocessing industry, improving productivity and shortening development cycles, leading to cheaper drugs reaching the clinic faster with potentially improved safety and efficacy (Fazenda et al., 2013).

Successful operation of a bioprocess depends on measuring culture conditions and the response of the organism to these conditions. A wide range of chemical parameters modulate cell physiology but few are routinely measured on-line such as temperature, pressure, dissolved oxygen, pH, stirring speed, and gas and liquid flow rates, which are commonly referred to as process variables (Schügerl, 2001). Real-time monitoring sensors for bioreactors are ideally fast response, sensitive, specific, non-destructive and robust (Classen et al., 2017). These should generate multi-analyte data without analyte consumption, not require sampling, not interfere with culture metabolism and be resistant (Vojinović et al., 2006).

The interaction between molecular bonds and electromagnetic waves can be detected and give useful information. Specifically for bioprocess monitoring and control, these spectroscopic sensors are typically in the spectral ranges from UV to MIR, which also includes Raman and fluorescence spectroscopy. Each spectral technique excites different molecule types. UV/Vis, NIR, and MIR spectroscopy primarily provide measurement of the light absorption of molecules, as well as their light scattering. Fluorescence spectroscopy uses the light emitted from excited fluorophores, and Raman spectroscopy uses rare inelastic scattering effects. Using the whole spectrum makes nearly all important physical, biological, and chemical variables of a bioprocess accessible by spectroscopy. However, despite the improved performance and reliability of these kind of sensors, most are still highly priced and come with high maintenance requirements (Simon et al., 2015). For these reasons, their applicability to monitoring bioprocesses faces many challenges.

The first step of bioprocess development is to determine the most adequate expression system to be used. The most commonly used expression systems are bacteria, yeasts, and mammalian cells (Berlec and Strukelj, 2013). However, insect cells (such as *Spodoptera* and *Trichoplusia*), algae and fungal cells are also available. The selection of the most appropriate system is based on the best trade-off between the manipulation (cultivation) of the organism and the obtained quality and quantity of product. Thus, extended knowledge about the expression systems physiology is essential for its adequate cultivation, to which follows further purification of the pharmaceutical product.

The sustained growth of the biopharmaceutical market stimulates the investment in developing tools for a deeper understanding of the bioprocess (Classen et al., 2017). PAT (Process Analytical Technology) is a concept introduced by the Food and Drug Administration (FDA) to increase the efficiency and control of the production of pharmaceuticals (FDA 2004). Spectroscopic methods can be ideal for process monitoring as they are non-invasive, non-destructive, can be implemented in different configurations and give (potentially real-time) information about several variables. The complexity of biological molecules and processes create big challenges in the application of PAT to biopharmaceutical manufacturing. However, the number and diversity of studies on the topic should guarantee its successful application (Hong et al., 2018).

This chapter is structured as follows: expression systems are reviewed, spectroscopic methods are discussed, chemometric methods are presented and sources of variability are approached.

2.2 Expression Systems

2.2.1 *Escherichia coli*

Escherichia coli (*E. coli*) strains are one of the most commonly used industrial bacteria for fermentation processes and they can be grown to high cell densities (up to 190 g/L dry cell mass) through aerobic cultivations (Nakano et al., 1997).

E. coli is the least demanding expression system since it is cheap and simple, and possibly the most studied organism, which makes its physiology very well known (Berlec and Strukelj, 2013). Moreover, process fluids for *E. coli* are of low viscosity and behave in a Newtonian fashion, thus being optically simpler than, for example, filamentous cultures (Scarff et al. 2006). The drawbacks of this system are related to not being an eukaryotic one and thus not executing the appropriate protein folding and other post-translational modifications (Berlec and Strukelj, 2013).

E. coli has been employed to produce industrial enzymes (e.g. rennin, amylases, proteases and cellulases) but also therapeutic proteins (e.g. filgrastim, insulin, growth hormones, and interferons) (Eiteman and Altman, 2006).

2.2.1.1 Physiology

There are different environments where *E. coli* can be grown. In an anaerobic environment *E. coli* can still extract energy from substrates through fermentation (Iuchi and Weiner, 1996).

In *E. coli*, recombinant proteins are often expressed in the bacterial cytoplasm but they can be directed to other compartments such as the periplasm or, more rarely, secreted to the growth medium (Berlec and Strukelj, 2013). Different approaches for recombinant protein expression are necessary for each location.

High-cell density cultures are able to provide high concentrations of producing cells and thus high production of the protein of interest. Furthermore, these systems are cost-effective and environmentally friendly (Berlec and Strukelj, 2013). In order to achieve high concentrations levels, the growth medium composition has to be controlled to prevent unwanted metabolic effects, which requires real time process control (Berlec and Strukelj, 2013; Scarff et al., 2006). However, these high cell density systems can lead to filamentation, having an impact in the viscosity of the liquid phase, which implies vigorously aerated and agitated systems at industrial scale and, therefore, constituting a highly challenging area for spectroscopic applications (Berlec and Strukelj, 2013).

2.2.1.2 Cultivation Process

E. coli can be grown in different modes: batch, fed-batch and continuous-fermentation (Berlec and Strukelj, 2013). The nutrient feeding strategy during a fed phase is crucial for high cell density cultures. Different feeding strategies include a constant feed rate, an increased feed or an exponential feeding profile (Choi et al., 2006).

2.2.1.3 Process Monitoring

The primary analytes that should be monitored in order to obtain an understanding of metabolic state of *E. coli* fermentations are the carbon source (glucose, glycerol) and potentially toxic by-products (such as ammonium, and acetic acid concentrations). In a batch medium, the nitrogen source is usually an ammonium salt and/or as ammonium

hydroxide when in a feed. It is possible to obtain faster growth rates on ammonium than on most amino acids with the exception of glutamine (Macaloney et al. 1997). However, it has been shown that ammonium ion can inhibit *E. coli* growth when present in concentrations larger than 170 mM (Thompson et al., 1985).

Acetic acid is a common by-product in *E. coli* processes that reduces the biomass yielded from carbon source, as well as the growth rate. Ultimately this leads to lower product synthesis (Pan et al., 1987; Bech Jensen and Carlsen, 1990).

2.2.2 Yeasts

Yeasts are eukaryotic cells that are widely used for the expression of several proteins in vaccine and pharmaceutical production. The mechanism of protein expression in these microorganisms is close to the ones in mammalian cells. Compared with bacteria, yeast cells have significant advantages including growth speed, post-translational modification, secretory expression, and easy genetic manipulation. Furthermore, linearized foreign DNA can be inserted in a chromosome in high efficiency via cross recombination phenomena to generate stable cell lines (Daly & Hearn, 2005). Yeasts are able to produce high yields of proteins at low cost. They are able to produce proteins larger than 50 kDa, remove signal sequences and perform glycosylation (Sekhon, 2010). The most used yeasts are *Saccharomyces cerevisiae* and *Pichia pastoris*.

2.2.2.1 *Saccharomyces Cerevisiae*

S.cerevisiae is a known expression system. It can be grown in low cost media and involves simple genetic manipulation. It also brings the advantages of being an eukaryote: *S.cerevisiae* is able to do proteolytic processing, folding, disulphide bond formation, and post translational modifications without carboxylation. Since it has been a part of the human diet for centuries, this yeast is accepted as "generally recognized as safe"(GRAS), which facilitates regulation procedures. *Saccharomyces cerevisiae* is used in the manufacture of hepatitis B and human papillomavirus vaccines, both of which produce a protective immune response against wild-type viruses (Karbalaei et al., 2020).

Regarding its disadvantages, *S.cerevisiae* is unable to reach high cell densities and

exhibits limited secretion and irregular glycosylation (Berlec and Strukelj, 2013).

Instead of concentrating in a single protein or group of proteins, future studies on this yeast will rely on an 'omics approach (genome-, proteome-, transcriptome-, and/or metabolome- wide information) (Berlec and Strukelj, 2013).). High-throughput screening is already used to analyse cDNA over-expression libraries to identify, which gene may contribute to secretion of the biopharmaceuticals (Wentz and Shusta, 2007).

2.2.2.2 *Pichia pastoris*

Pichia pastoris (*P. pastoris*; syn.*Komagataella phaffii*) is an effective and versatile expression system used in the biopharmaceutical industry. More than 500 heterologous proteins have been expressed in *P.pastoris* which is licensed to more than 160 companies (Julien, 2006). It can synthesize proteins at high levels and achieve very high cell densities (>130 g/L), (Cereghino and Cregg, 2000; Berlec and Strukelj, 2013) and secrete the recombinant proteins, making the purification easier. *P.pastoris* is also able to perform post-translational modifications such as glycosylation, formation of disulphide bonds and proteolytic processing (Karbalaei et al., 2020).

This species exhibits many of the advantages normally associated with *Escherichia coli* expression systems (Morton and Potter, 2000) while overcoming many of the deficiencies associated with *Saccharomyces cerevisiae* systems. *P.pastoris* shows shorter and less immunogenic glycans, higher density cell growth and higher secreted protein yields than *S. cerevisiae* (Tran et al., 2017). Overall the main advantage of *P.pastoris* is the availability of engineered strains capable to perform humanized glycosylation, its better protein secretion efficiency, the high biomass yield and the existence of promoters such as pAOX1, which enables a tightly controlled methanol–inducible transgene expression (Niu et al., 2013).

Prior to humanized strains being developed, biopharmaceuticals produced by *P.pastoris* were already in the market. They were either not glycosylated (human serum albumin) or the glycosylation needed were just to attribute the proper folding (e.g. several vaccines). The availability of *P.pastoris* strains with human-type N-glycosylation constituted a major breakthrough (De Schutter et al., 2009). Up to that date, only

mammalian hosts could be used to produce proteins with humanlike N-glycosylation (Hamilton et al., 2006). However, as protein expression systems, yeasts present many advantages over mammalian cells, such as higher recombinant protein titers, shorter fermentation times, the ability to grow in chemically defined media instead of a complex one and, finally, they are also easier to manipulate (Hamilton et al., 2006; Macauley-Patrick et al., 2005). With these engineered strains, it is possible to obtain the correct glycosylation with the advantages of using yeast as expression system.

Some obstacles arise in *P.pastoris* application to large-scale such as the cost of downstream purification, the fact that certain recombinant proteins suffer proteolytic degradation, and the storage and handling of large amounts of methanol (Potvin et al., 2012).

2.2.2.2.1 Physiology

Methanol Metabolism

P.pastoris is a methylotrophic yeast, therefore it can be grown with methanol as the unique source of carbon and energy, as shown first by (Ogata et al., 1969). The first step of methanol utilisation pathway (Figure 2.1) of all methylotrophic microorganisms is catalysed by the enzyme alcohol oxidase (AOX) (Schenk et al., 2007). AOX has a poor affinity for oxygen and so must be expressed at high levels by the cell: this enzyme accounts for up to 35% of the total cell proteins and it is undetectable in cells that grow on glucose, ethanol or glycerol (Sreekrishna et al., 1997). This specificity and high levels have been exploited to express foreign proteins in *P.pastoris*, through isolation of the gene and promoter for AOX (Schenk et al., 2007).

As shown in Figure 2.1, the first step in the methanol pathway (MUT) that takes place in the peroxisome, is the oxidation of methanol to formaldehyde and hydrogen peroxide. The enzymes AOX and catalase, which degrades hydrogen peroxide to oxygen and water, are located within the peroxisome. A portion of the generated formaldehyde leaves the peroxisome and further oxidized to formic acid and carbon dioxide by two cytoplasmic dehydrogenases. These reactions are the source of energy for cells that grow on methanol.

The formaldehyde remaining in the peroxisome is assimilated to form cellular constituents by a cyclic pathway that starts with a reaction catalysed by the enzyme dihydroxyacetone synthase (DHAS).

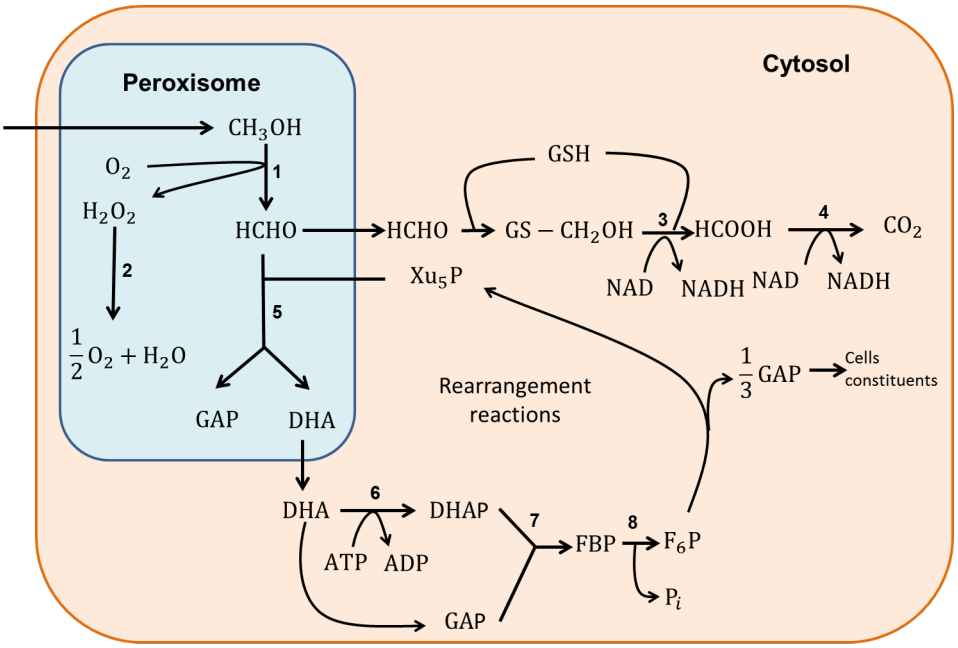

Figure 2.1: The methanol pathway (MUT) in *Pichia pastoris*. 1-alcohol oxidase (AOX); 2 – catalase; 3 – formaldehyde dehydrogenase; 4 – formate dehydrogenase; 5 – dihydroxyacetone synthase (DHAS); 6 – dihydroxyacetone kinase; 7 – fructose 1,6-biphosphate aldolase; 8 – fructose 1,6-bisphosphatase. From Cereghino and Cregg (2000).

Other Promoters

The glyceraldehyde 3-phosphate dehydrogenase (GAP) promoter (pGAP) can also be used to express heterologous proteins (Cos et al., 2006). However, this promoter is constitutively expressed which means that it constitutes an option only for proteins that are not toxic for the cell (Cereghino and Cregg, 2000).

The activity levels of the GAP promoter are higher with glucose than glycerol (one-third less) or methanol (two-thirds less) (Cereghino and Cregg, 2000). With this system there is no need to use methanol for induction, or to shift cultures from one carbon source to another; biomass and protein synthesis just occur simultaneously and they are directly correlated to pGAP-regulated gene dosage (Potvin et al., 2012).

2.2.2.2.2 Cultivation Processes

Usually, *P.pastoris* is grown in fed-batch mode instead of continuous cultivations in order to achieve high cell densities and to be easier to control. When pAOX1-regulated strains are used the cultivation is commonly divided into three phases: a glycerol batch and fed-batch stages, and a methanol induction phase.

The batch phase permits the consumption of the carbon source to reach the right level of biomass as fast as possible before the expression of heterologous protein. Usually the growth substrate is glycerol since growth rates on glycerol are higher than on methanol (Potvin et al., 2012). The duration of this stage is approximately 24 hours and it starts with a concentration about 40 g/L of glycerol (Cos et al., 2006).

The second stage is a glycerol fed-batch and starts when the initial glycerol of the batch phase is consumed. This is detected by a sharp increase in the dissolved oxygen. The goal of this phase is to increase biomass levels and to de-repress the AOX1 promoter by not having an excess of glycerol. There are different approaches to the feeding mode: at a constant rate of glycerol or an exponential one (Cos et al., 2006). The length of this phase depends on the desired biomass concentration prior to methanol induction (Potvin et al. 2012).

Finally, methanol is fed to induce expression of the recombinant protein. Methanol then functions both as carbon source and as pAOX1 inducer (Potvin et al., 2012).

Scale-up constraints

High amounts of heat are generated when methanol is metabolized, due to its heat of combustion (-727 kJ·C-mol^{-}1). This constitutes a challenge for large scale processes (Jungo et al., 2007; Niu et al., 2013). It creates the need of rapid and efficient heat dissipation, which is not trivial in a large bioreactor where the ratio of surface area to volume is small. If the fermenter temperature increases, it affects the productivity and quality of the recombinant protein (Jungo et al., 2007).

In an aerobic culture, the production of heat is correlated with oxygen consumption. High cell density cultures of *P.pastoris* have a high demand of oxygen. The challenge is to reduce oxygen consumption without affecting the protein productivity. One option

that has been discussed in the literature is the use of co-substrates, with particular focus on sorbitol (Niu et al. 2013), to reduce the use of methanol in the recombinant protein expression phase. A mixed feed strategy enables the cells to avoid dependence on the slow metabolization of methanol which then functions mainly as an inducer (Cos et al. 2006). Glycerol, glucose or sorbitol can be used as co-substrates without loss of protein productivity. However, sorbitol is the most promising one since it is a weak reducer and constitutes a non-repressing carbon for pAOX1 and also an energy source (Niu et al. 2013). Due to these reasons, the oxidation flux in the peroxisome is decreased which leads to less oxygen consumption and heat production. According to Niu and co-workers, 61% of sorbitol goes through the TCA cycle and the rest is used for biosynthesis at 0.5 C-mol/C-mol methanol (Niu et al. 2013).

2.2.2.2.3 Process Monitoring

Cultivation-level factors play an important role in the yield of heterologous proteins, particularly temperature, pH and dissolved oxygen (DO) concentration, since they affect the molecular mechanisms and cell growth (Potvin et al. 2012). As key process variables, methanol concentrations, protein production rates and cell density should be controlled to reach an optimal recombinant protein production (Potvin et al. 2012).

Methanol is a key parameter in *P.pastoris*. While too low concentrations might not be enough to induce the AOX promoter (Cereghino and Cregg, 2000), too high concentrations might be toxic to the cells (Zhang et al., 2000). Methanol is used at a constant rate in the induction phase for the production of the heterologous proteins (Chiruvolu et al., 1997).

2.2.3 Mammalian Cells

Mammalian cells are more complex and expensive to maintain but they enable the highest degree of protein quality and most authentic post-translational modifications. From 2006 and 2010, there were 58 approved biopharmaceuticals of which 32 were produced from mammalian cells (Walsh, 2010).

The most widely employed mammalian cell line for expression of therapeutic pro-

teins is that from Chinese hamster ovary (CHO) cells (Datta et al., 2013). However, there are a number of others mammalian cell lines such as baby hamster kidney mouse (BHK), myeloma-derived NS0 (NS0), human embryonic kidney (HEK-293), and the human retina-derived (PER-C6) (Kim et al., 2012). Currently, CHO cells are used to produce biopharmaceutical compounds, monoclonal antibodies, and Fc-fusion proteins Karbalaei et al. (2020). The reasons for CHO being so frequently used are its many advantages (Kim et al., 2012). First, they have been demonstrated as safe hosts, making it probably easier to obtain approval by regulatory agencies. Second, the low specific productivity of other mammalian cells can be partially overcome in CHO cells by gene amplification. Third, they are able to perform efficient post-translational modifications, which are both compatible with and bioactive in humans. Finally, they can be easily adapted to grow in regulatory-friendly serum-free suspension conditions which is preferred in large-scale cultures. Besides the higher costs involved in CHO production due to slow cell growth, expensive growth media and culture conditions (continuous CO2 supply, expensive transfection reagents), CHO cells are also more prone to potential contamination from virus, which limits its use in large-scale production Karbalaei et al. (2020)

2.2.3.1 Physiology

In 2011, Xu and co-workers published the sequencing of the CHO-K1 genome, constituting a major milestone for cellular and metabolic engineering (Xu et al., 2011). Using 'omics has several advantages. First, it can lead to a better understanding of the relationship between process conditions and glycosylation. It can also be used for expression optimization and to understand the relationship between cell engineering with growth and productivity (Datta et al., 2013).

2.2.3.2 Cultivation Processes

Mammalian cells can grow in adherent or suspension cultures. Adherent cultures can be grown in roller bottles or with spherical microcarriers. The bottles are filled to 10-30% of their capacity with medium and slowly rotated, allowing cells to adhere. This

movement will guarantee a regular wetting of the cells and that oxygen is supplied by the free space in the bottle (Berlec & Strukelj 2013). This cultivation mode can easily allow scale up since it only depends on the number of bottles in parallel, although the ratio of cell to volume is much lower than what can be achieved in an optimized stirred-tank reactor process (Wurm, 2004). Microcarriers can be maintained as a suspension in stirred tank bioreactors, which can be easily scaled-up. Two examples of biopharmarceuticals that use such a system are follicle stimulating hormones and virus vaccines (Wurm 2004).

Suspension cultures allow much higher cell densities which make them very popular in biopharmaceutical production of recombinant proteins. CHO cells are able to grow in single-cell suspension so they can be grown in batch, fed-batch or perfusion modes (Wurm 2004; Berlec & Strukelj 2013).

In order to have a sustained culture growth and an effective recombinant protein production, the cell culture media composition is of major importance. Over 50 years, an essential component of mammalian cell growth media was fetal bovine serum (FBS). It has however been discarded due to its uncharacterized nature, the risk of transmitting adventitious agents (as for example bovine viruses) and its cost (Berlec and Strukelj, 2013). Nowadays, each biopharmaceutical producer has its own proprietary optimized chemically defined media that include peptides, growth factors, proteins, lipids, carbohydrates and small molecules.

2.2.3.3 Process Monitoring

In mammalian cell cultures, nutrients and medium composition need to be tightly controlled as they influence the glycosylation pattern of the product. One example includes glutamine and glucose that have to be supplied in sufficient but not excessive quantities (Berlec & Strukelj 2013). The accumulation of toxic waste (such as ammonia and lactate) generated from these energy sources has to be controlled as they affect cell growth and product quality (Kim et al. 2012).

2.3 Spectroscopic Methods

Spectroscopy is the interaction between matter and electromagnetic radiation. At molecular spectroscopy level, different chemical bonds vibrate at different wavelengths. The effects of interactions between matter and light include absorption, emission, and scattering. These can then be detected by different spectroscopic methods, such as Near-infrared (NIR), Mid-infrared (MIR), Raman, or fluorescence. The type of information and possible applications of some of the spectroscopic methods are discussed below.

Based on the previous literature survey on expression systems, a summary of the analytes involved that are interesting to monitor in the bioprocess is presented in Table 2.1.

Table 2.1: Carbon and nitrogen sources and possible toxic products for each expression system addressed

Organism	**Carbon source**	**Nitrogen source**	**Possible toxic products**
E. coli	Glucose, Glycerol	Ammonium or ammonium hydroxide (NH_4OH)	Ammonium Ethanol (by-product) Glycerol (by-product)
S.cerevisiae	Glucose	Ammonium sulphate Peptides and free amino acids. Glutamic acid	Ammonium Ethanol (by-product) Glycerol (by-product)
P.pastoris	Glycerol, Methanol	Ammonium hydroxide (NH_4OH), Peptone, Amino acid mixture	High levels of methanol (3.7 – 20 g/L)
CHO cells	Glucose	Glutamine	Lactate (by-product) Ammonia (N-waste product)

2.3.1 NIR

Near infrared spectroscopy (NIRS) corresponds to the operational range of 700-2500 nm (14300 - 4000 cm-1) of the electromagnetic spectrum. All absorption bands are the result of overtones or combinations of the fundamental vibrations seen in the mid-

infrared. The higher the overtones, the weaker the intensity and combination bands resulting from the sum of two or three fundamental bands. Mainly vibrations of those functional groups with covalent bonding are observed, i.e. groups with X-H (–CH, –OH, –SH and –NH). While not all molecules absorb in this region of the spectrum, the presence of X-H bonding ensures that NIR measures the majority of fundamental components in bioprocesses (Scarff et al., 2006).

Light can be absorbed, transmitted, reflected and scattered (as shown in Figure 2.2). Absorbance values cannot be directly measured thus spectral acquisition can be based on transmission (T) and reflectance (R). The absorbance values can be obtained by calculating log (1/T) and log (1/R), respectively.

Transmittance (T) is the fraction of incident light which is transmitted. In other words, it is the amount of light that "successfully" passes through the substance and comes out the other side. It is defined as T = I/Io, where I = transmitted light ("output") and Io = incident light ("input").

Reflectance (R) measures the ratio of light intensity reflected by a sample to the light reflected by a background or reference reflective surface. The light that is reflected contains a specular component and a diffuse component. While specular reflectance does not provide much information about composition, diffuse reflectance is very useful and widely used for turbid liquids, slurries and solids. This makes it applicable in bioprocess applications for high cell density processes or ones where the biomass has a complex structure.

Transflectance (T*) combines transmittance and reflectance measurements. In this spectroscopy the light is transmitted through a sample and the unabsorbed radiation is reflected back from a mirror or a diffuse reflectance surface placed at the end of the probe, which doubles the pathlength.

The acquisition mode of the spectra depends on the optical properties of the broth. Transmission is commonly used for low cell density bioprocesses, and reflectance is used for fermentations with strongly light-scattering effects (Hall et al. 1996). In addition, reflectance mode restricts the use of wavelengths beyond 2000 nm due to water absorption band located at 1940 nm which begins to saturate (Hall et al., 1996),

making the transmittance mode more appealing.

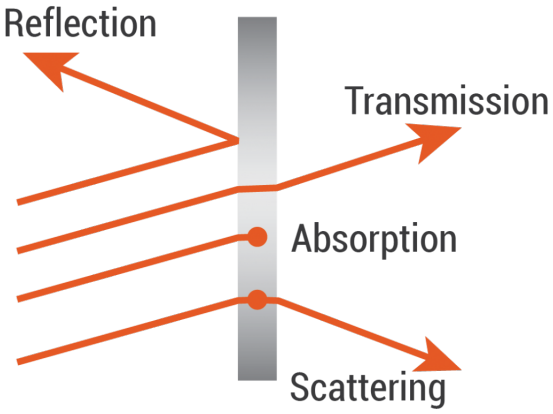

Figure 2.2: Modes for light travel: transmission, reflection and absorption.

The measurement approaches associated with NIR are off-line, at-line and on-line (Cervera et al. 2009). An offline measurement implies that the sample is removed from the process and analysed later. If the sample is analysed immediately and in close proximity to the process stream, the measurement is named at-line. An online measurement does not involve manual handling of samples collected from the process since the spectra are collected directly. Online measurements can be divided into two sub-categories, as depicted in Figure 2.3: *in-situ*(also named in-line) or *ex-situ*. The *in-situ* approach, also known as in-line, consists of locating the spectroscopic probe into the fermentation broth and spectral data flows through the optical fibres into the spectrometer. The two other configurations shown in Figure 2.3 are *ex-situ*. The middle one describes a probe outside the vessel measuring through the wall of the bioreactor. The bottom one describes a flow-through cell where a sample is collected from the bioreactor to be scanned in the spectrometer and returned to the bioreactor.

Reducing production time and a general "time-to-market" is of critical importance to industry. In continuous manufacturing, the quality of the products has to be assessed in real-time. Traditionally, often destructive and time-consuming off-line methods which provide information only hours after sampling are used. Real-time monitoring and control constitute an efficient mean of identifying and reducing variation, managing process risks, relating process information to critical quality attributes (CQAs)

and determining process improvement opportunities (Ündey et al., 2010). Although the at-line measurements represent an improvement over traditional off-line methods and are close to real-time analysis, the ideal approach is to monitor on-line, preferably *in-situ* (Arnold et al., 2002). Although, as pointed out in Arnold et al. (2002), the evolution from at-line to *in-situ* implies some losses as some wavelength regions are lost due to adverse signal to noise ratio. The light intensity is reduced which results in loss of sensitivity. Additionally, probe fouling can occur as well as other contributors to noise, e.g. gas phase effects (bubbles), temperature changes in some processes instead of the equilibrated samples of at-line mode, and vibrational effects due to agitation.

General advantages and disadvantages of NIR spectroscopy are listed in Table 2.2.

Regarding bioprocess monitoring, the more important difficulties that arise in the application of *in-situ* NIR are the rheological characteristics of some bioprocesses, high cell densities, and vigorous gassing and stirring rates of aerobic microbial cultures (Scarff et al. 2006).

Applications

Some examples of NIR applications on monitoring of different expression systems are given below.

E. coli

High cell density culture (HCDC) refers to approximately 10 times the normal cell density of a simple batch culture. If *E. coli* normal cell density is about 5-10 g/L, HCDC should be 50-100 g/L (Lee, 1996; Chang et al., 2014). However, these values are still low enough to allow collection of spectra in transmittance mode (Hall et al., 1996).

Macaloney et al. (1994) reported the collection of transmittance spectra from shake flask fermentations in order to monitor biomass and glycerol. For the biomass model, multiple linear least-squares regression (MLR) was applied and for the glycerol model the second derivative of the spectra was used. On another publication, the same group developed robust MLR models for ammonium, acetate, glycerol and biomass in fed-batch cultivations of *E. coli*, also using transmittance through NIR at-line (Macaloney

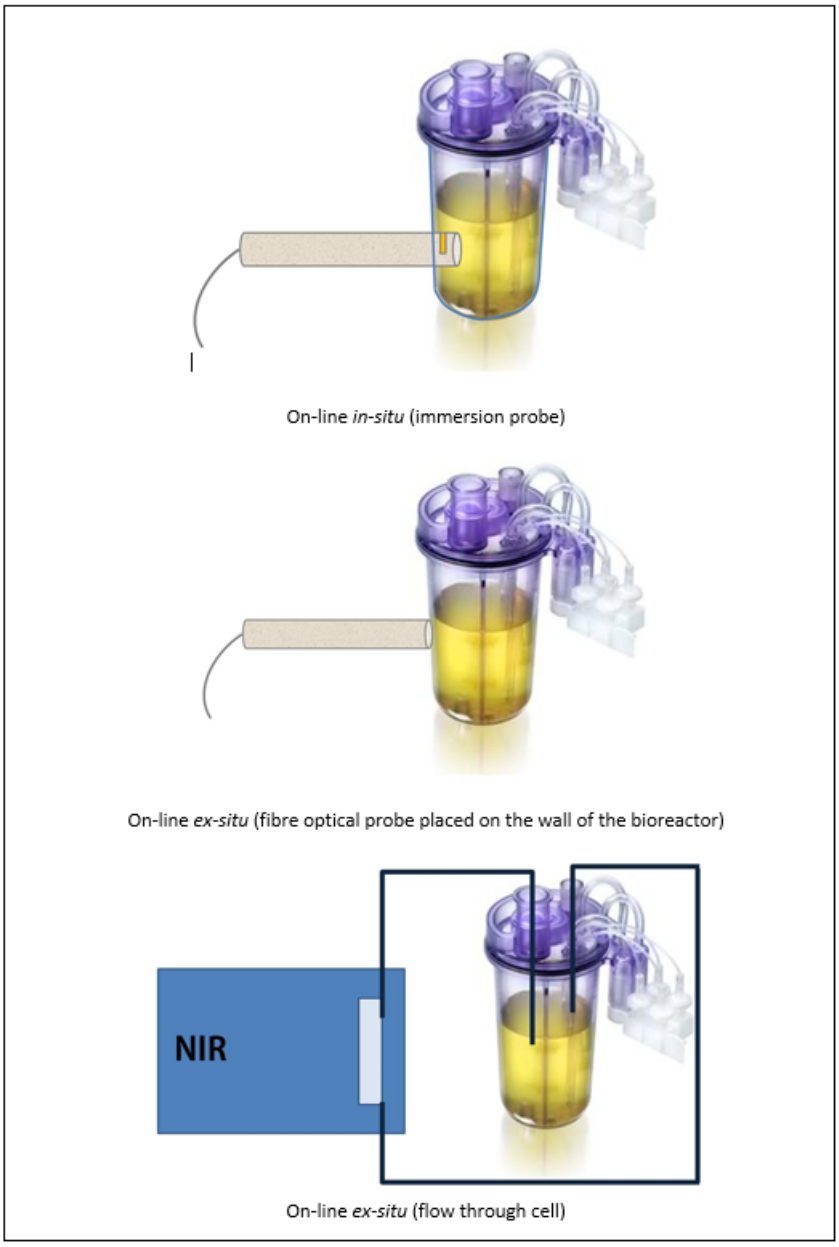

Figure 2.3: On-line measurement configurations: on-line in situ and ex-situ (probe and flow through cell).

Table 2.2: Advantages and disadvantages of NIR spectroscopy. Based on Jamrógiewicz (2012), Cervera et al. (2009) and Scarff et al. (2006).

Advantages	**Disadvantages**
No sample pre-treatment requirements	Low sensitivity (high detection limit) – components in low concentrations (below 0.1-1 ppm) may not be visible
No sample alteration (sample can be re-used after measurement)	High initial investment for the instrumentation
Fast acquisition of spectra ($\leq$1 min)	Development of calibration models requires high trained personnel
Collection of several spectra on the same object makes it possible to obtain more representative sample composition and more accurate result of analysis	Accurate and robust calibration requires large data set convening large variation
Low-cost analysis – no chemical reagents needed; single operator can analyse a large number of samples	Requires continuous maintenance of the calibration data set
Ability to predict physical and chemical properties of the sample	It is difficult to transfer calibrations, even between instruments from the same manufacture
Multipurpose method – several constituents of the same sample can be measured at the same time	Requires an accurate chemical and physical analysis of reference samples
Measurements can be carried out at/on-line	Non-selectivity – compounds active in this region can mask variation arising from the interest analyte(s)
Can be used for particulate level and bulk level with proper calibration	Hydrodynamic conditions in the fermenter change the spectra: position of the probe, agitation rate and aeration rate must be kept constant during calibration and validation

et al., 1994). Arnold et al. (2002) applied NIR, both at-line and *in-situ*, to an industrial fed-batch and PLS (partial least squares) models for biomass were built based on the second derivative of the transmittance spectra.

S.cerevisiae

Several studies have shown NIR applications in *Saccharomyces cerevisiae*. Cavinato et al. (1990) developed a MLR from the second derivative of the spectra collect ex-situ (the fibre-optic probe was placed on the outside of the glass-wall of the fermenter) to determine ethanol concentration during the time course of a fermentation.

The same group used the same NIR approach to monitor biomass in a range of 1-60 g/L and obtained an average standard error of prediction of 1.6 g/L (Ge et al., 1994). These results were valid only if the air flow and the stirring rates remained constant throughout the fermentation. Finn et al. (2006) described the use of NIR for monitoring biomass, glucose, ethanol and protein content in a high cell density fed-batch bioprocess of *S.cerevisiae*. Spectra were collected at-line in transmission mode and robust models were obtained despite the complex matrix.

Recently, *in-situ* monitoring has been described by Corro-Herrera et al. (2016). By using an immersion transflection NIR probe (XDS Process Analytics from FOSS-NIRSystems) and data from 10 fermentation batches, the authors were able to develop PLS models for biomass, glucose, ethanol and glycerol. The standard errors of calibration (SEC) for biomass, ethanol, glucose and glycerol were 0.212, 0.287, 0.532, and 0.296 g/L and standard errors of prediction (SEP) were 0.323, 0.369, 0.794, and 0.507 g/L, respectively.

P.pastoris

Crowley et al. (2005) investigated a high cell density fed-batch process of *Pichia pastoris* and NIR was used to at-line monitor concentrations of biomass, methanol, glycerol and product. The models were built using linear regression (LR) or partial least squares (PLS) and the acquisition mode used was transmission for glycerol, methanol and biomass and reflectance for methanol and product. Second derivative, Standard Normal Variate(SNV) and Multiple Scatter Correction(MSC) were investigated as pre-

treatments.

CHO

Mammalian cell culture processes involve low agitation rates, a modest gas phase, and clean homogeneous non-viscous broths. These characteristics result in a less challenging application of NIR as it is less likely that the quality of the spectra would be affected (Roychoudhury et al., 2007).

Arnold et al. (2003) used *in-situ* NIR to monitor a process of CHO fermentations in an adapted 2-L bioreactor. An immersion (transmission) probe was used and calibration models were developed for glucose, lactate, glutamine, and ammonia. The second derivative of the spectra was used in all models and SNV was also used in all models with exception of that for glucose.

In the work described by Roychoudhury et al. (2007), models were developed for glucose and lactate by using *in-situ* transflectance fibre-optic probes. They compared the use of one single probe with multiplexed measurements and also conducted a signal intensity study to determine the effect of each optical component (spectrometer channel, probe design, mirror optical properties) on the overall variability in signal and on spectra acquired. Differences in probe design were found to be the factor that most contributes to optical variation in signal. Multiplexed measurements present several advantages compared to one single probe as multiple vessels can be monitored simultaneously instead of gathering a calibration set from several runs in just one bioreactor. Therefore the time needed for the calibration process is considerably reduced. Furthermore, multiplexed systems allow model transferability which means that models developed from small bioreactors can be employed to large scale. Although this study showed a slight degradation of multiplexed models when compared to single probe models (less than 1.4% for both analytes) the models incorporated inter-probe variability and they were fit for purpose.

Chen et al. (2011) proposed a novel calibration model for the analysis of complex spectral data sets arising from multiplexed probes. This Multiplex Calibration Model (MCM) was applied on data collected *in-situ* on six bioreactors by using six transflectance fibre-optic probes. PLS models with different common preprocessing

methods (OSC, SNV and MSC) were developed for prediction of glucose. However, the proposed MCM model outperformed all the other models by achieving a 54% reduction in the RMSEP values for the test samples.

Clavaud et al. (2013) have shown the ability of NIR in large scale for manufacturing of biopharmaceuticals. They have developed models based on twelve 500 L-scale cultivations of CHO cells to predict glucose (and other culture parameters), using a Fourier transform near infrared (FT-NIR) multiplex process analyser instrument (Clavaud et al., 2013).

2.3.2 MIR

The Mid-Infrared region (MIR) of the electromagnetic spectrum corresponds to the vibrational measurements in the range of 2500 to 25000 nm (4000 – 400 cm^{-1}) (Siesler et al., 2002). It covers fundamental vibrations of most of the common chemical bonds (Lourenço et al., 2012). The region above 1500 cm^{-1} corresponds to absorptions bands assignable to a number of functional groups and at low energy, below 1500 cm-1, a series of absorption bands resulting from vibrations of the molecule as a whole are shown (Siesler et al., 2002). Therefore, the region below 1500 cm^{-1} shows absorption bands characteristic of the compound in question and no other compound. This is known as the "fingerprint region" (Siesler et al., 2002). Moreover, absorption bands from the molecular skeleton (e.g., C-C, C=C, and C=O) or functional groups containing heavier atoms (e.g. C-Cl and C-N) in the fingerprint region are already into the second or third overtone above 4000 cm^{-1} (Lourenço et al., 2012). Thus, for bioprocess monitoring, MIR provides enhanced sensitivity and selectivity (Roychoudhury et al., 2006).

Fourier transformation (FT) is commonly employed to solve problems associated with signal stability, and measurement noise when applying conventional MIR for process measurements (Roychoudhury et al., 2006) FT spectrophotometers are useful in the determination of component concentrations in complex matrices such as a typical fermentation fluid (Doak and Phillips, 1999).

The attenuated total reflectance (ATR) technique transformed MIR into a method with little or no sample preparation and allowed the collection of high-quality spec-

tra from aqueous samples with multiple analytes present (Doak and Phillips, 1999; Roychoudhury et al., 2006). The ATR crystal, through which the light is reflected, is composed of a material with high refraction index (Lourenço et al. 2012). The sample is placed in optical contact with the ATR crystal and the light partially enters the sample where it can be absorbed, resulting in attenuated reflected light. These probes are particularly suited to measurements in highly absorbent samples, for which standard transmission probes cannot be used (Lourenço et al. 2012). In addition, the short, reproducible pathlength of the device enables the measurement of spectra without the complication of water bands. This obscures certain portions of the MIR-spectrum, as the spectral contribution of water is simply eliminated by spectral subtraction (Acha et al., 2000; Roychoudhury et al., 2006).

The spectral collection modes available are the same as in NIR: off-line, at-line or on-line. Again, the mode should be chosen according to the optical properties of the fermentation broth, the required analyte selectivity and sensitivity, the duration of the run, and the monitoring and control requirements (Roychoudhury et al. 2006). As before, the approach that enables an accurate determination of analytes in real-time is on-line monitoring.

Notwithstanding the advantages of MIR, NIR is easier to apply *in-situ* as the NIR optical fibres are robust and inexpensive when compared to MIR (Vojinović et al., 2006). The materials available to construct fibres that can efficiently transmit MIR radiation (e.g. chalcogenide) do not effectively cover the full mid-infrared range. These fibres are expensive, easily damaged and have poor transmission characteristics when compared to NIR fibres. Since they can be easily damaged, they are restricted in length (1-5 m) and have to be greatly armoured and protected from physical strain (Roychoudhury et al. 2006). In addition, some materials can obscure some regions as they have strong absorptions (e.g. chalcogenide glasses near 2000 cm^{-1}).

A summary of advantages over NIR as well as limitations is present in Table 2.3.

Table 2.3: Comparison between MIR and NIR (based on Roychoudhury et al. (2006)).

MIR	**NIR**
Broader wavelength range (MIR absorbances are based on fundamental vibration modes of molecules)	Limited multiplexing advantage
Inherent potential to include process variations through chemometrics	Fragile and expensive fibres
Ability to quantify analytes present at limiting levels	Sample presentation
Ability to distinguish between analytes of similar structure	Detector saturation
Potentially identify aspects of product quality (e.g. folding patterns)	

2.3.3 Raman

Raman spectroscopy is based on inelastic scattering of incident light by a sample, shifted in frequency by the energy of its characteristic molecular vibrations (Kneipp et al., 1999). The incident light is a monochromatic laser light, typically producing light in the visible (e.g., 532 nm) or near-infrared (e.g., 785 nm and 1064 nm) range (De Beer et al. 2011). The size of the wavelength shift and the number of different shifts that will occur depend on the molecules vibrational levels (Jestel 2005). It is often considered a complementary technique to NIR, as bonds that are strongly IR-active will be weak Raman-active and vice-versa. Highly symmetric molecules, particularly homonuclear diatomic species such as C–C, C=C, N=N, S–S, generate strong Raman scatter and are correspondingly weak in infrared region (Jestel, 2005).

The major advantage of Raman spectroscopy is its flexibility as it can be used for solid, liquid and gas samples or slurries, and gels without sample preparation (Pons et al., 2004; Lourenço et al., 2012). The spectra are not sensitive to water and can provide data about the composition, chemical environment, and structure of the sample, enabling structural, qualitative, and quantitative analysis (Lourenço et al., 2012).

However, the signal tends to be weak and some spectral interference can arise from fluorescence of some biological molecules in this region (Lourenço et al. 2012). Therefore, the critical decision in the selection of a Raman instrument is the selection of

a laser wavelength to maximize the signal and minimize fluorescence (Lourenço et al. 2012). Another disadvantage is that high power lasers are necessary which increases the complexity and cost of the instruments; not all appropriate wavelengths are available and high laser power can destroy delicate or dark samples (Pons et al. 2004; Lourenço et al. 2012). Also, use of these lasers requires a moderate skill from the operator in the alignment and operation of the equipment compared with NIR or MIR techniques (Sivakesava et al., 2001).

Recently, Raman has seen great developments, its associated costs have decreased and it is expected that it becomes more robust, to have higher throughput, smaller, easier to use, and eventually more automated (Buckley and Ryder, 2017). Even though not as widespread as NIR, an increasing number of studies have recently been published which confirm the ability of this spectroscopy to fully satisfy the requirements of a PAT tool with respect to monitoring and control (Jenzsch et al., 2017; Kozma et al., 2017). The work from Abu-Absi et al. (2011) was the first to report the implementation of Raman spectroscopy for bioprocess as they monirored glucose, lactate, glutamine, glutamate and ammonium during a 500 L-scale, fed-batch CHO cell cultivation.

Berry et al. (2016) have also measured and controlled glucose concentration in CHO fed-batch cultivation, through the use of Raman, to keep glucose concentration below a specific level to avoid glycation of the target protein. They have also studied model transferability using a small scale (5 L), a pilot scale (200 L) and a manufacturing scale (2000 L) bioreactor.

Esmonde-White et al. (2017) have done a good review on the topic and considered Raman a valuable PAT for fermentation or cell culture bioprocess monitoring and control (Esmonde-White et al., 2017).

2.3.4 Fluorescence

Fluorescence spectroscopy is restricted to species which fluoresce. A fluorescent molecule is excited by irradiation with light because it absorbs a photon and emits light of lower energy afterwards. Fluorescence spectroscopy is a useful tool for bioprocess monitoring since there are several biological compounds that fluoresce on excitation by visible or

near-UV light (e.g., proteins, enzymes, coenzymes, vitamins, and primary or secondary metabolites from microbial growth) (Lourenço et al. 2012). This spectroscopy has the advantage, over vibrational spectroscopies, of being applicable to high concentrations of biomass (Musmann et al., 2016). In some cases, it can also show higher sensitivity and selectivity Lourenço et al. (2012).

2.3.5 Comparison Studies

Some studies have compared the above described spectroscopical methods. A review from Musmann et al. (2016) compared spectroscopy applicability for high-throughput characterisation of mammalian cell cultures in automated cell culture systems. Raman and NIR are advised for the determination of metabolite concentration and the combination of the results of both techniques results in complementary information. Given that no single method is perfect, and it always depends on the type of application, there is no single recommendation. However, the authors note that Raman offers the greatest robustness, highest information density, a wide range of applications, and the broadest range of automated measurement types. On the other hand, the cost of a Raman system is higher than a NIR instrument.

Kozma et al. (2017) have compared NIR and Raman spectroscopy by developing a model using a shake flask CHO cultivation and testing the model's predictive ability of online spectra collected in 10L and 100L bioreactors (Kozma et al., 2017). The model based on NIR spectra could predict the trend of glucose concentration but without sufficient accuracy for bioreactor monitoring. However, the Raman based model performed better and was able to predict both cultivations scales with an error of approximately 4 mM (0.72 g/L).

Trunfio et al. (2017) have compared NIR, MIR, Raman and fluorescence spectroscopy to characterize the variability of wheat hydrolysates, used in CHO cultivations, and to provide evidence that the classification of good and bad lots of raw material is possible (Trunfio et al., 2017). Models built with fluorescence and NIR spectra yielded lower prediction errors than the models built from middle infrared and Raman spectra. From the two best techniques, the final recommendation from the authors is NIR as

these models performed slightly better.

Finally, Li et al. (2018) studied the ability of *in-situ* Raman and Near Infrared (NIR) spectroscopies to predict the concentration of viable CHO cells, glucose, lactate, glutamine, ammonium and antibodies in bioreactors. They found that Raman spectroscopy was better in predicting concentrations of glucose, lactate and antibody, while NIR spectroscopy predicted glutamine and ammonium ion concentrations better.

2.4 Chemometric Techniques

Chemometrics is the science of extracting information from chemical instruments by data-driven means, mirroring other sciences such as econometrics and psychometrics. By chemometrics, trends in spectra can be detected and relationships between spectral changes and, for example, changes in analyte concentrations can be measured (Cervera et al., 2009). Spectroscopic methods generate very large amounts of data which makes chemometrics a requirement for the data analysis.

A flow diagram of typical steps used in chemometrics is shown in Figure 2.4. Once a dataset of spectra is collected it should be randomly split into calibration and validation datasets. Most samples (two-thirds) should be part of the calibration set, while the validation set should be a third of the total collection of samples. Selection of the appropriate spectral pretreatment and variable selection is the second step of this procedure. Iteratively, a multivariate regression model should be obtained for each tested pretreatment. The validation dataset is then predicted by the model and comparative results should be obtained.

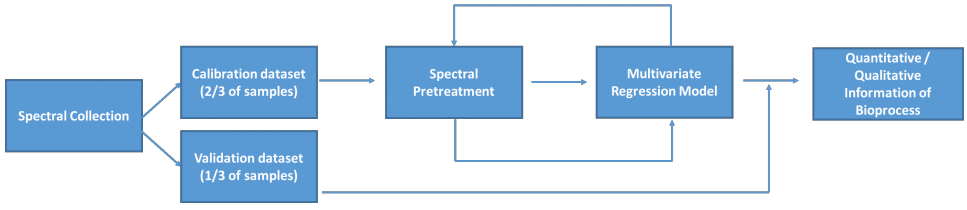

Figure 2.4: Chemometrics flow diagram to extract information about the bioprocess from spectroscopic data.

2.4.1 Regression model approaches

The most common and straightforward modelling approach is Single or Multiple (SLR or MLR) Linear Regression. Both depend upon a simple relationship between detected peak height at a certain wavelength and analyte concentration. This approach has mostly been used for modelling dominant analytes with few interferences (Scarff et al. 2006), e.g. oil as described in Vaidyanathan et al. (2000) and Arnold et al. (2000), and biomass in Crowley et al. (2005). These models were derived from few wavelengths, at most, and thus baseline and peak drifts can strongly impact the results extrapolated from the SLR or MLR. Complex fluids with strong absorption and scattering nature require more advanced multivariate techniques that can deal with such complexity (Scarff et al., 2006). The most commonly used techniques are principal component analysis (PCA) and partial least squares (PLS).

PCA can be described in matrix notation or geometrically. In matrix notation and for the case of spectral data, a two-way array is generated when spectra are collected over time. Such data can be arranged in a matrix (X) with K variables and N objects. For spectra, the variables are wavelengths and the objects are samples measured at different time points. PCA is a projection method that extracts the systematic variation found in X. The matrix is decomposed into a sum of matrix products, where one matrix is called scores (T of size PCs times N) and the other is called loadings (P of size PCs times K). The variation not described by the conducted PCA model is found in matrix E (which is called residuals). In matrix notation, PCA can be describe as:

$$X = T.P^T + E \tag{2.1}$$

Geometrically speaking, a PCA summarises the variation of the data by identifying directions in the original data. These directions are estimated as linear combinations of both the original variables and the observations, which makes PCA a bilinear model (Esbensen et al., 2002). The first created direction, called the first principal component (PC1), describes the maximum variance found in the original data. The second direction, named the second principal component (PC2), is orthogonal to the first PC

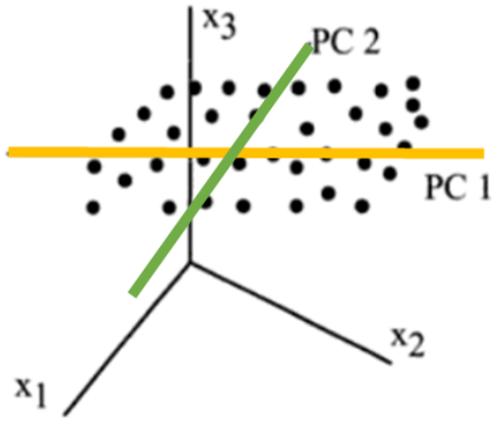

Figure 2.5: A two component model, where the first and second principle components (PC1 and PC2) pass through the average-point of the data cloud. The first PC is the yellow line and the second PC is the green line.

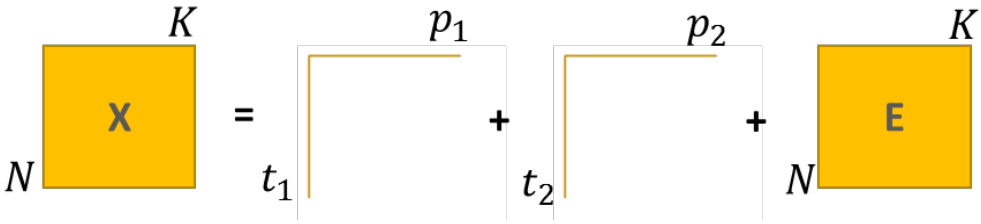

Figure 2.6: A PCA model decomposing matrix X into a score vector (T), a loading vector (P) and residuals (E).

and the process continues until descriptive variation is found (Esbensen et al., 2002). This is illustrated in Figure 2.5 for two components.

An example of a two-component PCA model is shown in Figure 2.6 in which an original space of K by N is described in a two-dimensional subspace. In the Figure, t1 and t2 are the scores vectors and p1 and p2 are the loading vectors for PC1 and PC2.

The score value (p) of the first component is defined by the projection of the original position of a sample onto PC1 (specifically, the direction determined by the loading vector for PC1), as well as for the second score value. (Esbensen et al., 2002).

The obtained scores values and loadings will elucidate trends and relationships that were initially obscured in the spectral data.

Partial Least Squares (PLS) expands on the PCA by relating two data matrices: X

(spectra) and Y (the reference data, e.g. analyte concentration). Dependent variables (Y) are used to decompose independent variables (X) and the goal is to determine a small number of latent variables that can predict the property of interest by using the spectral data as efficiently as possible (Wold et al., 2001; Lourenço et al., 2012).

Regarding the risks in applying these techniques, one can point out the fact that in fermentation systems there is a stoichiometric relationship between analytes. This may imply that a PLS model of one analyte may not be based on the spectral characteristics of that analyte alone or even at all (Roychoudhury et al., 2007; Warnes et al., 1996). This characteristic of fermentation system datasets is often termed co-linearity (Riley et al., 1998). Careful examination of the spectral contributors to the models can help identify where and to what extent it is occurring. Alternatively, the regions where specific analytes are contributing can sometimes be unambiguously identified by spiking with known amounts of the analyte (Riley et al., 1997). Adaptive calibration can be used to remove or reduce the co-linearity in the data by artificially altering the concentration of one analyte independently of the others (Riley et al., 1998). Neither approach is suitable for on line or *in-situ* spectroscopy since it relies on sample manipulation.

2.4.2 Model Performance

A good predictive model should be accurate, by showing low errors of prediction, and robust, which can be evaluated by comparing the different errors of estimation. The root mean square error (RMSE) is obtained by the sum of squares of the residuals of Y divided by the degrees of freedom (as shown below in Equations 2.2 and 2.3), where A is the number of PLS components and A_0 is 1 or 0, depending on whether Y is centred or not. Depending on the dataset used, this error can be named root mean square of estimation (RMSEE), when the calibration dataset is used for the calculation, or root mean square of prediction (RMSEP), when an independent "prediction" dataset is used. A robust model is able to predict different datasets within the calibration range with similar accuracy thus RMSEP should be similar to RMSEE. A good predictive model should also show a high coefficient of determination (R2) and randomly distributed

residuals ($Y_{obs} - Y_{pred}$) to ensure overfitting has been avoided.

$$RMSEE = \sqrt{\frac{\Sigma(Y_{obs} - Y_{pred})^2}{n_{calibSet} - A - A_0}} \tag{2.2}$$

$$RMSEP = \sqrt{\frac{\Sigma(Y_{obs} - Y_{pred})^2}{n_{predictionSet}}} \tag{2.3}$$

2.4.3 Calibration, External Validation and Outlier Detection

In the calibration stage, both X and y need to be known. The property to be modelled (e.g. concentration) constitutes the y vector and the spectra variability forms the X matrix. The calibration dataset should include all possible variations and the model should not be used to predict outside the calibration range. Variability in spectra can arise from different causes (e.g. sampling procedure, differences in raw materials). Thus, the model should be developed to describe the desired variability while not being sensitive to common variability sources of the process (Henriques et al., 2009).

Ideally, there is also a true independent set to perform external validation i.e. from a different batch, unseen by the model, which will be a better indicator of how the model will perform afterwards (Henriques et al. 2009).

One important thing to have in mind while modelling is that, frequently, there are observations which are far from the rest of the main dataset cluster; these are called outliers (Naes et al., 2002). They might be due to variability in the measurements or they may indicate experimental error, and they could reduce the accuracy of the model. Some statistics such as the Mahalanobis distance or Q residuals allow the identification of outliers.

2.4.4 Spectral Pre-treatments

Pre-treatments are mathematical methodologies that are usually applied to spectra to minimize variability unrelated to the property to be modelled. For example, NIR is sensitive to chemical and physical properties and if an analyte concentration is to be modelled, it is important to apply appropriate preprocessing to minimize physical

effects before calibration of the model. In this case, the success depends on the ability of the mathematical treatment to separate light scattering from light absorbance (Huang et al., 2010).

Bioprocess fluids are especially challenging for NIR measurements due to matrix effects, large absorption bands, and light scattering differences (Roychoudhury et al., 2007). The most commonly used techniques for bioprocess samples are discussed below.

Mean-centering

Mean-centering is a normalization pre-treatment that consists of subtracting the response of each variable from the mean response of that variable over all samples in the data set. It enhances response variations since it removes the absolute intensity information from each of the variables (Lourenço et al. 2012). Furthermore, it also reduces the final model complexity, often reducing the number of variables to be employed.

SNV

The standard normal variate (SNV) is obtained by correcting each absorbance variable by the mean absorbance of the spectrum; this result is then divided by the standard deviation of the variables under investigation. This way, additive and multiplicative effects and shifts are corrected, then each spectrum is scaled, which generally results in a cleaned-up spectrum of noise and interferences (Crowley et al. 2005). The subtraction of the mean corresponds geometrically to the projection of the points that represent raw spectra orthogonally onto the plane (Fearn et al., 2009). The division by the standard deviation results in scaling all spectral components to the same length (Fearn et al., 2009).

MSC

Multiplicative scatter correction (MSC) removes scatter effects, both multiplicative and additive components which improve linearity, reducing the number of components needed in the regression model (Lourenço et al. 2012). It can be applied to the parts of the spectrum affected by light dispersion (Lourenço et al. 2012). For each sample, MSC fits a straight line to the spectral reading against the corresponding points of the mean spectrum by least squares. It then uses the slope and intercept of the line to correct

the scattering effects on the spectrum (Warnes et al., 1996; Crowley et al., 2005). MSC can be useful when dealing with moderate to high cell density fermentation processes. It can reduce the scattering caused by biomass concentration levels as shown in studies by Crowley et al. (2005) (high cell density *P.pastoris*) and Roychoudhury et al. (2007) (CHO cells).

However, MSC assumes that offset and multiplicative spectral effects are much larger than effects from changes in sample chemistry which can result in poor modelling results when this is not the case and instead chemical-based variations are greater (Bakeev, 2010).

Second derivative

It is common, especially for biomass monitoring, to deconvolute the overlapping peaks and reduce baseline shifts in the raw spectra by taking the 2nd derivative of the spectra (Arnold et al. 2002). Although the signal-to-noise ratio (SNR) might be sacrificed with this technique, the quality of the peaks is important, while the SNR is naturally low in NIR spectra due to the physics of the system (Arnold et al. 2002). This can be checked by running a performance test prior to the use of NIR (Arnold et al. 2002).

Table 2.4: Reasons for using some techniques and their pitfalls.

Problems	**Solution**	**Problems that can arise**
High cell density	MSC	
Baseline drift, Deconvolute overlapping peaks	Second derivative	Signal-to-noise ratio (SNR) can be sacrificed (it can be checked with a performance test prior to using NIR).
Erroneous noise or interferences	SNV	

2.5 Sources of Variability in NIR Bioprocess Modelling

NIRS is the most mature spectroscopic method and therefore the most widely used in bioprocesses. Numerous advantages were pointed out and exemplified, however NIRS should be applied with caution, considering that many sources of variability exist that can be manifest in spectral data unknown to the user.

Fermentation processes require complex model development for prediction of analytes. Bioprocesses involve very challenging matrices, commonly vigorous stirring and aeration rates, or liquid phases with high viscosities and non-Newtonian behaviour. In addition, these rheological properties may vary the cultivation time course, from batch-to-batch and with bioreactor scale (Rodrigues et al., 2008; Cervera et al., 2009). Also, samples from bioprocesses typically contain gas, solid, and possibly more than one liquid phase (Vaidyanathan et al. 2000). As hydrodynamic conditions affect the spectra, there is a consensus that the position of the probe in the fermenter, the agitation rate and the aeration rate must be kept constant during calibration and validation runs (Cervera et al., 2009).

Besides physical aspects, the chemical composition of the sample matrix itself is also an important source of sample variance in bioreactor monitoring (Rhiel, Ducommun, Bolzonella, Marison and von Stockar, 2002). A matrix is the set of all compounds that may influence the measurements of the compound to be determined (Massart et al., 1988).

Other sources of variability include the flex angles of the fibre bundles that result in a different light path. If there is gas intrusion in the measuring field it will interfere with the spectra collection. In addition, it is necessary to be careful with excessive light scattering, e.g. resulting from high biomass content, as it may saturate the measurement.

Multiplex systems are useful but may involve more sources of variability. Through multiplexed measurements, the extensive calibration process is reduced, as several fibres are used to monitor several bioreactors at the same time. These systems allow for model transferability but variability between optical probes should be addressed and

carefully verified. Roychoudhury et al. (2007) evaluated the effect of the optical characteristics of the different probes, mirrors and spectrometer channels on the calibration models, by evaluating the impact of these factors upon signal intensity. From these, the probe was the most influencing factor; thus, the optical compatibility between probes in multiplexing applications must be ensured. They also reported that more variance occurs at lower wavelengths, which can be explained for these regions being less energetic (Cervera et al. 2009). Even though the probe design revealed to be the most relevant, all these factors revealed some effect in the signal intensity, which means that they all can contribute to introducing errors in NIR signals if not consistent during the runs.

Other common sources of variability are the result of the lengthy processes. Vaidyanathan et al. (2000) reported the importance of ensuring the robustness of the model over an extended period, since factors such as batch-to-batch variability, e.g. in substrate feedstock, can contribute to spectral variations. To do so, this group evaluated the model with an external validation set from two years after the development of the model.

Once again, Vaidyanathan et al. (2001) focused on uncontrolled variations over time, such as instrumental drift, and therefore conducted a six-month experiment. With external validation, useful models and anomalies within them were identified. Models for weak absorbers were vulnerable to changes in the matrix while the change in scale affected the models probably depending on the morphology of the microorganisms. The mechanical disaggregation of biomass decreases particle size which theoretically increases light scattering and thus influence the measured spectra which can then affect the performance of a chemometric model that was previously built for a fermentation process. The study also shows that changes in morphology influence the spectra to a lesser degree in the longer-wavelength NIR region (1600–2350 nm), than in the visible or short-wavelength regions.

2.6 Outlook

The purpose of this thesis is to expedite biopharmaceutical research and development by using spectroscopy during the early stages of development. Biopharmaceuticals can be produced through different expression systems which are all characterised by their unique set of analytes. The goal of bioengineering is to optimize the bioprocess to maximize the yield of the biopharmaceutical being produced. This requires a thorough understanding of the expression system being used. The first step is to understand these systems' physiology and their related key analytes as well as the best methodologies for their cultivation.

By real time process monitoring, spectroscopy can allow an even deeper understanding of the bioprocess. This however requires an understanding of sources of variability that come into play during the cultivation process. Chemometrics techniques can help to separate noise from meaningful signal, allowing spectroscopic methods to monitor the cultivation process correctly.

These lessons will guide the following chapters. In characterising the fermentation matrix, the next chapter will focus on *E. coli* as an expression system and test at-line and in-line instruments while focusing on different sources of variability involved in these settings. The chapter after that will focus on *P.pastoris*. Chemometrics lessons will be applied throughout all chapters.

Chapter 3

Characterization of the Fermentation Matrix by NIR Spectroscopy

3.1 Introduction

The main goal of this thesis is to critically evaluate the feasibility of implementing spectroscopic based real-time monitoring, in a range of biological systems used for biopharmaceutical production. The key challenge in doing so is to accurately identify and measure various sources of variability which can be done using two different NIR instruments: at-line and in-line.

At present, a much better understanding of how these impact the spectra and, consequently, the ability to predict analyte levels is required. Ultimately, actions could be then taken to either reduce or eliminate the sources of variability prior to spectral acquisition or to develop novel ways of dealing with the interfering factors in the calibration models' post-spectral acquisition.

This chapter seeks to set a baseline, by evaluating the effects of different fermentation matrix components on the spectral signal. This is done by starting from the simplest setting, i.e. water, while building up to more complex backgrounds such as buffer, fresh media, spent media or presence of cells, using different concentrations of analytes and backgrounds. For the in-line setting, the additional variation stemming from process variables is also considered. Glucose and lactate were the selected analytes to be quantified in the fermentation matrix. Glucose is the main source of carbon for many expression systems and lactate is a common byproduct that can become toxic to cells and inhibit production of the biopharmaceutical. Real time information on the amounts of these analytes in the fermentation medium would allow for an adequate level of glucose and low levels of lactate to be kept in order to maximise the cells growth.

These different sources of variability were analysed using PCA and PLS models. By analysis of the loadings of the developed PCAs, the different effects in the spectra were attributed to their respective components of the fermentation matrix. Second, by developing individual PLS models for each background, the accuracy of such a system was quantified. Finally, one of these models was selected for each analyte and tested as a potential "universal" model that can predict the analyte across different backgrounds. Developing such a universal model would allow for a straightforward implementation

of spectroscopy into the early stages of bioprocess development.

The remainder of this chapter is structured as follows. The next section will describe the materials and methods that were used in the set-up of the research design. After that, the results would be presented prior to establishing a conclusion.

3.2 Materials and Methods

This section presents the set-up of the research design. Chemicals used and prepared mixtures are listed on Table 3.1. Cells were obtained through a fermentation of *E. coli* for which the bioreactor, media composition and instruments used will be described.

Table 3.1: Manufacturer and reference of each material used for this study.

Identification	Reference	Manufacturer
Agar	LP0011	Oxoid
Ammonium citrate dibasic	09831	Fluka - Sigma
Ammonium sulphate	A4418	Sigma
Calcium chloride dihydrate	C/1560/53	Fisher
Cobalt (II) chloride hexahydrate	25559-9	Sigma
Cupric sulphate pentahydrate	C-7631	Sigma
Di-Potassium hydrogen phosphate	1.05104.1000	Merck
Disodium EDTA dehydrate	E1644	Sigma
Ethanol	34870	Sigma
Ferric chloride (Iron(III) chloride hexahydrate)	236489	Sigma
Glucose, anhydrous	0188	Amresco
Lactic Acid solution	252476	Sigma
Magnesium sulphate heptahydrate	M1880	Sigma
Manganese sulphate monohydrate	M-7634	Sigma
Phosphate buffer (Dulbecco A)	BR0014G	Oxoid
Polypropylene glycol 2000	297776T	VWR
Potassium dihydrogen phosphate	26936.320	VWR
Sodium hydroxide	5588-1	Sigma
Sodium phosphate monobasic dihydrate	04269	Sigma
Zinc sulphate heptahydrate	Z-0251	Sigma

3.2.1 Mixtures preparation

Two biologically relevant analytes, glucose and lactic acid, were scanned in different concentrations and backgrounds. Four different levels of concentrations of glucose and lactic acid were used, as presented in Table 3.2. A stock solution of 50 g/L of each analyte was prepared. This translates in molar concentrations of 278mM for glucose and 555 mM for lactic acid.

Table 3.2: Glucose and Lactic Acid used in different levels, in g/L and in mM.

Level	Glucose		Lactic Acid	
	g/L	mM	g/L	mM
Zero	0.00	0.00	0.00	0.00
Low	1.25	6.94	0.58	6.44
Medium	2.50	13.9	2.50	27.8
High	10.0	55.5	10.0	111

The backgrounds tested were water, phosphate buffer, fresh media, spent media along with washed cells content. Spent media and washed cells were obtained by running a fermentation of *E.coli* as described below. Online spectra were collected *in-situ* during the run. The final solutions were then randomly scanned with an at-line NIR system. Spectra collection was randomized with respect to the concentration of the analytes to incorporate other non-relevant sources of variation such as temperature or instrument drift.

3.2.2 Batch run of *E. coli*

A batch run of *E. coli* was carried out in a stainless steel 15 L SIP bioreactor (Biostat® C-DCU, Sartorius) shown in Figure 3.1) and spectra was collected during the run.

The reactor was equipped with four internal baffles (1.5 cm x 57 cm) and three six-blade Rushton turbine impellers with adjustable height. Their height was adjusted

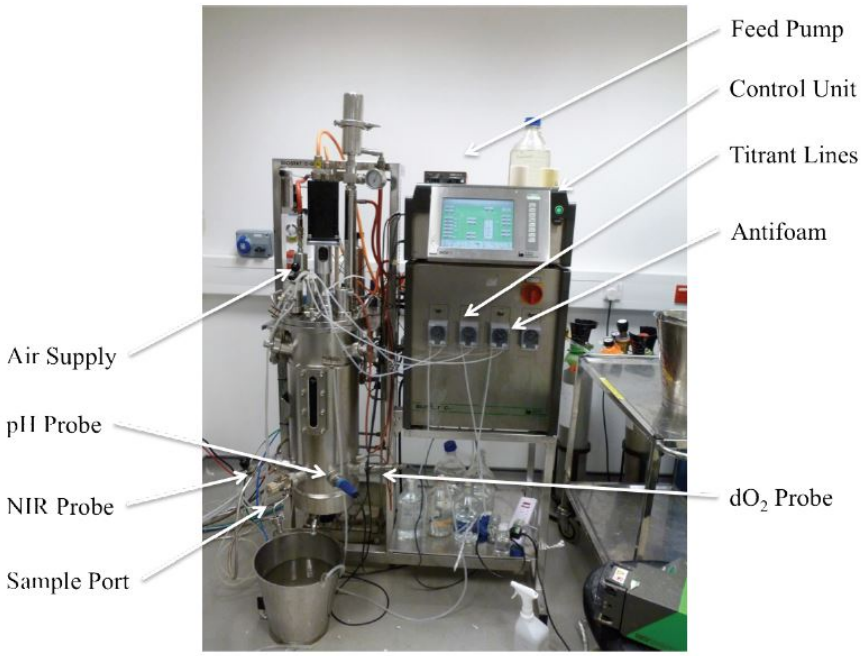

Figure 3.1: Biostat C-DCU system illustrating the locations of the key components and features of the bioreactor system.

in such a way that all the impellers would stay with the maximum distance between each other but still immersed. Temperature control is provided via a water jacket and aeration via a circular annular sparger located at the bottom of the vessel.

Four ports in the bottom of the reactor housed the pH probe (Mettler Toledo Ltd., Leicester, UK), dissolved oxygen (dO2) probe (Mettler Toledo Ltd., Leicester, UK), sampling port and near infrared spectroscopic probe (XDS Process Analytics Microbundle Multiplexer Instrument, FOSS NIRSystems, Maryland, USA).

3.2.3 Media composition

The composition of the medium used was based on Åkesson et al. (1999)) as shown in Table 3.3. A total volume of 10 L growth medium was prepared for batch fermentation. The media components (excluding those identified in Table 3.3) were weighed out and added to a volume of 9 L distilled water. The solution was stirred until all media components were completely dissolved. This was then transferred to the bioreactor

system and *in-situ* sterilised at 120°C for twenty minutes.

Table 3.3: Media components contents. The ones marked with * were added after sterilisation of the other components *in-situ*.

Media component		
Ammonium sulphate	$(NH_4)_2SO_4$	2.0 g/L
Potassium phosphate dibasic	K_2HPO_4	14.6 g/L
Sodium phosphate monobasic dihydrate	$NaH_2PO_4 .2H_2O$	3.6 g/L
Ammonium citrate dibasic	$HOC(CO_2H)(CH_2CO_2NH_4)_2$	0.5 g/L
Polypropylene glycol		0.1 mL/L
Glucose*		11.9 g/L
1M Magnesium sulphate*		2 mL/L
Trace Elements*		2 mL/L

Table 3.4: Trace elements solution composition – components and respective concentrations in g/L.

Component		C (g/L)
Calcium Chloride Dihydrate	$CaCl_2 .2H_2O$	0.5
Ferric Chloride /// Iron(III)Chloride Hexahydrate	$FeCl_3. 6H_2O$	16.7
Zinc Sulfate Heptahydrate	$ZnSO_4 .7H_2O$	0.18
Copper Sulfate Pentahydrate	$CuSO_4 .5H_2O$	0.16
Manganese Sulfate Monohydrate	$MnSO_4 .H_2O$	0.15
Cobalt Chloride Hexahydrate	$CoCl_2. 6H_2O$	0.18
Disodium EDTA dihydrate /// Ethylenediaminetetraacetic acid disodium salt dihydrate	$C_{10}H_{14}N_2Na_2O_8 .2H_2O$	22.3

The glucose solution was autoclaved separately and it was prepared with 119g of glucose and water added up to 996mL. This solution was added aseptically along with the other media components that required filter sterilisation (through a syringe using 0.2 µm cellulose acetate filter).

Two one-litre conical flasks, containing 300mL of media each, were prepared and autoclaved. For this, glucose was used from a stock solution of 50% (w/v) glucose that was previously prepared and autoclaved. For a total of 600 mL of media, 120 mL of salt

solution (Table 3.5), 1.2 mL Trace elements solution (Table 3.4), 1.2 mL magnesium sulphate 1M and 12mL of glucose 50%(w/v) were put together. The pH was checked, then each flask inoculated with a cryovial of *E.coli* from the cell bank and incubated at 37°C, and shaken at 250rpm. Once the optical density was read approximately at 1 (against a water background) at a wavelength of 600 nm, one of the culture flasks was utilised as inoculum of the 10 L medium.

The temperature of the bioreactor was kept constant at 37°C, the stirrer speed was set at 350 rpm, the air flow was controlled at 1vvm (10slpm) automatically and the pH was not controlled.

Table 3.5: Salt solution composition.

Component		C (g/L)
Ammonium Sulphate	$(NH4)_2SO_4$	10
Potassium Phosphate Dibasic	K_2HPO_4	73
Sodium Phosphate Monobasic Dihydrate	$NaH_2PO_4 .2H_2O$	18
Ammonium Citrate Dibasic	$HOC(CO_2H)(CH_2CO_2NH_4)_2$	2.5

3.2.4 Glucose determination

For determination of glucose concentration in the media, an YSI 2900 Biochemistry Analyzer (Xylem Inc., Ohio, USA) was used (Figure 3.2). This system uses enzymatic biosensors immobilized in an enzymatic membrane, which are specific for different analyte classes. Glucose oxidase is one of the biosensors available which oxidises glucose to hydrogen peroxide. The resulting hydrogen peroxide is further oxidized at a platinum electrode surface polarized. The created electron flow produces a current that is proportional to the glucose concentration.

This system can determine glucose up to 25 g/L, depending on the sample volume with a precision of 2% (CV,n=10). It is also possible to quantify L-lactic acid up to 2.67 g/L with a CV(n=10) of 2%.

Figure 3.2: YSI 2900 Biochemistry Analyzer.

3.2.5 Near-Infrared Spectroscopy System

Two different NIR instruments were used. An at-line spectrometer was used to obtain NIR spectra of the prepared mixtures and an in-line spectrometer was used to collect spectra in the bioreactor, during the fermentation process, and of the mixtures that were prepared inside the bioreactor.

3.2.5.1 At-line instrument

A 6500 NIR spectrometer (FOSS-NIRSystems Inc., Maryland, USA) was used to capture spectra in the transmittance mode with a cuvette of 0.5 mm pathlength by using the sample transport module. An average of 32 diffuse reflectance spectra referenced with 32 co-added scans of air were considered. Spectra were collected at the wavelengths of 1100 to 2498 nm with 2 nm resolution (700 wavelengths in total). Five final spectra (each resulting from 32 scans) for each loaded sample were collected each time. Laboratory temperatures were kept between 20 and 25°C and relative humidity was always between 45 and 65%.

3.2.5.2 In-line instrument

For spectra collected in-line in the fermentor, a stainless steel Interactance Immersion probe, connected to a XDS Process Analytics Microbundle Multiplexer Instrument

Figure 3.3: At-line instrument used: 6500 NIR (FOSS NIRSystems, Inc., USA) with sample transport module.

(FOSS NIRSystems, Inc., Maryland, USA) was used. This type of probe is suitable for transmission-style measurements as the beam is directed through the sample to a shielded mirror, that the beam could be reflected (back through the sample again) to the receiving channel. An optical slit of 0.5 mm was used, which results in an effective pathlength of 1 mm. Spectra were collected from 800 to 2200 nm with 0.5 nm resolution (2800 wavelengths in total). Each collected spectrum is the result of 32 co-added scans which results in a measurement frequency of 19sec per spectrum.

Prior to collecting the spectrum, a NIST (United States National Institute of Standards and Technology) traceable reference material (serial number R99P0079) is used for reference, while using a reflectance probe. A reference procedure was carried out and subsequently a correction factor was applied to compensate for the differences between the probes. The in-line collection was then performed with the immersion probe.

3.2.5.3 Data analysis/model development

All infrared spectra acquired were exported from the proprietary software, Vision (version 4.0.1.0, Foss NIRSystems), into Matlab application [R2012b] (MathWorks, USA) and SIMCA 13.0 (Umetrics, Sweden) for manipulation and modelling of data. Parameters determined by SIMCA were set at a confidence level of 95% and p-value of

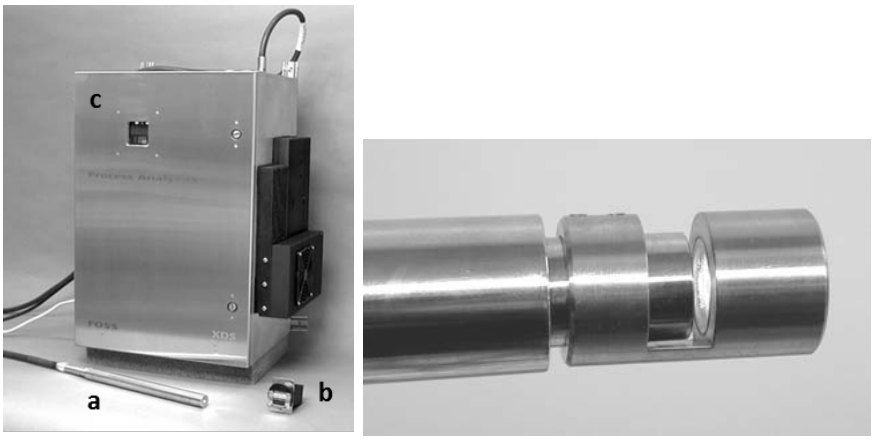

Figure 3.4: In-line instrument used: XDS Process Analytics Microbundle Multiplexer Instrument (FOSS NIRSystems, Inc., USA): reflectance probe (a), reference case (b), spectrometer (c). Right hand side shows the Interactance Immersion probe.

0.05.

3.2.6 Experimental design

To aptly identify and measure various sources of variability in the spectra of a fermentation matrix, this chapter follows a bottom up approach. Starting with water, four additional backgrounds were used, i.e. buffer, fresh media, spent media and one that consisted of a mixture of the analytes (glucose and lactate) spiked with a solution of washed cells in water. This approach allowed for a step-wise monitoring of various sources of variability in the spectra. A two-level design with two factors was used to have a data of each analyte against each solvent and mixtures of both analytes in water. Low and high levels of the solvent were also prepared with exception of the analytes in water, for which three levels were used. A three-level design considered two factors at low, intermediate and high levels, which facilitates investigation of a quadratic relationship between the response and the respective factors. However, this approach required a large number of runs. Therefore, the middle level was not used. The prepared samples, shown in Table 3.6, were scanned in the at-line NIR instrument.

Table 3.6: Experimental design

Factor 1	Factor 2	N samples
Glucose	x	3
Lactate	x	3
Glucose	Lactate	4
Buffer	x	2
Buffer	Lactate	4
Buffer	Glucose	4
Fresh Media	x	2
Fresh Media	Glucose	4
Fresh Media	Lactate	4
Spent media	x	2
Cell content	x	2

3.3 Results and Discussion

This section develops a universal model that can predict concentrations of analytes across different backgrounds. The flow of a typical chemometric methodology, as outlined in section 2.4, provides a useful structure to frame the discussion. The collected samples, appropriate wavelength and pre-processing techniques were discussed first; prior to examining the fermentation process. Then, the model development follows the bottom-up approach by using solutions of increasing complexity. From there, a universal model was tested. While the above description follows an at-line approach, an in-line approach was also utilised. The bioreactor was employed with various media components while different supplementary solutions were fed into the vessel, while in-line NIR measurements were collected. Glucose and lactate were the studied analytes in this bottom-up approach as they are relevant in different bioprocess settings: glucose is the main source of carbon for many expression systems and lactate is a common byproduct that can become toxic to cells and inhibit production of the protein of interest. However, this approach is a proof-of concept and it could be applied with other different analytes.

3.3.1 Fermentation of *E. coli* and Sample Selection

Figure 3.5 shows the results of the fermentation of *E. coli*. The main goal was to obtain spent media and cellsto support the experimental design on different matrices. The top plot shows the growth curve based on absorbance readings at 600nm, while the bottom plot shows the glucose concentration in the supernatant, measured with the YSI system.

Sample s4 was selected for further processing, since it showed good levels of biomass (abs_{600nm} approximately 5units) and the YSI reading of glucose concentration was 6.61±0.16 g/L. This value was then accounted for when developing the models.

During the course of the fermentation process the same procedure was followed for all the collected samples:

- To have enough volume for all mixtures, a 45ml sample aliquot was collected per

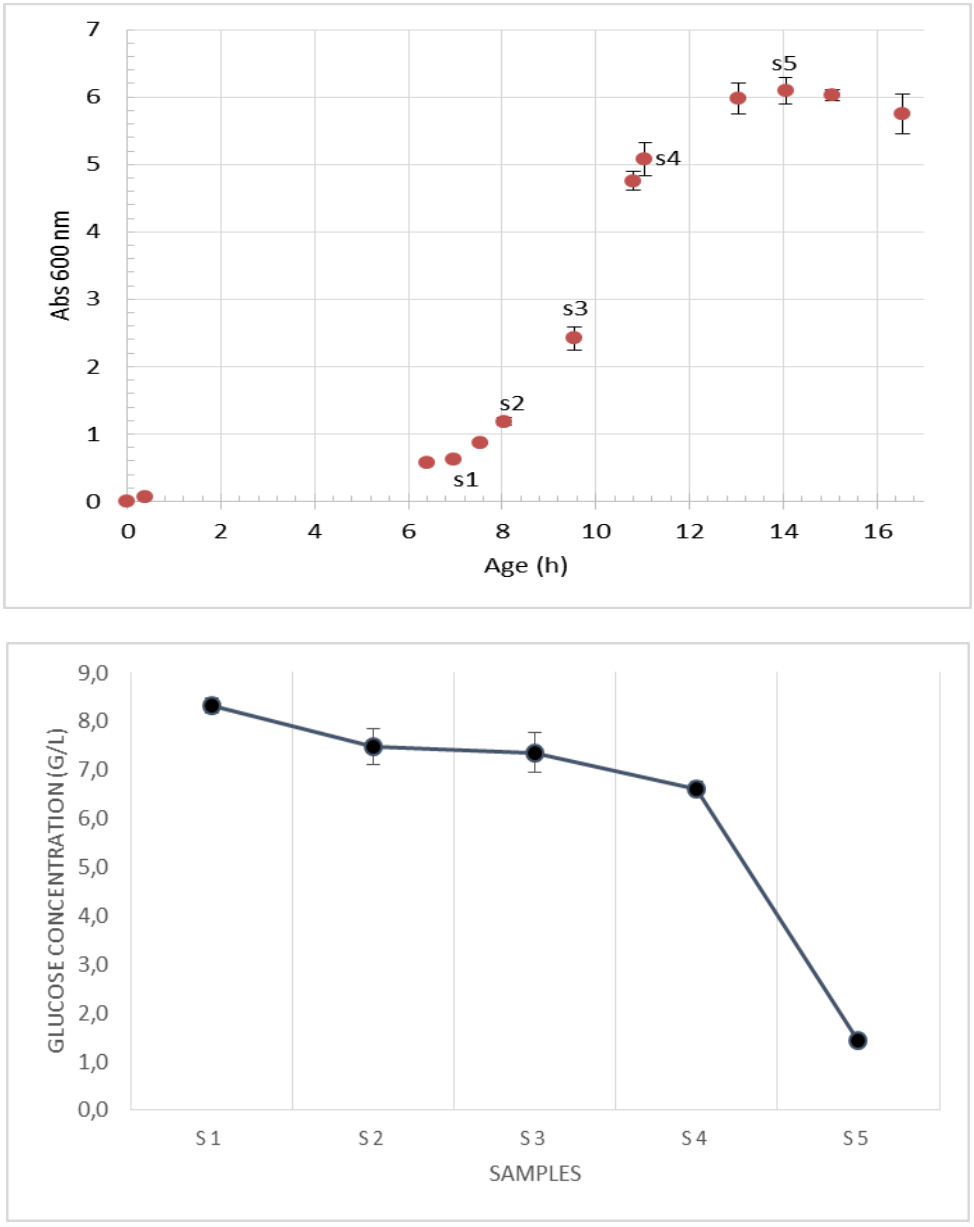

Figure 3.5: Absorbance measurements at 600 nm obtained on samples from the fermentation of *E. coli* (top plot). Five of these samples were selected for glucose measurements (g/L) using the YSI analyser (marked s1 to s5) and the results are plotted for each sample (bottom plot). Finally, only sample s4 was selected for further processing since it had interesting amounts of both biomass and glucose in its medium.

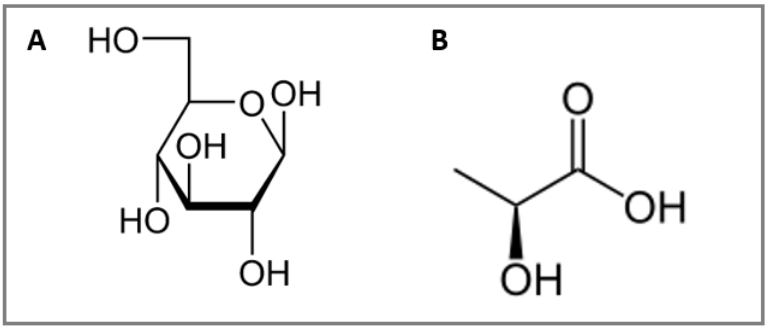

Figure 3.6: Structures of Glucose (A) and Lactate (B).

falcon tube. Five falcon tube samples were used.

- The absorbance was read at 600 nm
- The cells were spun down at 9000rpm and 4°C for 5min
- The supernatant/spent media of all five tubes were collected and immediately frozen
- The cells were washed with water and spun down twice (water/spin down/water/spin down)
- A suspension in water was prepared for a given absorbance and used as stock solution to spike the mixtures
- The final absorbance of the final mixture was read at 600nm
- Cells suspension and supernatants were stored immediately at -20°C.

3.3.2 Wavelength Selection

After the raw NIR data were collected, the second step involves a chemometric process to select the appropriate wavelength and preprocessing technique (Figure 2.4). Wavelength selection for further model development was based on *a-priori* knowledge about the structure of the molecules.

As mentioned before, NIR bands correspond to specific vibrations of chemical bonds (shown in Figure 3.7) and C-H vibrations are mainly captured in the combination band region (2150 to 2450nm) and in the first overtone region (1620 to 1780nm). Spectra of the molecules in these two regions are plotted in Figure 3.8 and Figure 3.9, respectively. There are stronger bands in the combination region and weaker bands in the first

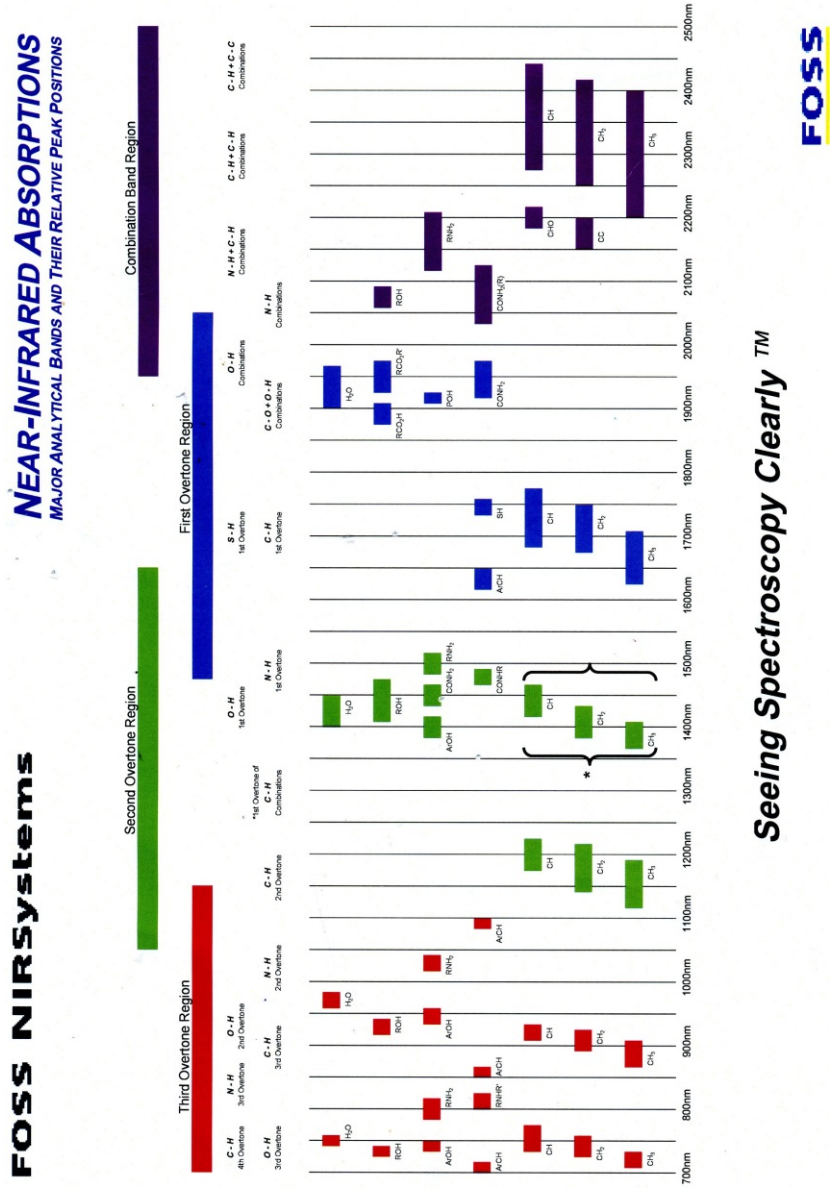

Figure 3.7: Major analytical bands and their relative positions (supplied by FOSS NIRSystems).

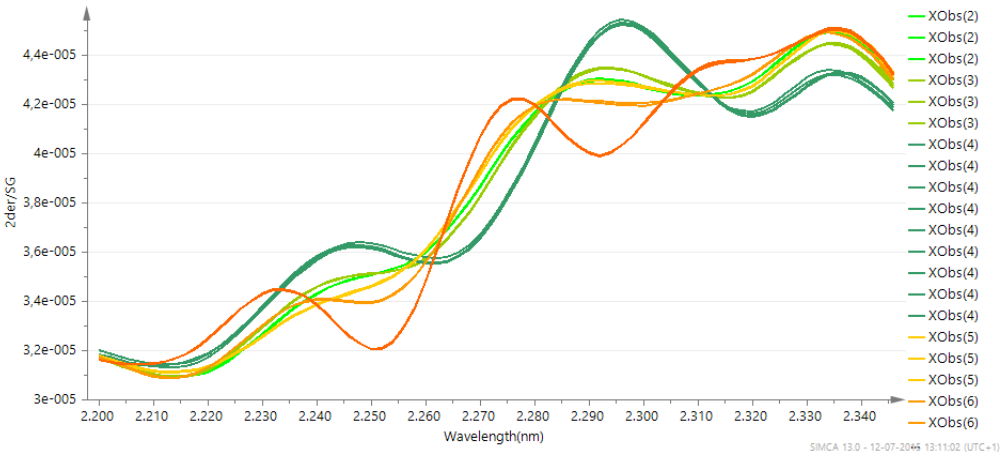

Figure 3.8: Combination overtone region: stronger bands were detected. Spectra of glucose solutions (1.25, 2.5 and 10g/L) correspond to yellow, orange and red lines, respectively. Spectra of lactate solutions (0.58, 2.5 and 10g/L) are plotted in lime, light green and dark green, respectively.

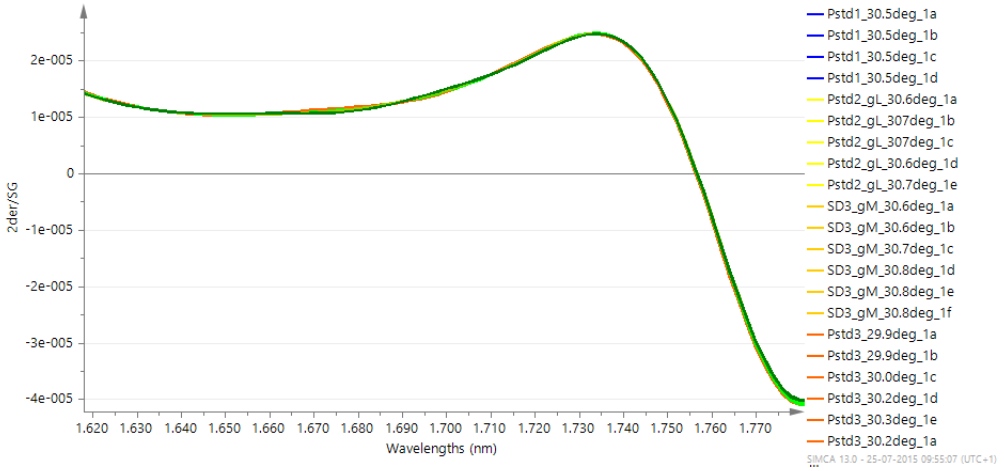

Figure 3.9: First overtone spectra region: weaker signals were detected. Spectra of glucose solutions (1.25, 2.5 and 10g/L) correspond to yellow, orange and red lines, respectively. Spectra of lactate solutions (0.58, 2.5 and 10g/L) are plotted in lime, light green and dark green, respectively.

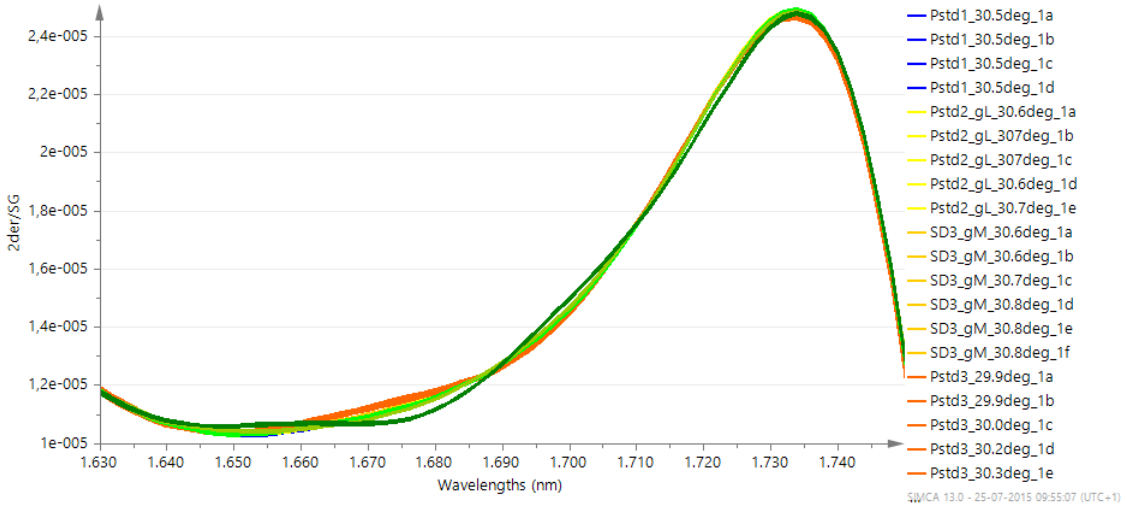

Figure 3.10: Spectra of pure analytes in water for the selected wavelength region. Concentrations of glucose are represented in grades of orange and concentrations of lactate in grades of green.

overtone region.

Even though the combination band region would be very useful on the quantification of glucose and lactate, the end goal of this study was to implement an in-line probe in a fermentation vessel to obtain real-time information and ultimately be able to control for glucose levels in the medium and neutralise the lactate above toxic levels. The wavelengths above 2000nm were not considered. This spectral region is not usable when using optical fibres because silica interfere, resulting in a loss of signal (Yu et al., 2012).

Through visual comparison of the various spectra from different concentrations of the analytes (Figure 3.10), together with attempts of using different ranges (Table 3.7), wavelengths 1630-1750 nm were selected.

Table 3.7: Results from PLS models with two latent variables, for the two analytes. The errors selected models are shown in bold.

Root Mean Squared Errors (g/L)	Glucose 1620-1780 nm	Glucose 1630-1750 nm	Lactate 1620-1780 nm	Lactate 1630-1750 nm
RMSEE	0.707	**0.477**	0.590	**0.561**
RMSEcv	0.965	**0.685**	0.610	**0.616**

In addition to selecting the optimum wavelength, two common preprocessing techniques were also investigated: first derivative and second derivative. The results of using these techniques were compared with the use of raw spectra in the selected region of 1630 to 1750 nm (Table 3.8).

The choice of the appropriate pre-processing technique is highly relevant. Modelling based on raw spectra facilitated baseline shifts definition. Derivatives are most useful when there are several closely overlapping peaks and were used to improve resolution. The apparent peak width could be reduced, while the noise is also substantially amplified (Brereton, 2003). After an initial investigation of the raw data, first and second derivative pre-processing techniques were also applied. For every tested PLS model, the data was mean-centred.

Table 3.8: Results from PLS models with two latent variables for each preprocessing technique applied to each analyte on the wavelengths 1630-1750 nm. Smaller errors resulting from using a 2nd derivative are highlighted.

Root Mean Squared Errors (g/L)	Glucose Raw	Glucose 1st Derivative	Glucose 2nd Derivative	Lactate Raw	Lactate 1st Derivative	Lactate 2nd Derivative
RMSEE	3.43	0.741	**0.477**	3.48	0.799	**0.561**
RMSEcv	3.39	0.815	**0.685**	3.34	1.01	**0.616**

3.3.3 Model Development

It is difficult to obtain relevant information about complex matrices using NIR. This technique has a high detection limit due to the complex nature of spectra that includes overtones and combination bands from various vibrational energy levels (Luypaert et al., 2007). In a complex matrix all these transformations result in a broad range of overlapping absorption bands, which makes the extraction of accurate analyte specific information a real challenge. Extra difficulties arise when the analytes of interest are contained in mixtures with similar chemical species present at similar concentrations (Rhiel, Cohen, Murhammer and Arnold, 2002).

3.3.3.1 PCA Models

PCA models allow the dissociation of spectral features that are not easily detected by the eye. The aim of this section is to detect differences between matrices by using the scores plots and loadings plots of these developed models. Since PCA does not correlate spectra with a Y-variable, it highlights variability in the spectral, free from forced correlations.

The loadings plot of the principal components describes correlations between variables and the components. A small (positive or negative) values indicate a weak relationship between the variables and the components. Through comparison of the loadings plot with the scores plot, it is possible to identify spectral regions that are responsible for the discrimination of respective elements of the matrix through the corresponding position on their scores plot. With this method, it was investigated if the PCs can deconvolute the effect of the different matrices in the analytes' spectra, which would allow these differences to be accounted for, further on in the model development. This section focuses on samples containing lactate to investigate the effect of the different matrices backgrounds on its spectra.

3.3.3.1.1 Buffer

A PCA model using the samples of two levels of lactate (labelled with numbers -1 and 1) in two levels of buffer (colours) was developed. The scores plot is shown in Figure 3.11.

As it is part of the definition of PCA model's development, the first principal component (x axis) captures most variance of data and in this case, it is possible to hypothesise that it is related to lactate content. While the second component only explains 4.9% of the data variability, it seems to separate buffer content in the samples. For validation, the loadings plot obtained from the PCA were compared to the loadings plot of the analyte in water (Figure 3.12), which proves that the first PC is describing lactate variation. The second PC describes baseline variation which can be attributed to the presence of buffer. However, there are several sources of variability that also

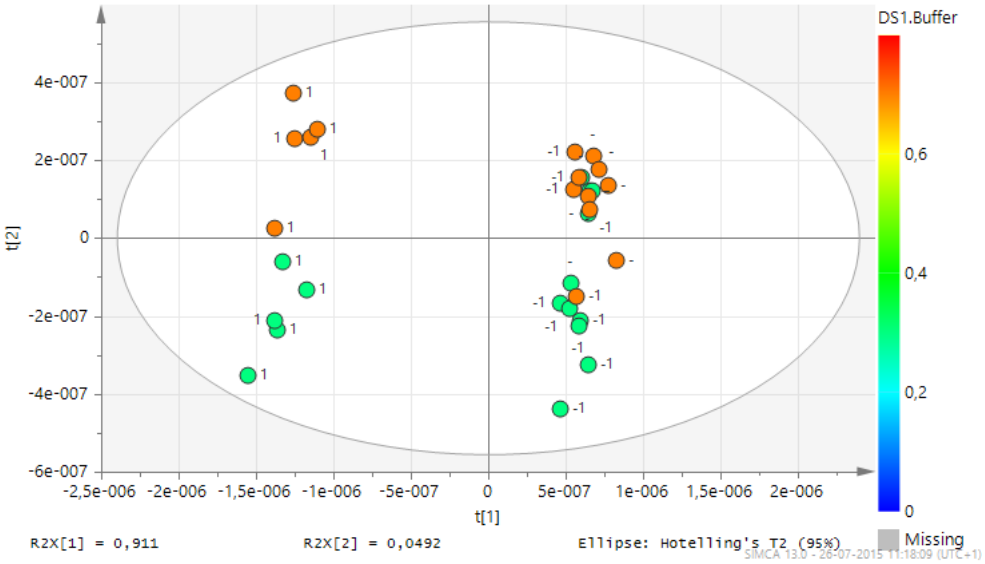

Figure 3.11: Scores plot of PCA model for lactate samples in two different percentages of buffer: in green 30% and in orange 70%. Numbers represent the level of lactate: low (-1), high (1) or no lactate (-).

result in baseline effects, such as temperature.

Figure 3.13 provides a visual comparison of the spectrum and its resulting second derivative. It shows that any extremities in the raw spectrum are extenuated in the second derivative, which translates into higher signal-to-noise ratio.

The derivatised spectra were also visually checked to understand if it is possible to detect the differences within buffer percentages. Samples with high lactate content are shown in Figure 3.14 and low lactate content (lower peaks) in Figure 3.15.

3.3.3.1.2 Fresh media

The same approach was used for fresh media samples as for buffer. The scores plot (Figure 3.16) reveals the expected clustering on lactate content and media content. Fresh media constitutes a more complex matrix than buffer, with each of the components having their respective NIR spectra. However, the selected wavelengths for this study should only be sensitive to C-H and S-H vibrations. In these conditions, signals mainly from glucose (at 11.9 g/L in the fresh media) could be observed but residual amounts of EDTA (Table 3.4) from the trace elements solution were detected as well.

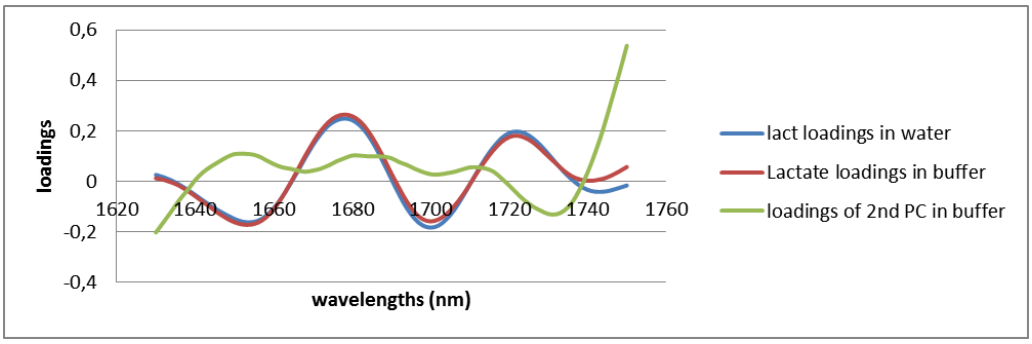

Figure 3.12: Comparison of loadings plots of the occurrence of lactate in water (blue) and buffer from the PCA model of which the scores plot is shown in Figure 3.11 (first PC in red, second PC in green).

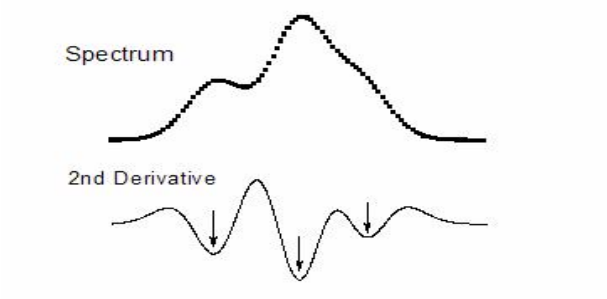

Figure 3.13: Visual comparison between the spectrum and its resulting second derivative. Higher peaks in raw spectra result in deeper troughs on its second derivative form with a consequent lower shoulder after the peak.

The second principal component (2PC, t[2], on the y axis) revealed a correlation with lactate content while a comparison of the loadings plot shown in Figure 3.17 prove that relation. The loadings of the second PC match the loadings of a PCA model based only on solutions of lactate in water. The first principal component (1PC, t[1], x axis) correlates with media percentage, decreasing along the x axis. However, this percentage decrease of media also implies a decreasing concentration of glucose. The trend variant plot of glucose in water (Figure 3.18) show that it is likely that there is a strong effect from glucose, but it is likely that there are other components in the media, causing a matrix effect.

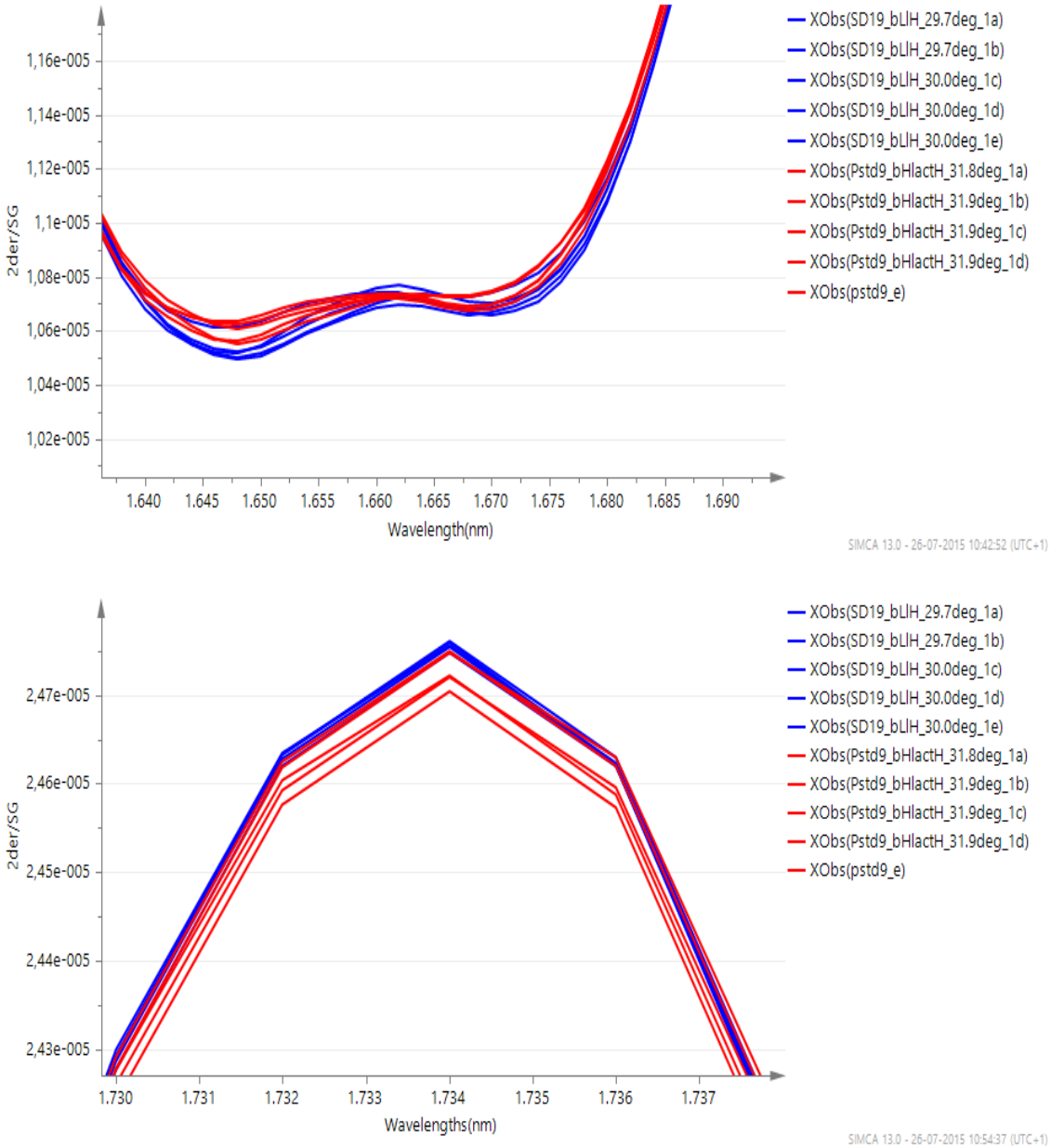

Figure 3.14: Spectra (with second derivative applied to it) of solutions with high lactate content in low percentage of buffer (in blue) and high percentage of buffer (in red), focusing on different wavelengths.

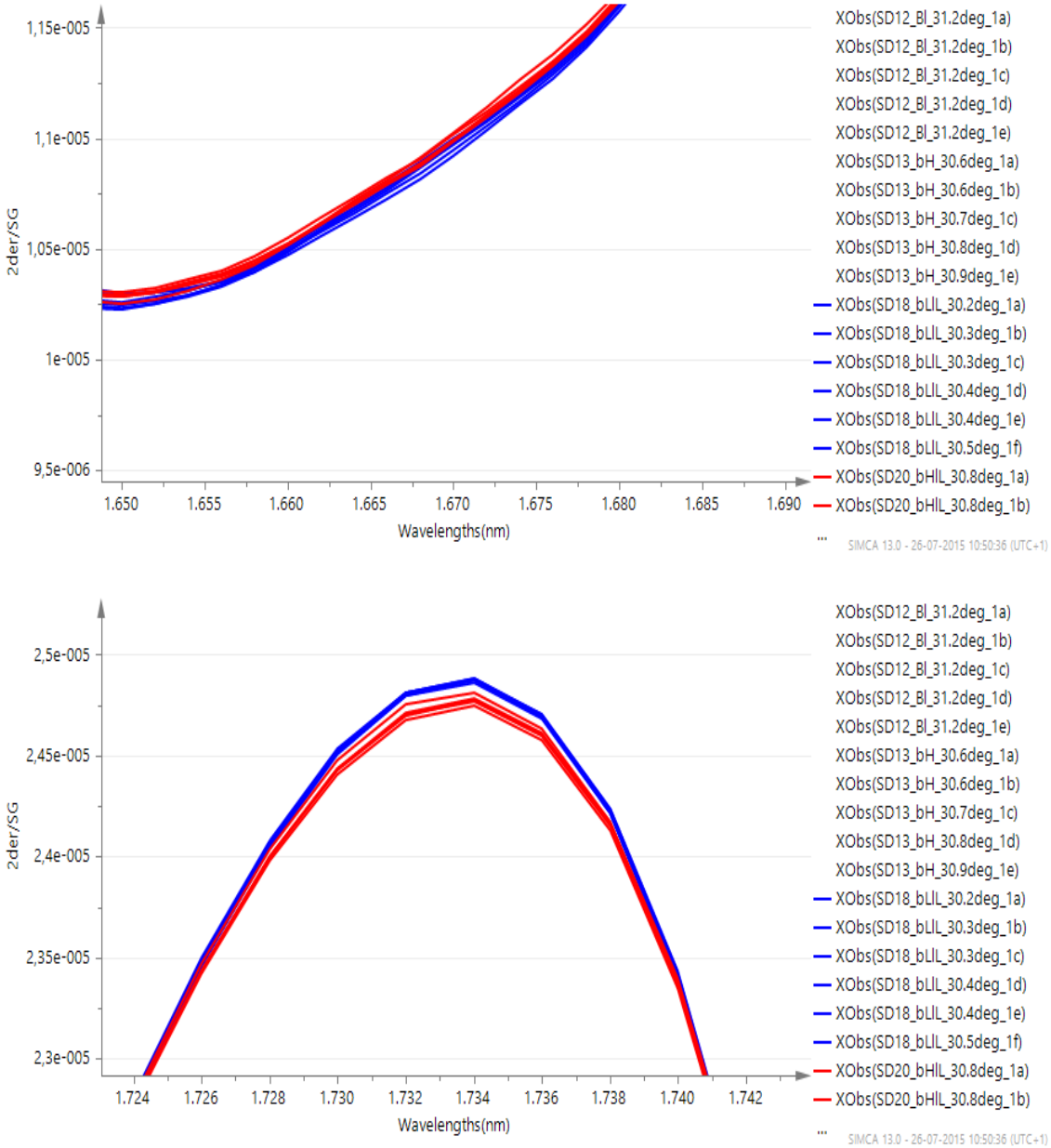

Figure 3.15: Spectra (pretreated with second derivative) of solutions with low lactate content in low percentage of buffer (in blue) and high percentage of buffer (in red), focusing on different wavelengths.

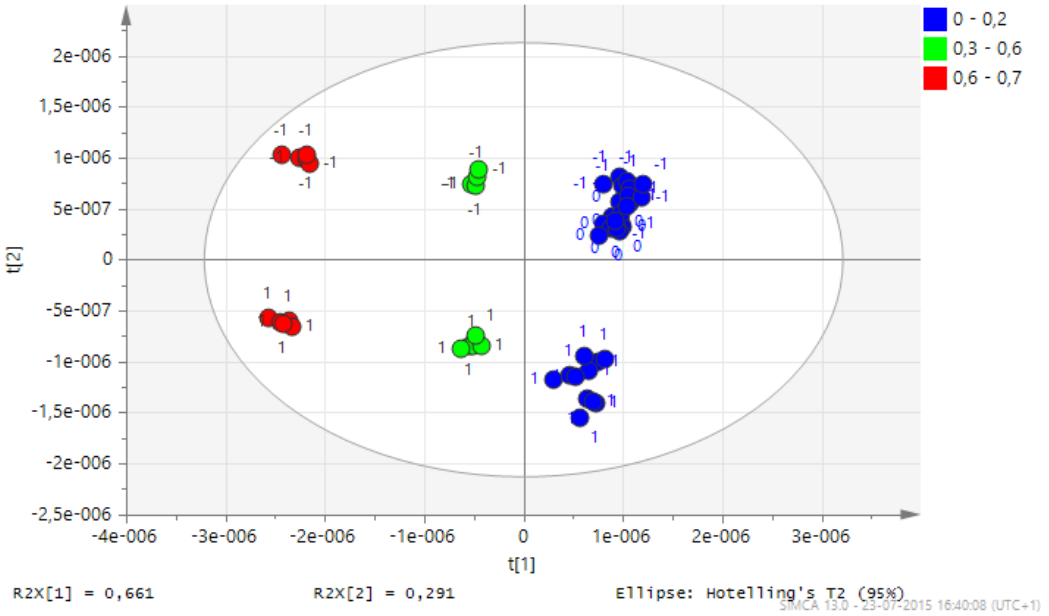

Figure 3.16: Scores plot of a PCA model developed from a dataset of low (-1), medium (0) and high (1) levels of lactate in water (blue scores) and in fresh media containing 11.9 g/L of glucose (30% in green and 70% in red).

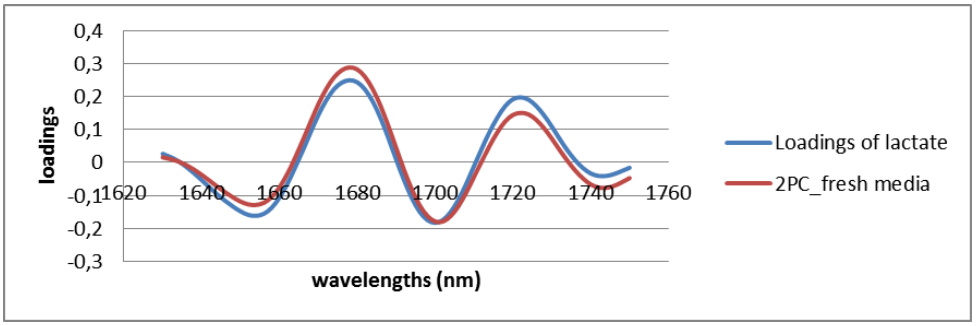

Figure 3.17: Comparison between the loadings of the second principal components (in red) with loadings from lactate in water (blue).

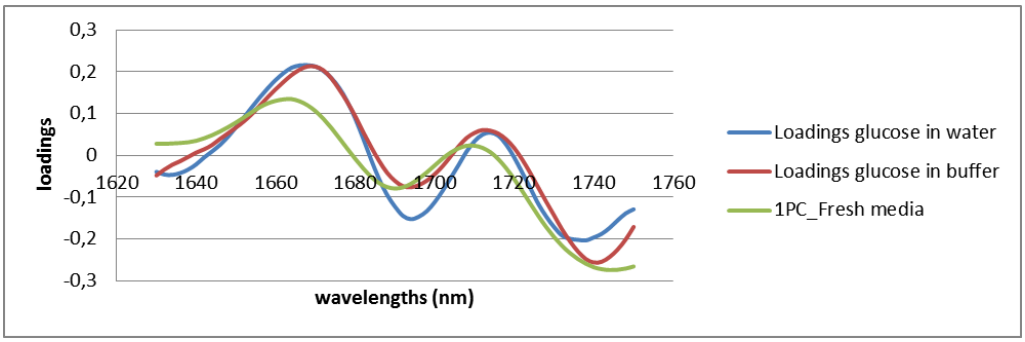

Figure 3.18: Comparison between the loadings of the first principal components (in red) with loadings from glucose in water (blue).

3.3.3.1.3 Spent Media

Spent media (or supernatant) should have a similar composition to the fresh media, without the analytes that were already consumed at this point of the fermentation run, and with the secondary metabolites that might have been produced by the cells, or for example some cell debris. The scores plot of this PCA model is shown in Figure 3.19.

A comparison of the loadings (Figure 3.20) of lactate in various backgrounds showed a correlation of loadings from lactate in water with the second principal component (2PC), as indicated by the scores plot in Figure 3.19.

From comparison of the loadings plot (Figure 3.21), the first principal component is more related to the loadings of fresh media than to glucose in water. The shift relatively to the fresh media is likely to be due to the formation of other metabolites during the course of the fermentation, e.g. acetic acid, which would vibrate on C-H region as well. Another possibility would be that of the presence of cell debris – it is possible that cell debris did not get spun down in the centrifugation step and might have influence the collected spectra.

3.3.3.1.4 Presence of Cells

As discussed before in the Introduction (Chapter 1), the presence of cells increases the turbidity of the sample, which increases the scattered light and consequently reduces the amount of light reaching the detector (Henriques et al., 2009). The resultant

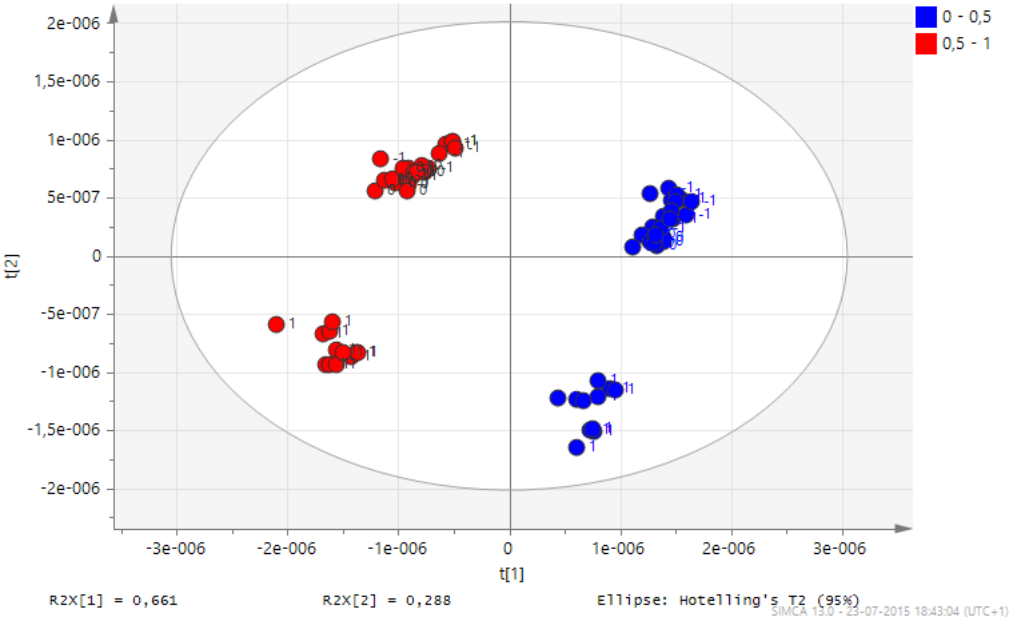

Figure 3.19: Scores plot of a PCA model developed from a dataset of low (-1), medium (0) and high (1) levels of lactate in water (blue scores) and in spent media (30% in green and 70% in red).

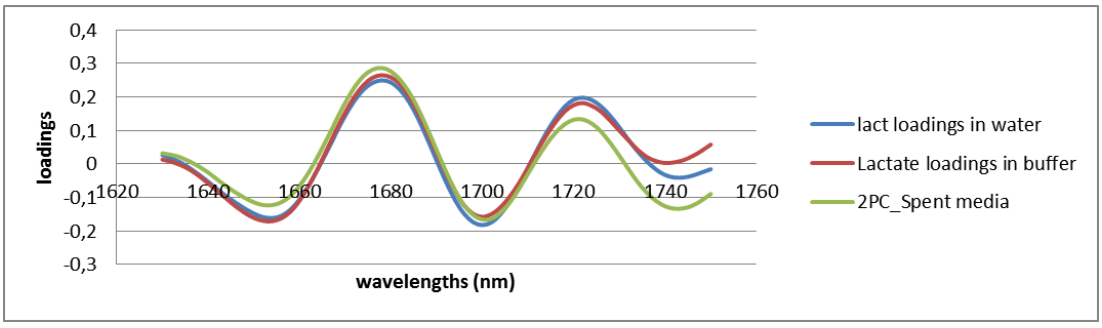

Figure 3.20: Comparison between the loadings of the second principal component (in green) with loadings from lactate in water (blue) and in buffer (red).

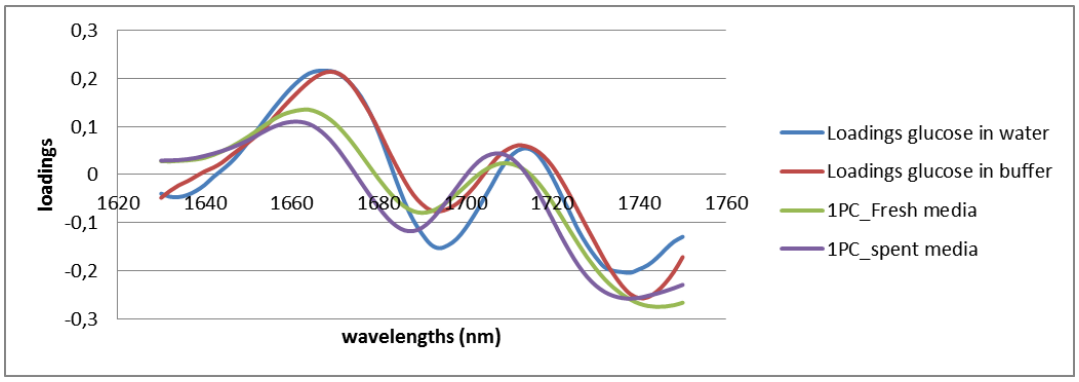

Figure 3.21: Comparison between the loadings of the first principal component of the PCA on spent media (in purple) with loadings from a PCA of glucose in water (blue), a PCA of glucose in buffer (red) and the first PC of the PCA in fresh media (green).

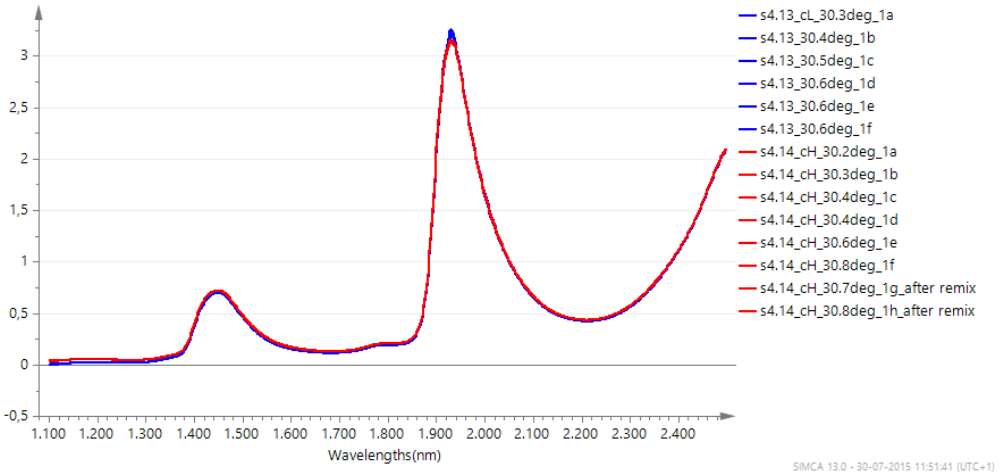

Figure 3.22: Raw spectra of lower cell content in water (blue samples) and higher cell content (red samples).

notable feature on the raw spectra is a baseline shift. Applying a first derivative as pre-processing technique removes most of physical effects, enhancing chemical information for the modelling steps (Henriques et al., 2009).

Figure 3.22 shows the raw spectra of samples with low cell content (in blue) are compared with spectra from high cell content samples (red). There is a clear baseline shift up due to the presence of cells but also a lower peak at 1932nm due to less water content in the sample with more cells, resulting in less OH bonds vibrations.

The second derivative of the same spectra (Figure 3.23) reduced the baseline shift observed in the raw spectra but it is still possible to find differences between the same

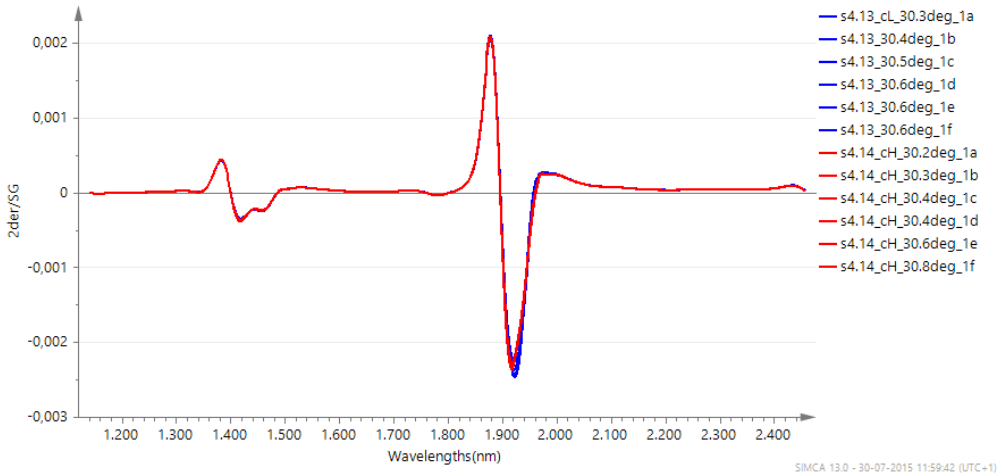

Figure 3.23: Spectra (with second derivative and smoothing filter applied) from samples with low cell content (blue) and high cell content (red).

concentration of the analyte in water with low content of cells (blue) or high (red), which proves that chemical information is being kept.

The scores plot of a PCA model developed for these samples is presented in Figure 3.24 and the loadings in Figures 3.25 and 3.26.

Applying an MSC made it possible to reduce the effect of the scatter produced by the cells. The result is shown in the scores plot presented in Figure 3.27. The second principal component indicated the concentration of lactate. After the MSC filter was applied to the second derivative data, the scores were no longer segregated by the number of cells in the samples as previously presented but the second principal component was instead related to the lactate concentration. From the comparison of loadings plot presented in Figure 3.28, there is now a similar behaviour of the PCs to the models of lactate in water or in buffer.This PCA model was able to predict other scores independent of the cell content. As the MSC corrected the effect of cell concentration, the effect was clearly indicated within the information shown by the comparison of loadings in Figure 3.28.

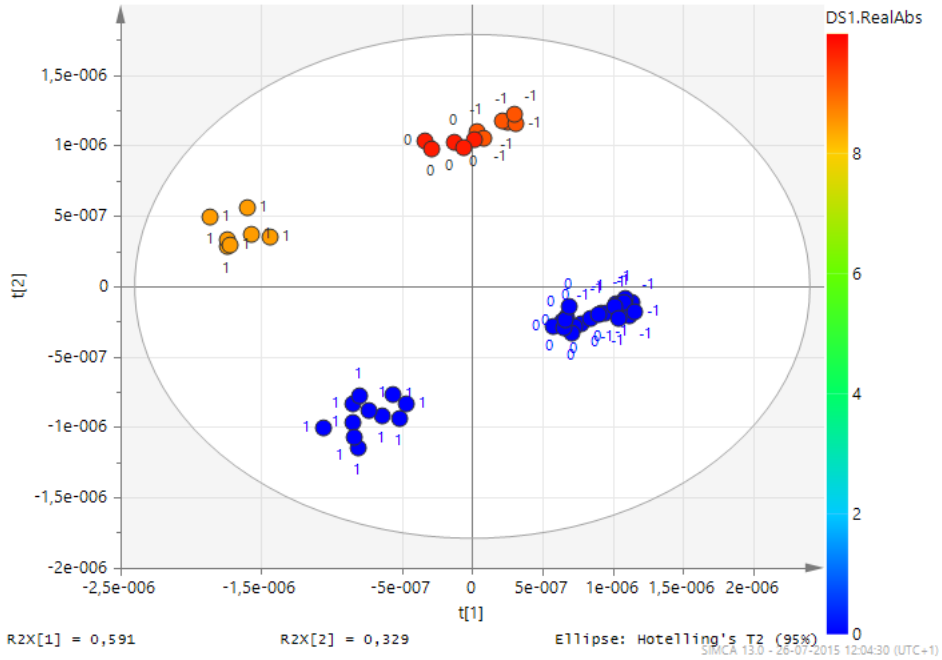

Figure 3.24: Scores plot of a PCA model developed from a dataset of low (-1), medium (0) and high (1) levels of lactate in water (blue scores) and in the presence of cells (0 - blue, 2.5 - orange, 7.0 - red colours). The grade of "real absorbance" was the reading at 600nm after the mixtures were prepared.

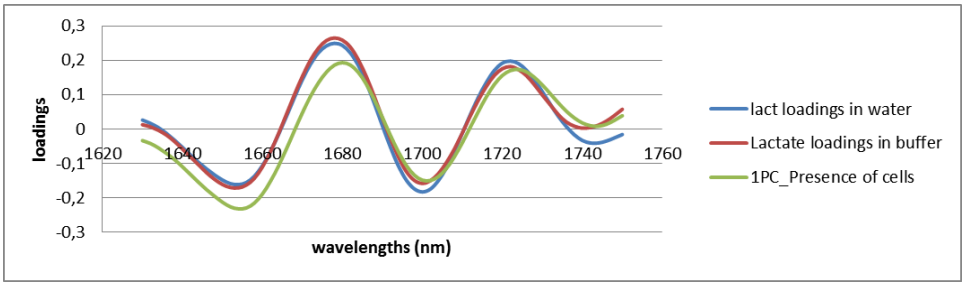

Figure 3.25: Comparison between the loadings of the first principal component of the PCA of mixtures of lactate in the presence of cells (in green) with loadings from a PCA of glucose in water (blue), a PCA of glucose in buffer (red) and the first PC of the PCA in fresh media (green).

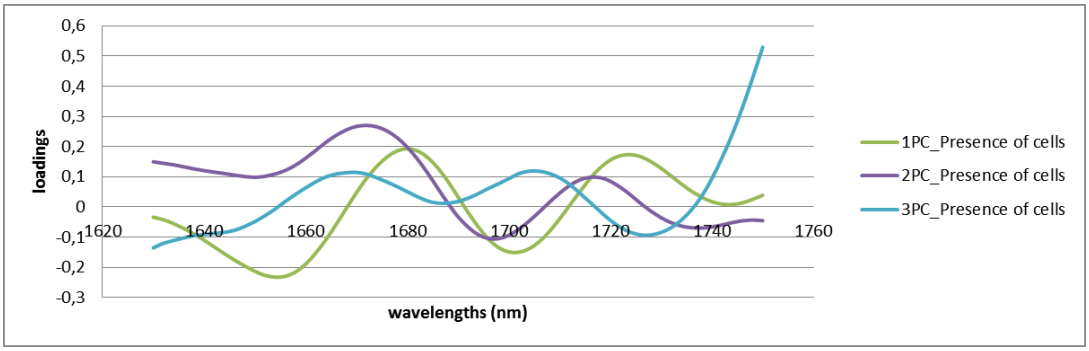

Figure 3.26: Comparison of the loadings of the three principal components of the PCA of mixtures of lactate in the presence of cells: first (in green), the second (in purple) and the third (in blue).

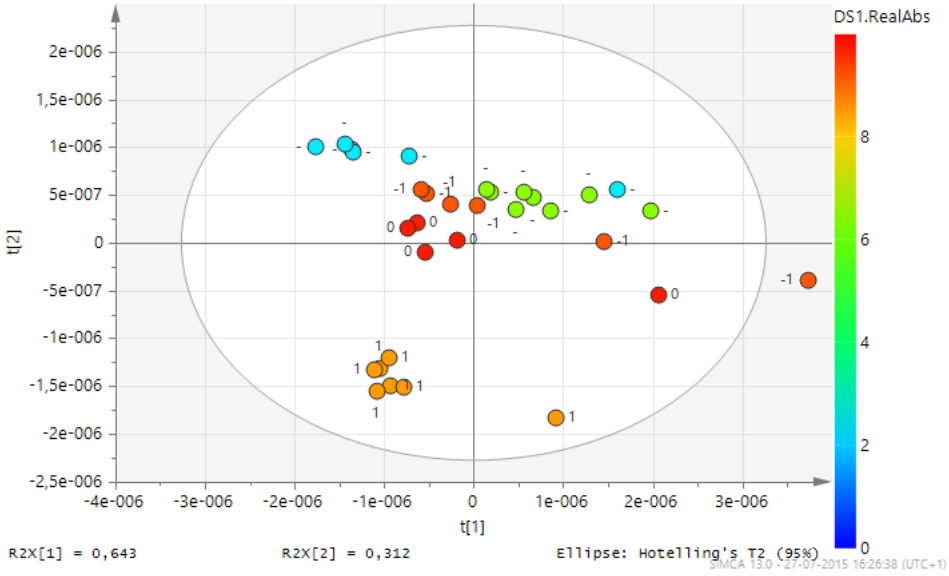

Figure 3.27: Scores plot of a model to which an MSC correction was applied on the data. The colours of the scores are equivalent to the absorbance reading at 600 nm following the scale as shown from 0 (blue) to 10(red).

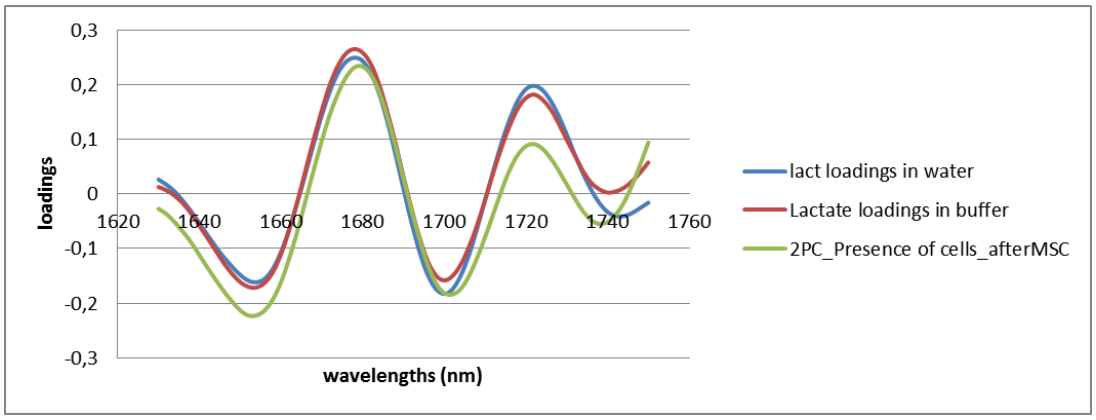

Figure 3.28: Comparison of the loadings of the second principal component of the PCA shown in Figure 3.27 with the loadings of lactate in water (blue) or in buffer (red).

Table 3.9: Summary of the PLS models developed for Lactate concentration in the different tested backgrounds.

Background	Range (g/L)	R^2 (Obs vs Pred)	RMSEE (g/L)	RMSEcv (g/L)	Residuals R^2
Water	0.00/ 0.58 – 10.0	0.981	0.595	0.615	0.984
Buffer	0.00/ 0.58 – 10.0	0.991	0.381	0.410	0.974
Low %Fresh media	0.00/ 0.58 – 10.0	0.969	0.693	0.695	0.957
High %Fresh media	0.00/ 0.58 – 10.0	0.982	0.532	0.543	0.976
Spent Media	0.00/ 0.58 – 10.0	0.960	0.826	0.839	0.992
Cells	0.00/ 0.58 – 10.0	0.976	0.741	0.734	0.990

3.3.3.2 PLS Models

Individual calibration models for each of the backgrounds were developed. For each model only pure solutions of the respective analytes to be quantified were included. The results for lactate are presented in Table 3.9 and for glucose in Table 3.10.

The residuals R squared results for a fitted line to the Normal Probability plot, which displayed the standardised residual on a double Log scale, were generated by SIMCA software. This is the raw residual divided by the residual standard deviation (RSD). Therefore, if the residuals are random and normally distributed, the normal probability plot of the residuals has all points lying on a straight line and is the R squared of a fitted line closer to the unit.

The measurement errors for lactic acid concentration in buffer were ±0.4 g/L or

Table 3.10: Summary of the PLS models developed for Glcose concentration in the different tested backgrounds.

Background	Range (g/L)	R^2 (Obs vs Pred)	RMSEE (g/L)	RMSEcv (g/L)	Residuals R^2
Water	0.00/ 1.25 – 10.0	0.965	0.806	0.937	0.977
Buffer	0.00/ 1.25 – 10.0	0.982	0.609	0.612	0.977
Low %Fresh media	0.00/ 3.57-13.8	0.881	1.47	1.46	0.847
High %Fresh media	0.00/ 8.33-18.3	0.956	0.735	0.892	0.985
Spent Media	0.00/ 1.98-14.6	0.849	1.79	1.83	0.932
Cells	0.00/ 1.25-10.0	0.649	2.47	2.45	0.923

4.44 mM, which agreed with millimolar results expected for an NIR spectral dataset (Rhiel et al., 2002). This error is however lower than in water. Spectra obtained for analytes in water revealed more variability than in samples with a certain percentage of buffer in the mixture. These errors are higher than those presented by Rhiel et al. (2002) but the models used here were simpler than those from the publication, as they were built with only two latent variables to avoid overfitting and without using any variable selection. The dataset built on spent media revealed the highest errors. This result agreed with the indicated by the PCA models. It was expected that this would represent the most difficult matrix amongst those tested in this study. As earlier pointed out, spent media contained high levels of glucose, which could confound the effect of lactate in the spectra due to some cell debris that might have been left in the solution after centrifugation while, the factor with greater impact could be due to the formation of other primary metabolites during the course of the fermentation. *E.coli* produces acetic acid ($C_2H_4O_2$) which will also have vibrational bands on the selected wavelengths and would be overlapping with lactate bands. The dataset for the analytes in presence of cells afforded similar error values between fresh and spent media, suggesting that light being scattered by the cells did not produce greater effect than the other matrices. As for the presence of lactate, the buffer showed a positive effect on the prediction of glucose concentration, and gave smaller errors of ± 0.6 g/L or 4.99 mM in comparison with samples in water. The accuracy of estimations diminishes with the presence of more complex compounds in the matrix.

Residuals and 'observed vs. predicted' plots indicate that there is no outliers in-

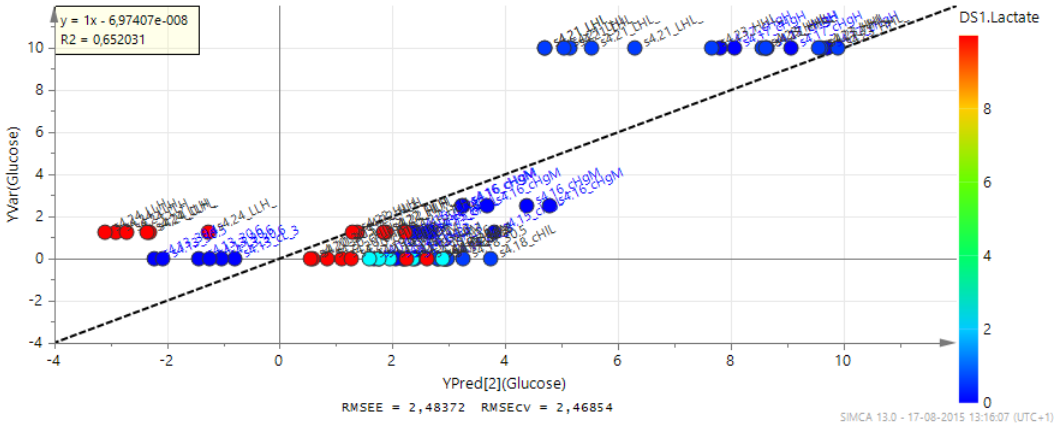

Figure 3.29: Predicted glucose concentrations (x axis) against observed values (y axis). The points are coloured by increasing concentration of lactate present in the mixture (gradient colour label, from 0 g/L in dark blue, to 10g/L in red). The labels of each point are the name of sample s4 plus the specific treatment.

cluded in the model, with the exception of the model in cells, which requires further work. From Figure 3.29, the mixtures containing a lower percentage of cells are being underpredicted. A first approach would be to develop different models for each content level of cells.

This section shows that it is possible to extract accurate analyte specific information in each of the tested datasets. Lactate estimations revealed always millimolar level accuracy. Errors were higher when spent media was present in the mixtures, resulting in errors of 9.3 mM. From the tested backgrounds, cells had a significant effect on the prediction of glucose, most likely due to light scattering by the cells.

It was assumed that the errors obtained for the analytes in water and buffer represent the lowest possible for the system used. Further studies will seek to reduce the errors associated with other PLS models to the errors levels obtained in the simple water/buffer scenario.

The used method is the one currently used which requires extensive calibration procedure for each matrix and conditions. The following chapter focuses on an attempt of developing a universal model that would work across a range of backgrounds.

3.3.4 Universal Model Selection

Ultimately, the goal of this thesis is to apply this knowledge to an automated system, such that the end-user would not have to execute heavy calibration procedures. Instead, a model would be able to predict across a range of different matrices and conditions. The tested hypothesis in this section is that such a model can be built in a simple matrix, as is the case of water or buffer.

As shown in the previous section (3.3.3.2), the PLS models developed for the analytes in the presence of buffer resulted in lower prediction errors. Accordingly, buffer was the selected background to compare with the other matrices. Two PLS models (one for each analyte concentration, Y) were developed.

To quantify the influence of each matrix background on the determination of the concentrations of these two analytes, testing the prediction ability of this model across a different range of conditions, all the other datasets (analytes in water, fresh media, spent media, in the presence of cells) were predicted through the model of samples in buffer. The measurement used for this comparison was the mean squared error (MSE) and the obtained results are summarised in Tables 3.11 through 3.13.

The MSE measures the average of the squared errors, which in turn are the differences between estimated or observed values and predicted values.

$$MSE = \frac{1}{n}\Sigma(Observed.Conc - Predicted.conc)^2 \tag{3.1}$$

An MSE of zero would mean that samples are being predicted with perfect accuracy, which is practically impossible. Two different estimators can be compared using MSE values as the model with lower MSE value explains better the observations dataset, with minimized variance.

The results from fresh media reveal that high levels of lactate can be detected as well as in water in this matrix, using the model in buffer. However, the model fails for lower levels of lactate in media (either fresh media or spend media) which is likely to be due to high levels of glucose present in these matrices. This effect was observed before, by Rhiel et al. (2002), but the authors had only a limited number of low concentrations

Table 3.11: Mean squared errors for predictions of lactate concentrations on each matrix, using PLS model developed on buffer. The results for glucose are only presented in water since the other results revealed to be extremely high and therefore not feasible. The listed errors are only for samples spiked with either glucose or lactate – no mixtures were included.

		Water	**Fresh media**		**Spent media**		**Cells**	
			Low	high	low	high	low	High
Lactate	Low	0.15	3.33	8.93	x	17.86	x	4.49
	Med	0.14	x	x	x	7.29	x	3.89
	High	0.32	0.35	0.29	x	2.48	x	4.21
Glucose	Low	0.76						
	Med	0.70						
	High	0.12						

Table 3.12: MSEs for predictions of lactate concentration in the presence of glucose, in different background matrices.

		Water		Spent media				Cells			
				Low		High		low		high	
		GlcLow	Glc high	GlcLow	Glc high	GlcLow	Glc high	GlcLow	Glc high	GlcLow	Glc high
Lactate	**Low**	0.40	0.64	X	2.07	x	7.01	x	0.112	x	3.87
	High	0.61	1.17	1.43	x	1.48	x	1.58	x	2.33	x

in the dataset. One can argue that the lower level of lactate is likely to be under the detection limit, confounded by the presence of high levels of glucose.

The error of predictions in the presence of cells revealed to be independent of the concentration of lactate which means that a correction might be easier to obtain in this case.

For mixtures of lactate and glucose in different matrices the results of the errors of predictions are shown in Table 3.12 and Table 3.13, respectively.

Table 3.13: MSEs for predictions of glucose concentration in the presence of glucose, in different background matrices.

		Water		Spent media				Cells			
				Low		High		Low		High	
		LactLow	Lact high	LactLow	Lact high	LactLow	Lact high	LactLow	Lact high	LactLow	Lact high
Glucose	**Low**	0.59	1.00	x	18.2	X	81.9	X	2.49	x	53.2
	High	0.73	4.23	5.43	x	59.4	x	0.745	x	29.6	x

MSEs from lactate predictions when in the presence of glucose are generally better than the predictions of glucose in the presence of lactate, through the use of the model in buffer.

For both analytes, the errors of prediction decrease with higher spiked amounts of that analyte. Naturally, in the presence of fewer molecules, there are fewer bonds to vibrate which results in noisier spectra and consequently higher errors of predictions.

As shown before, spectral characteristics derived from glucose variation are not as prominent as the ones originated by lactate. In fact, the higher level of molar concentration of lactate is almost ten times higher than the maximum spiked concentration of glucose. This results in a blending of the effects of glucose concentration on the stronger effects of lactate variation. Lower concentrations of glucose might therefore not be detectable in the presence of lactate at the wavelengths used.

Spent media revealed to be the most challenging background tested for the buffer model. Matrix effects on spent media are not accounted for in the buffer model. Additionally, the model in buffer has been built including glucose concentrations up to 10 g/L. The higher level of spiked glucose exceeds this value which results in less accurate results. A robust NIR model must incorporate the full concentration range of the components being modelled (Workman, 2008). A larger range of calibration standards would therefore be required in this case.

Macaloney et al. (1997) made some similar observations while assessing the effect of changes in the C:N ratio in robustness of models for analytes concentration. Model performance was good if a similar feed formulation would be used although when new organic components are added they can create additional overlapping absorbances which might significantly change the bias and/or slope of a particular model. They also observed that statistics of biomass predictions were better when the concentration range being predicted was similar to the one used in the model.

3.3.5 In-line Approach to Universal Model Development

The previous section has described an at-line approach to reach a universal model. As discussed in section 2.3.1 there are however multiple ways to obtain NIR measurements.

While at-line provides valid information, it is based on manually collected samples which are of limited frequency. As such, it can be that the bioprocess has prematurely entered the death phase. In-line measurements provides real time information, allowing the user to intervene in the bioprocess when necessary. On the downside, given the nature of in-line probe being inside the fermentation vessel, the collected spectra are subject to the fermentor process variables, such as agitation, aeration and temperature.

This section will first measure the variability caused by process variables on spectra. The consequent model development deals with this variability by keeping process variables constant. As such, an in-line approach to the universal model is provided.

3.3.5.1 Effects of process variables

Process variables can have an effect in the collected NIR spectra. Temperature, agitation and aeration were analysed.

3.3.5.1.1 Temperature Effect

As temperature changes influence the vibration intensity of molecular bonds, it is highly relevant to know the effect of temperature variation in spectra. Pure spectra of water and glucose in water were analysed.

To analyse the temperature effect in water peaks, three different temperatures were compared from spectra collected in-line. Figure 3.30 shows three spectra at three different temperatures. It is possible to see that an increase of temperature (from blue colour to red) results in sharper peaks in the O-H first and second overtone regions (1400 nm and 950 nm, respectively) and a shift left (towards lower wavelengths) of the peak in the O-H combination region (1900 nm). The latter is a composition of different hydrogen bonding. When the temperature increases the number of free OH groups increase as well, so the relative absorbance of the other peaks increases while this peak shifts towards higher energy (lower wavelengths). This is in agreement with what was described by Wülfert et al. (1998).

For glucose solution, the same relation with temperature is visible 3.32 but it might be attenuated by the C-H bonds which also absorb in the region of O-H first overtone.

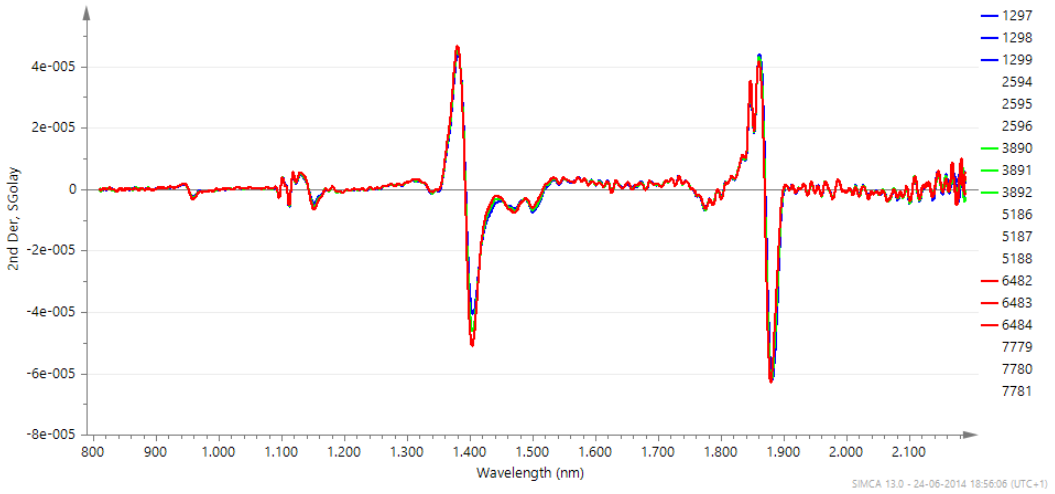

Figure 3.30: Second derivative of the online spectra of water at three different temperatures: 10°C (in blue), 20°C (in green) and 30°C (in red).

In order to investigate if the peaks in C-H region detected in water spectra might have resulted from glucose residues, both spectra were plotted together (collected at the same temperature, 20°C) and the peak of interest is shown in Figure 3.34. By comparison, the peaks result from glucose as they overlap in this region.

3.3.5.1.2 Agitation and Aeration

Different authors have stated that hydrodynamic conditions in the fermenter change the spectra thus the agitation rate and the aeration rate must be kept constant during calibration and validation runs (Cervera et al., 2009). Thus, it is fairly recognised that it has a relevant impact on the spectra but it is not often referred how it affects the spectra.

With the data collected with step changes, a PCA model of the second derivative spectra was developed. The scores plot obtained is presented in Figure 3.35, coloured by time evolution (top plot) and by stirring speed (11, 450 and 900 rpm) (bottom plot). Scores are clearly grouped by stirring speed thus the same agitation speed always resulted in similar spectral features. A relation between the spread of the scores and the agitation is noticed as well: lower agitation values resulted in clusters more concentrated and as the value of agitation increases the scores spread through the scores

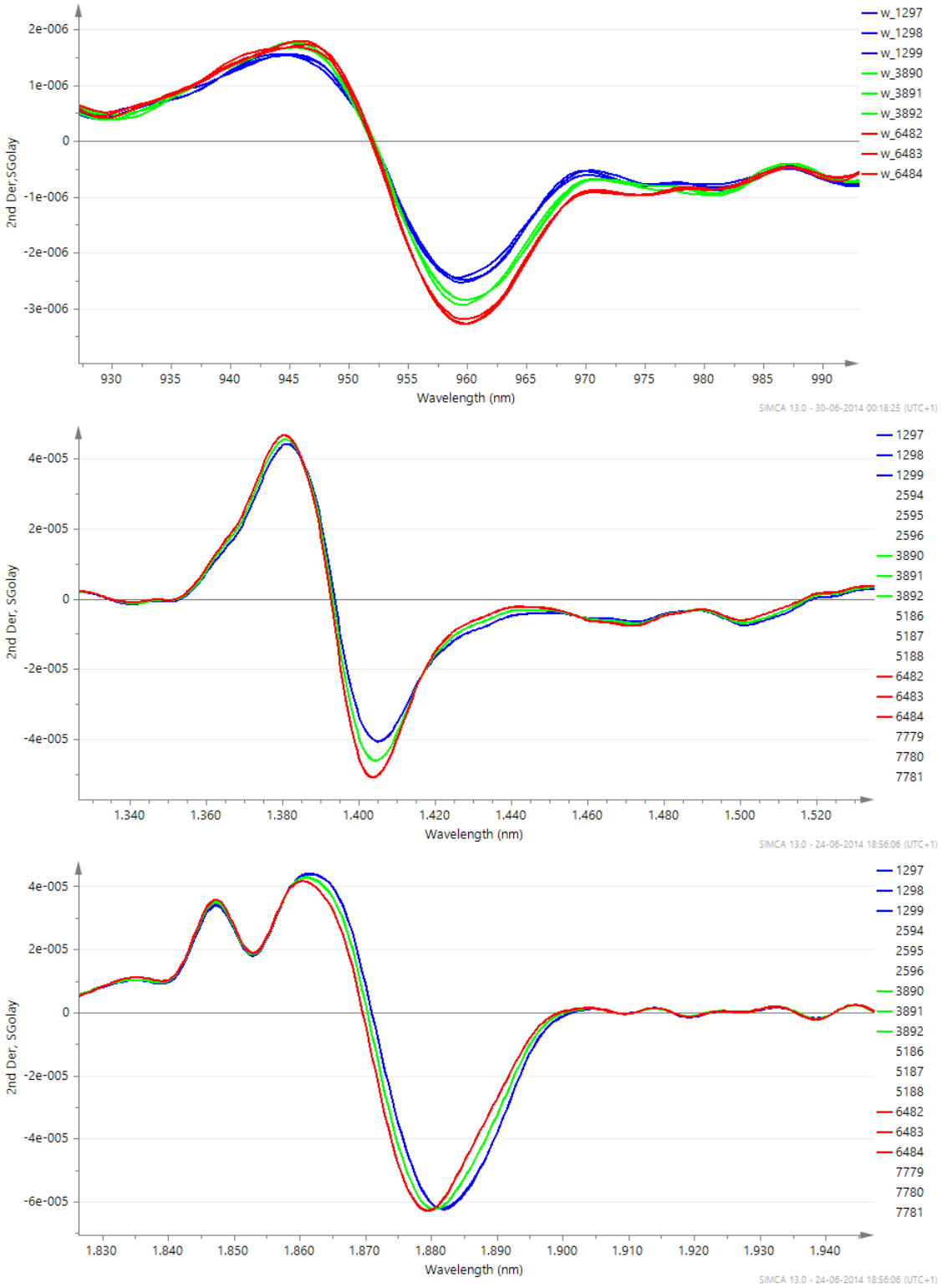

Figure 3.31: Closer look to stronger absorption peaks of water: second O-H overtone (top), first O-H overtone (middle), O-H combinations region (bottom).

Figure 3.32: Second derivative of the online spectra of 2.5 g/L glucose solution at three different temperatures: 10°C (in blue), 20°C (in green) and 30°C (in red).

plot.

Taking a closer look to raw spectra collected at the same point of aeration (Figure 3.36), it is clear the strong effect that agitation has in the baseline. It is also seen that high agitation leads to higher levels of noise.

The fact that scores are more spread when the agitation increases is related with the level of noise in spectra, less information is taken out of the spectra. Figure 3.37 shows spectra (treated with the second derivative) collected at 11 rpm (left side) is compared with spectra collected at 900 rpm (right side) in which the level of baseline noise has almost the same magnitude as the strongest water peaks. A low signal to noise ratio results in loss of sensitivity and that is also suggested by the scores plot with the spreading of the scores at higher agitation.

To characterise the effect of the air flow in the spectra, two different values of aeration at a constant value of agitation were collected and the raw spectra are shown in Figure 3.38.

The air flow showed to produce the same type of effect in spectra, although much less pronounced, that agitation produces. The population of bubbles increases with the air flow increase, which can help the transmission (Tamburini et al., 2003). However, if the bubbles also get bigger they may not enter the NIR probe slit and the light-

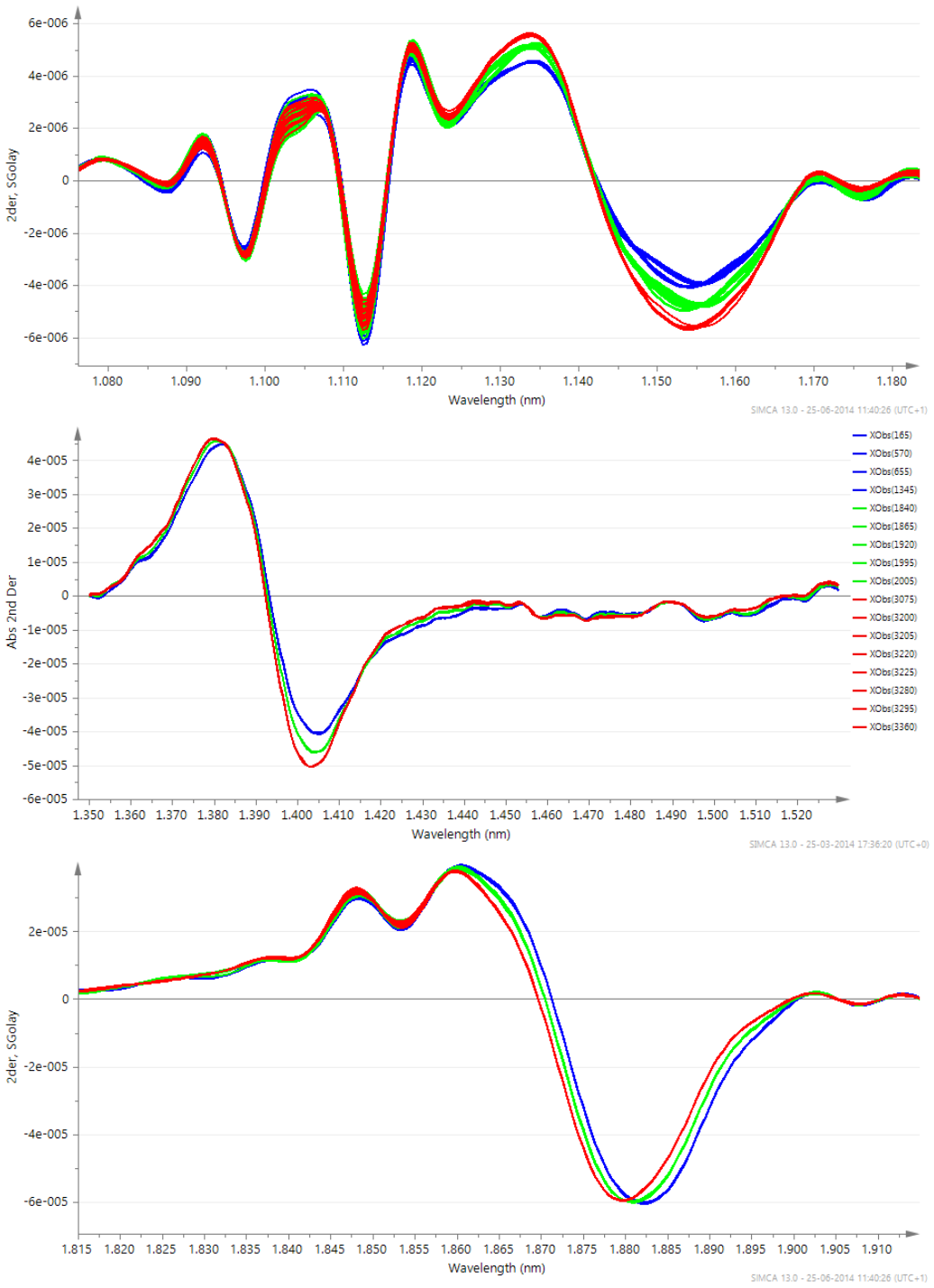

Figure 3.33: Closer look to stronger absorption peaks of water: C-H vibration second overtone (top), first O-H overtone (middle), O-H combinations region (bottom).

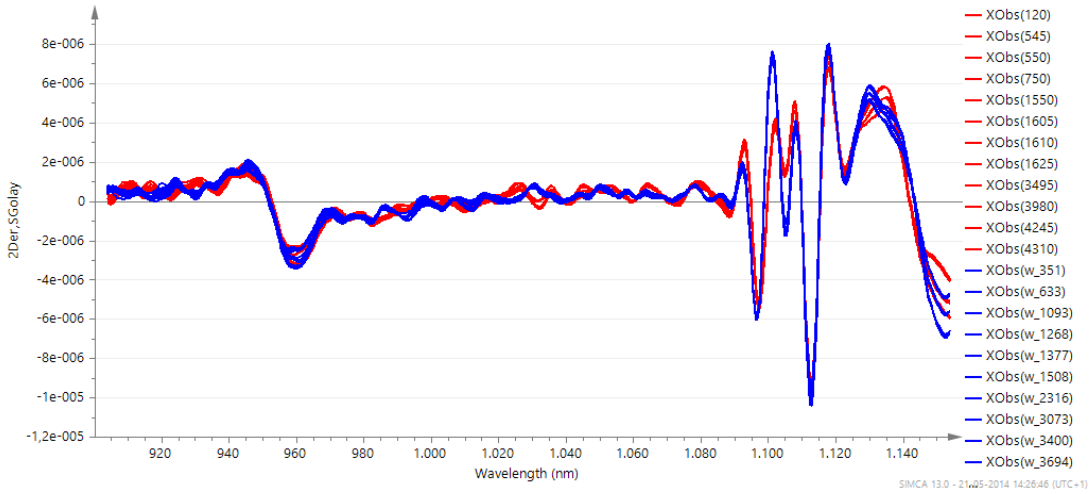

Figure 3.34: Online spectra of water (blue) and 2.5 g/L glucose solution (red). The region from 1080 to 1160nm corresponds to second overtone C-H vibrations.

scattering increases. These results are in line with the ones discussed by Cervera et al. (2009).

3.3.5.2 In-line Model Development

The goal was to develop a basic model in water for each analyte (lactate and glucose); extrapolate to predict the other datasets in different backgrounds: water, buffer, fresh media and supernatant. Once the effects of the different confounding factors are well known these can then be integrated into the original model. The calibration data was obtained by a constant feed of the analyte into the bioreactor. Setpoints of agitation and temperature were fixed and no aeration was used. For the PLS models, the wavelengths of 1600 - 1724.5 nm were selected and a MSC with first derivative was used as preprocessing. Lactate was increased up to a concentration of 5 g/L and glucose was increased up to 10 g/L.

The scores plot of the PLS model developed for glucose is shown in Figure 3.39 and a RMSEE of 0.31 g/L was obtained and an R-squared of 0.990. The first variable (t1) in x axis correlates to the increase of glucose concentration.

After this experiment, four additional ones were executed in the same vessel:

- glucose feed into the vessel containing a solution of lactate in water at 5g/L

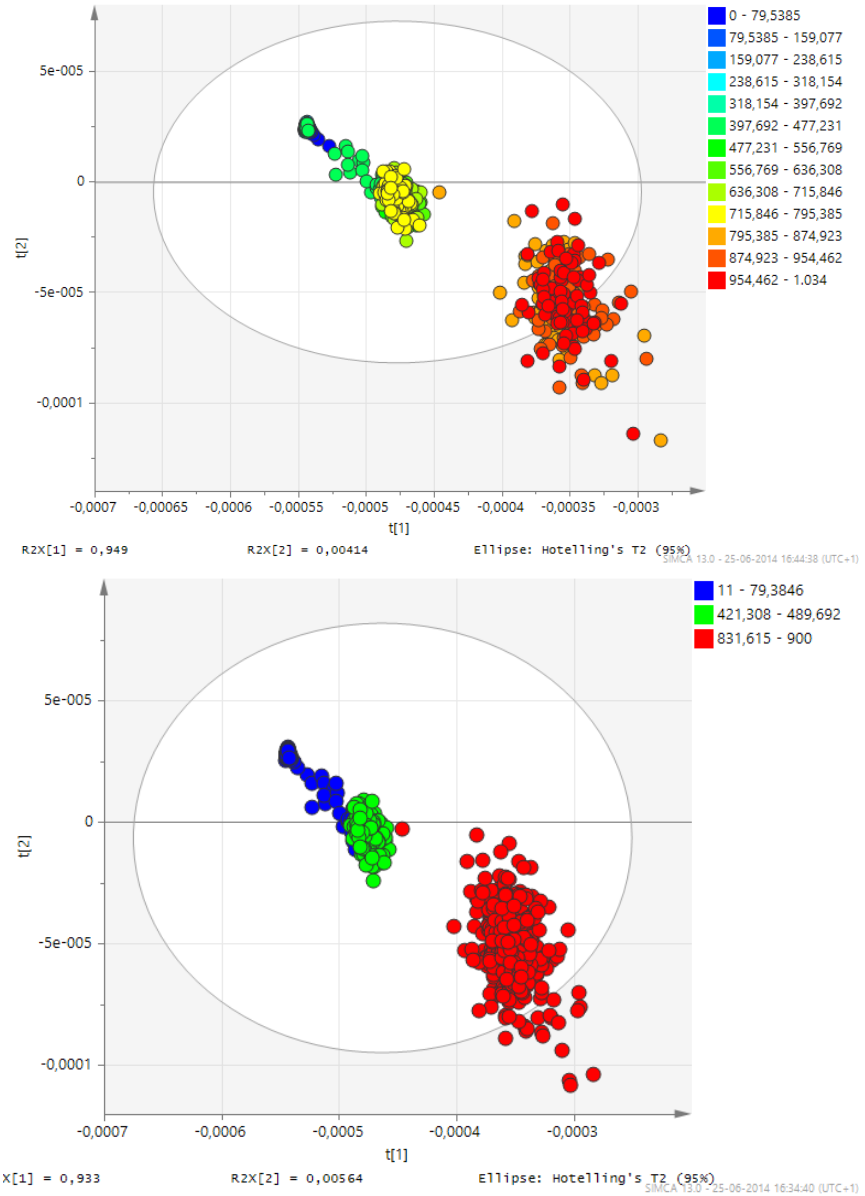

Figure 3.35: Scores plot of PCA model developed for collected spectra (treated with second derivative and smoothing filter) with step changes for agitation and aeration, coloured by time (top plot) and by stirring speed on the bottom plot (100 rpm in blue, 450 rpm in green and 900 rpm in red).

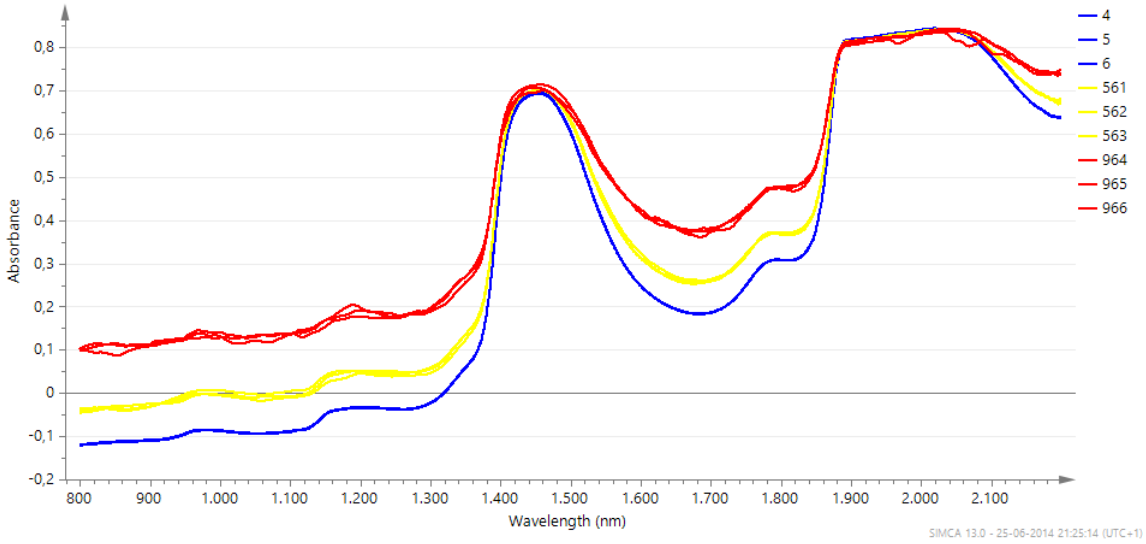

Figure 3.36: Raw spectra of glucose solution at 0.2 slpm and different stirring speeds: 11 rpm in blue, 450 rpm in yellow and 900 rpm in red.

- glucose feed into fresh media
- water feed into glucose solution
- glucose feed into solution containing cells and lactate

Figure 3.40 overlays the predicted and actual trends for glucose in these four experiments. In general, all the trends predicted from the model (green lines) follow the expected concentration in the vessel (blue lines), which is obtained by the volume of solution in the vessel and the amount of concentration being fed per unit of measurement.

Particularly, plots A and D show that the rate of change of glucose is accurately predicted in an environment with lactate. Plots B and C show some deviations from the expected trend. This could potentially be explained by the reference collected by the instrument every 30 minutes. Since it takes 19sec to collect one spectra, a total of 95 spectra have been collected after 30min. This corresponds to the first large deviation in plot B after the 95th spectra. Apparently, this procedure causes additional variation in the spectra that the model does not include yet. The same sort of deviations can be seen in plots C and D.

The scores plot of the PLS model developed for lactate is shown in Figure 3.41 and

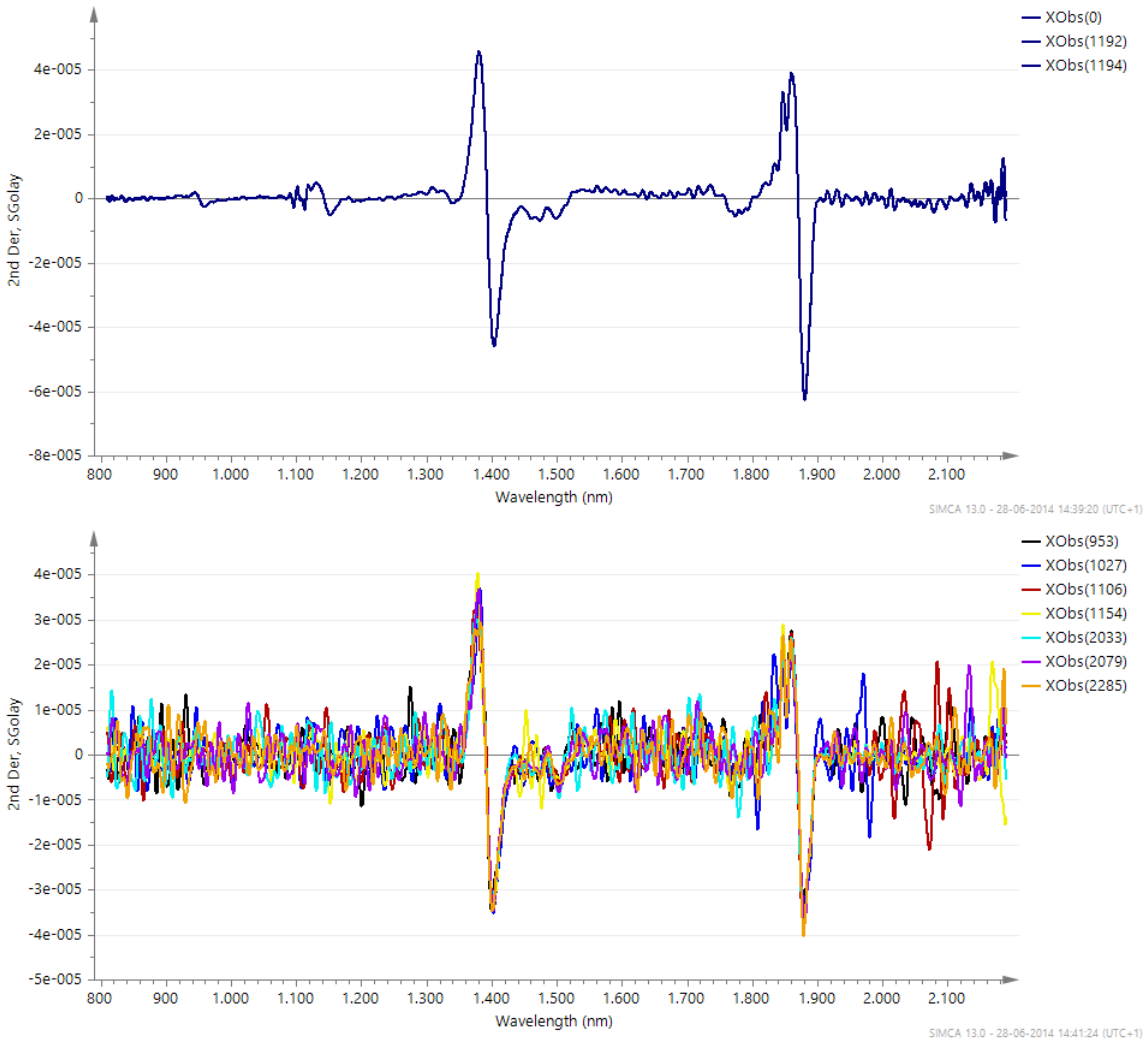

Figure 3.37: Second derivative with Saviztky-Golay smoothing filter of spectra collected at 11 rpm (top) and collected at 900 rpm (bottom).

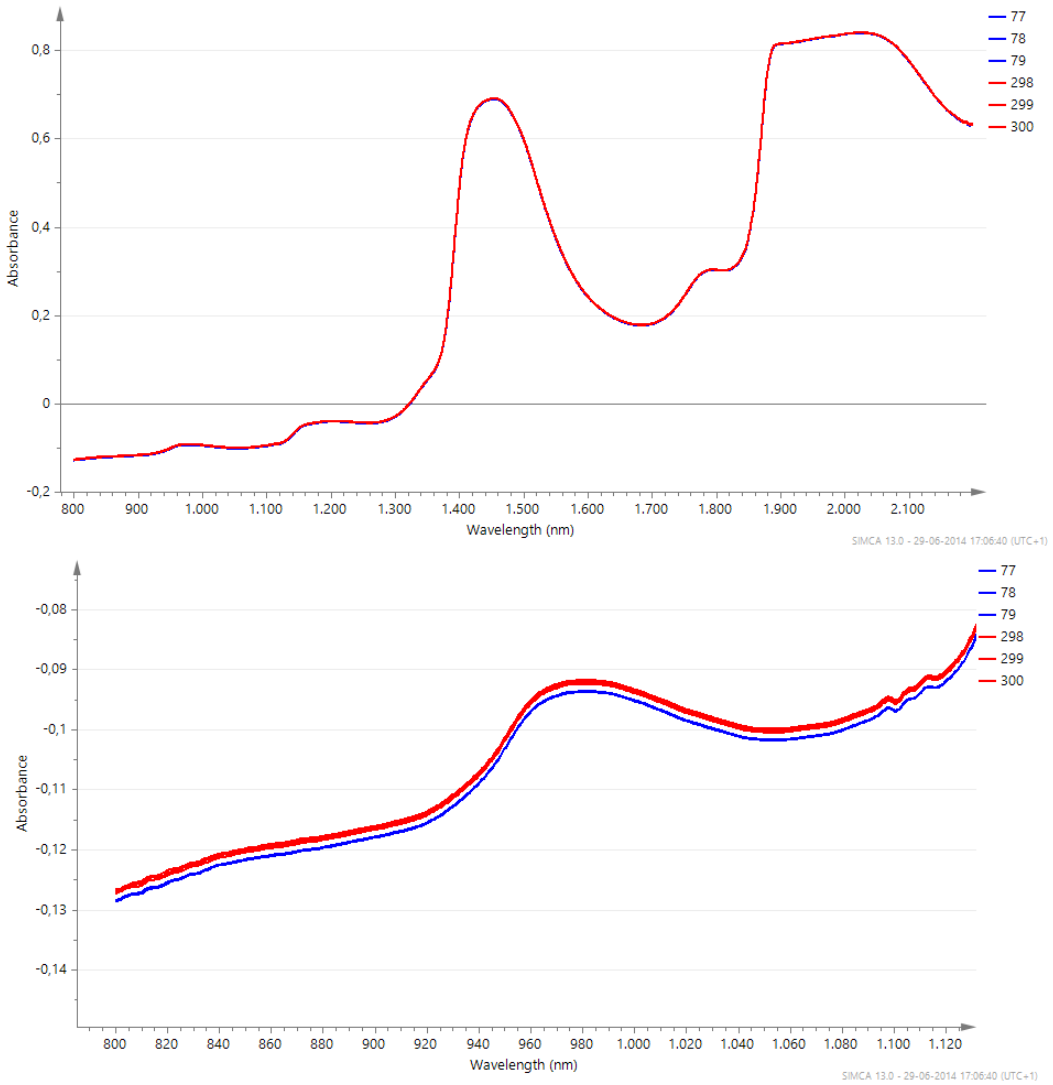

Figure 3.38: Raw spectra of glucose solution at 11 rpm and different air flows: 0.2 slpm in blue and 1.5 slpm in red. The full spectra is shown in the top plot and the bottom plot shows a close up of the first wavelengths in order to detect the difference in the baseline. Three spectra sampled at each of these two moments are shown.

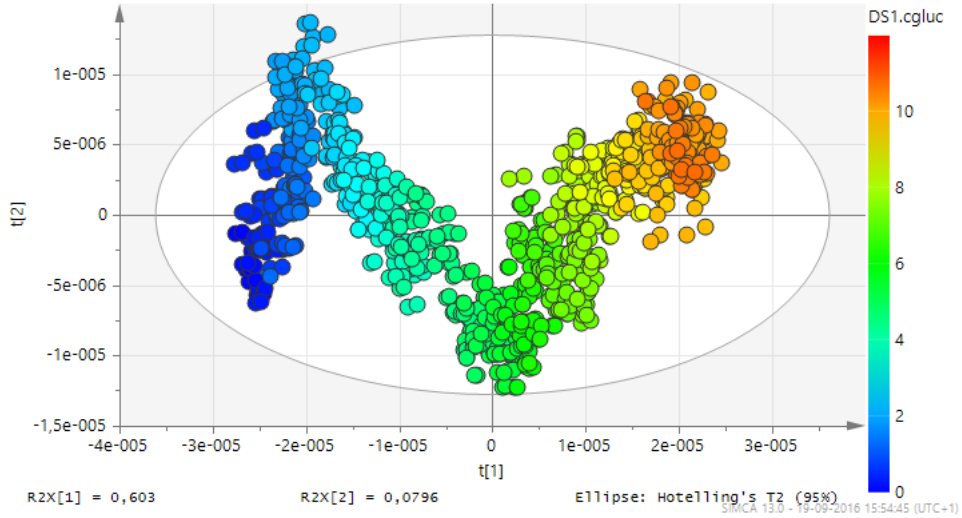

Figure 3.39: Scores plot of the PLS model developed for glucose concentration based on in-line spectral collection.

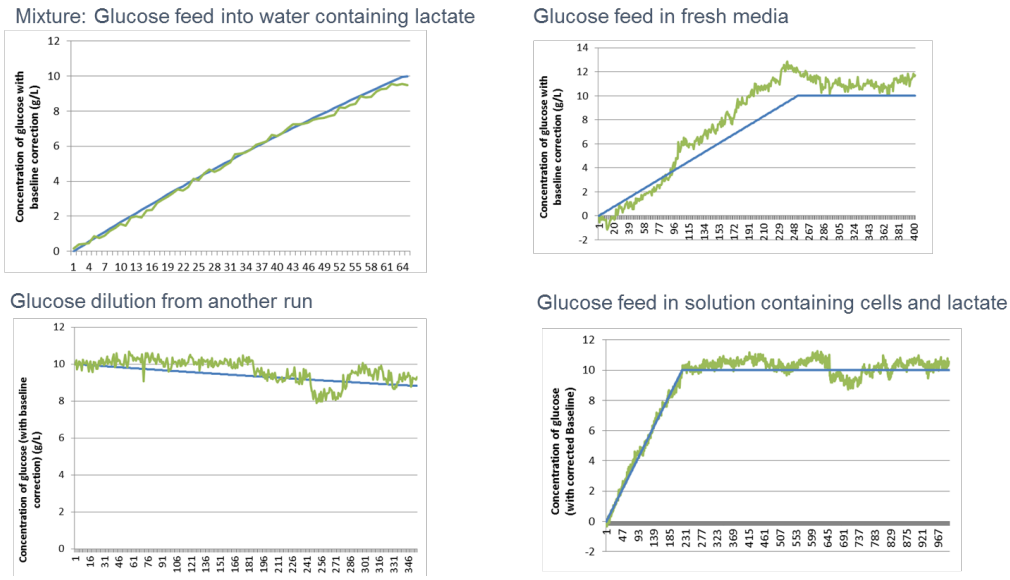

Figure 3.40: Trends of glucose based on the obtained PLS model. Green lines are predictions and blue lines are "observed" (expected based on the feed rate) concentrations of glucose.

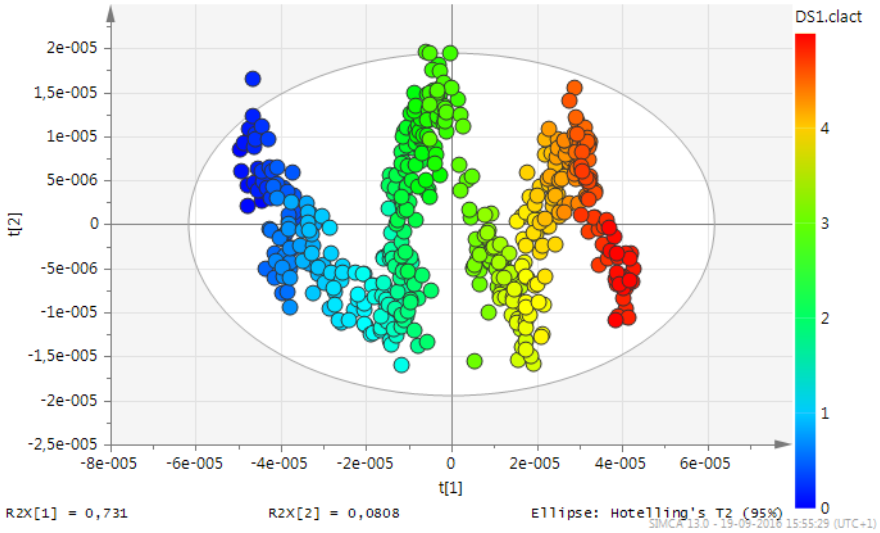

Figure 3.41: Scores plot of the PLS model developed for lactate concentration based on in-line spectral collection.

a RMSEE of 0.14g/L and an R-squared of 0.990 were obtained. The first variable (t1) in x axis correlated to the increasing of lactate concentration. Figure 3.42 overlays the predicted and actual trends for lactate in these four experiments below:

- Lactate feed in a different day (plot A)
- Lactate feed in fresh media, in sterilised conditions (plot B)
- Lactate feed in a solution of glucose in water (plot C)
- Lactate feed in solution with cells (plot D)

With exception of plot B, all the trends predicted from the model (green lines) followed the expected concentration in the vessel (blue lines). Plot B shows that the feed of lactate in fresh media cannot be accurately predicted by the simplified model of lactate in water.

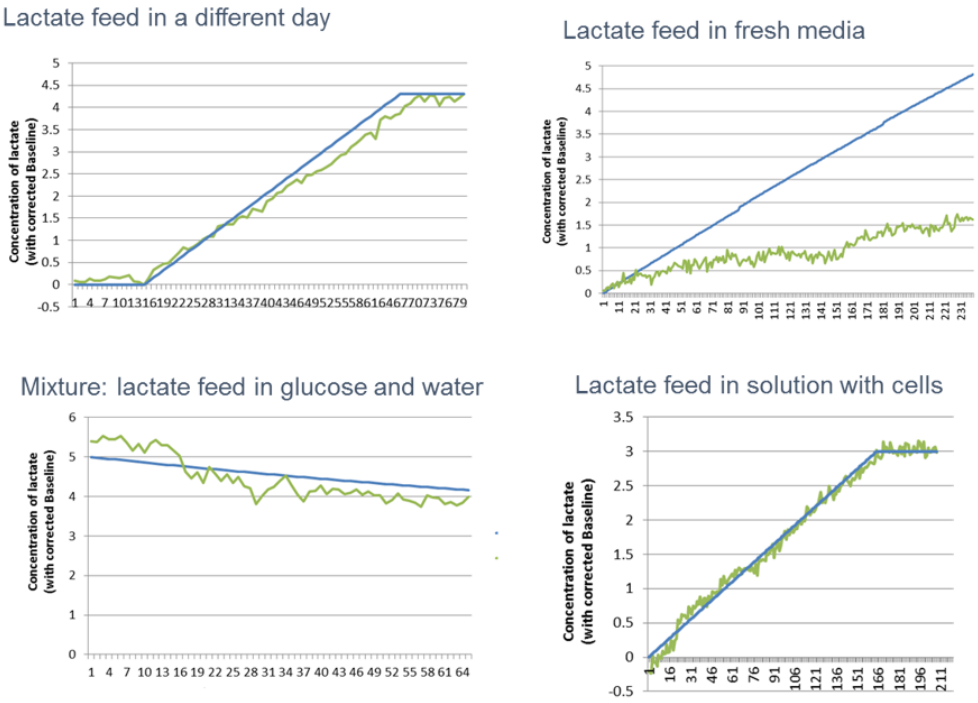

Figure 3.42: Trends of lactate based on the obtained PLS model. Green lines are predictions and blue lines are "observed", "ground truth" (expected based on the feed rate) concentrations of lactate.

3.4 Conclusions

This chapter focused on one of the main sources of variability involved in spectroscopic measurements in a bioprocess environment, which is the nature of the matrix. A detailed methodology was designed to obtain information about key analytes within a range of matrices of differing complexity.

The best wavelengths to describe the two analytes of interest in this study (glucose and lactate) were identified based on chemical knowledge and spectral observations. Also, since the goal is to evolve to online systems, the region between 1630-1750 nm was chosen as the most suitable.

The optimum spectral pre-processing was determined as the second derivative, with the data also being mean centred. This way the information content is enhanced, and the developed models are more able to pick up the desired information.

Utilising PCA models and thorough analysis of scores and loadings, the difference between matrices was demonstrated. A comparison of loadings plots allowed the deconvolution of spectra between individual analytes of interest and any matrix effects on the spectra. Thus, revealing which wavelengths were relevant in each case and this information can then be utilised in subsequent analysis.

However, care must be taken since the variance explained by the comparison of loadings plots could result, not only from the contents of the matrix, but also from other sources of variability such as temperature or instrument settings. Blank et al. (1996) mentioned that some noise from instrumental variation can become embedded in the calibration model due to correlation with the desired-property variation.

Individual PLS models were developed for each dataset and it was shown that within each dataset it is possible to predict glucose or lactate concentration. The highest errors of cross validation (RMSEcv) were ± 0.839 g/L for lactate (obtained on spent media) and ± 2.45 g/L for glucose (obtained in the presence of cells). The contribution from the analyte of interest in the spectra becomes less relevant as the complexity of the background - and consequently the variance in the spectra - increases.

With regards to the aim of having a model developed in a simple matrix which is able

to predict concentration levels on other matrices, the models for lactate quantification were better than the glucose models. Glucose models yielded no satisfactory results in terms of extrapolation, other than for samples in water.

A real-time approach can be developed through in-line measurements. It was shown that despite the added variation stemming from process variables, reliable trends can be predicted based on a simple model of glucose in water.

The effect of temperature to NIR spectra was proven not to be linear throughout the wavelengths. The vibration increase resulting from the higher temperature raises the number of free OH groups causing the peak in O-H combination region to shift towards the left (higher energy, lower wavelengths) and the relative absorbance of the other OH peaks increases.

Agitation and aeration produce the same effect in NIR: the baseline is shifted towards higher absorptions. However, agitation generates higher levels of noise in the collected spectra than aeration. If strong agitation is required for the fermentation, the spectral measurements might be compromised.

One option could be a combination of in-line and at-line strategies, given that the in-line gives real-time qualitative information on the analyte trends. It could indicate relevant moments to collect a sample for the at-line instrument. Further development of the in-line model would then allow for real-time control with feeds of glucose, for example.

This chapter has collected calibration datasets through an intensive workload, making it prone to human error. The next chapter makes use of automation to avoid manual mixture handling.

Chapter 4

Symbiosis of Automation and NIR Spectroscopy

4.1 Introduction

In developing an automated NIR system for small bioreactors, the previous chapter used the traditional method of calibration involved in model development, where the samples were collected and processed manually. That process can be time consuming and prone to human error.

Expanding on the previous chapter, this chapter describes the application of an automated liquid-handler to accurately prepare and scan the necessary mixtures for the calibration of a NIR-measuring system. Such automated processes can not only save time but also eliminate the human error involved in sample preparation. This would also allow microbiologists to readily benefit from spectroscopy measurements.

The development of a fully functional spectroscopy-based measuring system depends upon many factors, such as the selection of instrument, adequate configuration for the current application, determination of an appropriate calibration dataset, while model development can all create estimation bias and significantly alter the results. All these challenges are magnified if the measurement is done in bioprocesses. Culture broths consist of highly changing matrices as physical and chemical conditions can change throughout the bioprocess. Physical conditions such as temperature, pressure, particle size, flow regime and air bubbles, can lead to nonlinear effects on spectra resulting in failed predictions from linear models (Chen et al., 2004, 2013). More specifically, changes in particle size can affect the optical pathlength and mask the spectral variations related to differences in the chemical constituents (Jin et al., 2012); temperature fluctuations can cause broadening of the spectral bands of absorptive spectra of constituents in mixtures and nonlinear spectral shifts (Chen et al., 2011); air bubbles, flow rates, and solid impurities in the process flow impact the spectral absorption and baseline significantly (Wu et al., 2012). Chemical conditions could change during cultivation as well as between processes, as components of the media are consumed, and new ones are formed. The spectra of the resultant composition would affect the wavelength region that was previously selected and designed for an informative model. Other sources of unaccounted variability can result from changing the fibre optic probe,

or instrument aging or repair (Chen et al., 2013). Generally, the effect of perturbations on spectra can range from completely unknown to wavelength axis shifts, spectral shrinking/stretching, and nonlinear baseline shifts (Roussel et al., 2011).

This chapter targets the industrial context of developing an analyte measuring system by means of spectroscopy (NIR) through an automated system that can be used at early stages of bioprocess development.

Chapters 4 and 5 were developed within an internship at Sartorius Biosystems. This chapter describes an NIR measuring system that was developed from scratch. It included a prototype flow-cell, a diode-array spectrometer connected to transmission optical fibres and an existing automated liquid-handler that was adapted for this application. This flow-cell can eventually be implemented in bioreactors. The automated liquid-hander can be programmed to function overnight, while generating the appropriate calibration dataset for analyte prediction.

Having developed this prototype, the next step is to test the capability of the automated liquid-handler. In terms of hardware, the prototype must be able to accurately aspirate the appropriate stock solutions, mix, dispense to the flow cell and waste. In terms of software, the scripts for automation have to be developed, including a cleaning procedure that ensures non-cross-contaminated samples.

To test the predictive capability of the NIR instrument, different datasets were prepared by the automated system. Samples of pure analytes, binary and ternary mixtures were prepared with the liquid-handler, automatically scanned by the NIR instrument and discarded to a waste vessel. Ultimately, an optimized dataset for model development of analytes of interest in fermentation should be identified.

The structure of this chapter is as follow: it will commence by describing the hardware and software that were used to collect data in an automated manner. The methodology for data analysis and figures of merit used to evaluate the models obtained with the new configuration will be described. Then, followed by a description of steps to modify the hardware and development of a user-written scripts to collect spectroscopic data in an automated manner will be detailed. Finally, under the results section, an extensive exploratory data analysis will be presented with the goal of determining the

achievable accuracy on the quantification of analytes with this new system.

4.2 Materials and Methods

This section describes the configuration of the hardware used and how it was adapted for the desired application. It consists of hardware for spectroscopy (a miniaturized spectrometer, light source, power supply, optical fibres, a prototype flow-cell) in combination with hardware for sample preparation in an automated manner (pump valves, moving wheel, distribution valve, syringe pump, tubing). A second section describes the methods used for data analysis: software used for multivariate data analysis; spectral pre-treatments; figures of merit used to evaluate the obtained results of developed quantification models; and quantification of variability.

4.2.1 Hardware Set-up

The instrumentation used for data collection is shown in Figure 4.1. This comprised of a light source (B), responsible for generating the near infrared radiation, which goes through the sample and it is collected in the detector (A) and the collected spectrum is saved in the laptop through the USB cable. This radiation is transmitted through the optical fibres (E). The studied mixtures were prepared with an automated liquid-handler, the AM (analysis module) (C), and then sent through Teflon tubing to the flow cell (D).

4.2.1.1 *The automated liquid handler*

The modified analysis module (AM) consisted of a syringe pump, five pump valves connected to a dispensing wheel which feeds to a well vessel. A schematic diagram is represented in Figure 4.2.

4.2.1.2 *Spectroscopic Tools*

A light source from Avantes (AvaLight-Hal) was used, together with a miniaturized array detector (JETI Technische Instrumente GmbH). A classical spectrometer must

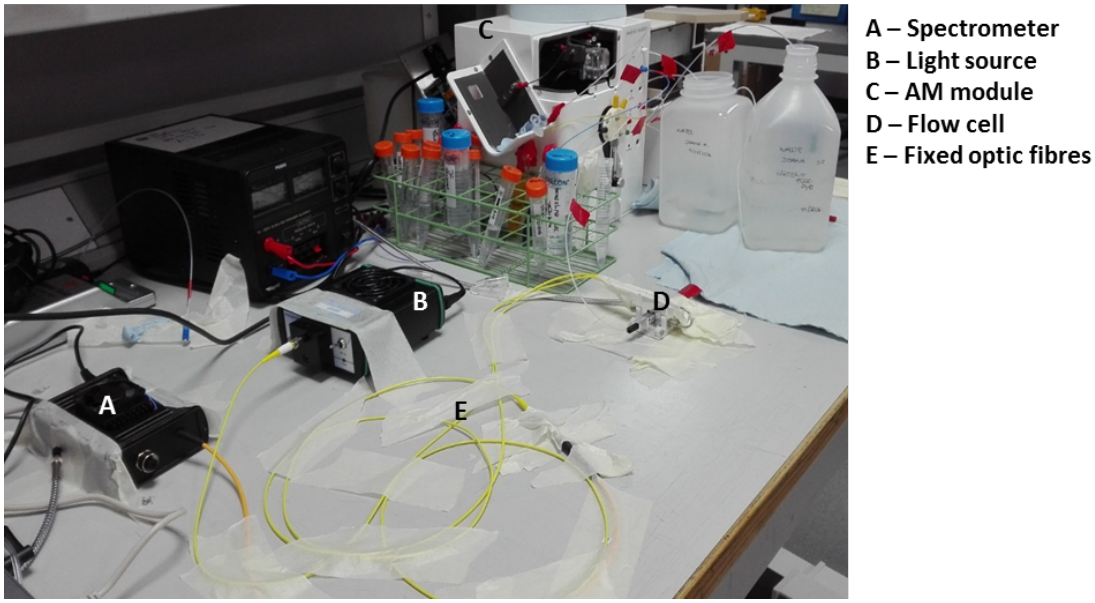

Figure 4.1: Hardware set-up for sample preparation and near infrared spectra collection. Legend in the picture describes the elements seen: A - NIR detector; B - light source, C - Analysis Module used for sample preparation (AM); D - flow-cell; E - fibre optics.

split multi-chromatic radiation into its spectral components and therefore it is composed by an input slit, a rotating dispersive element (prism or grating), an output slit and a single detector. The array spectrometer uses a detector array instead of a single one, which allows the implementation of fixed components. A classical spectrometer has higher sensitivity and lower stray light, but involves several drawbacks such as the moveable elements, the space consuming dimensions and the non-parallel measurement. These drawbacks were overcome with the implementation of array detectors (JETI 2005). Some disadvantages involved in the use of array detectors are their lower sensitivity (than a monochromator), the precision and resolution is also normally less than that of laboratory instruments with a rotating dispersive element (JETI 2005). The AD converter (which transforms the analog signal to a digital level) delivers the spectral signal in counts for each pixel. The pixels are numbered and these numbers have to be transformed into the corresponding wavelength.

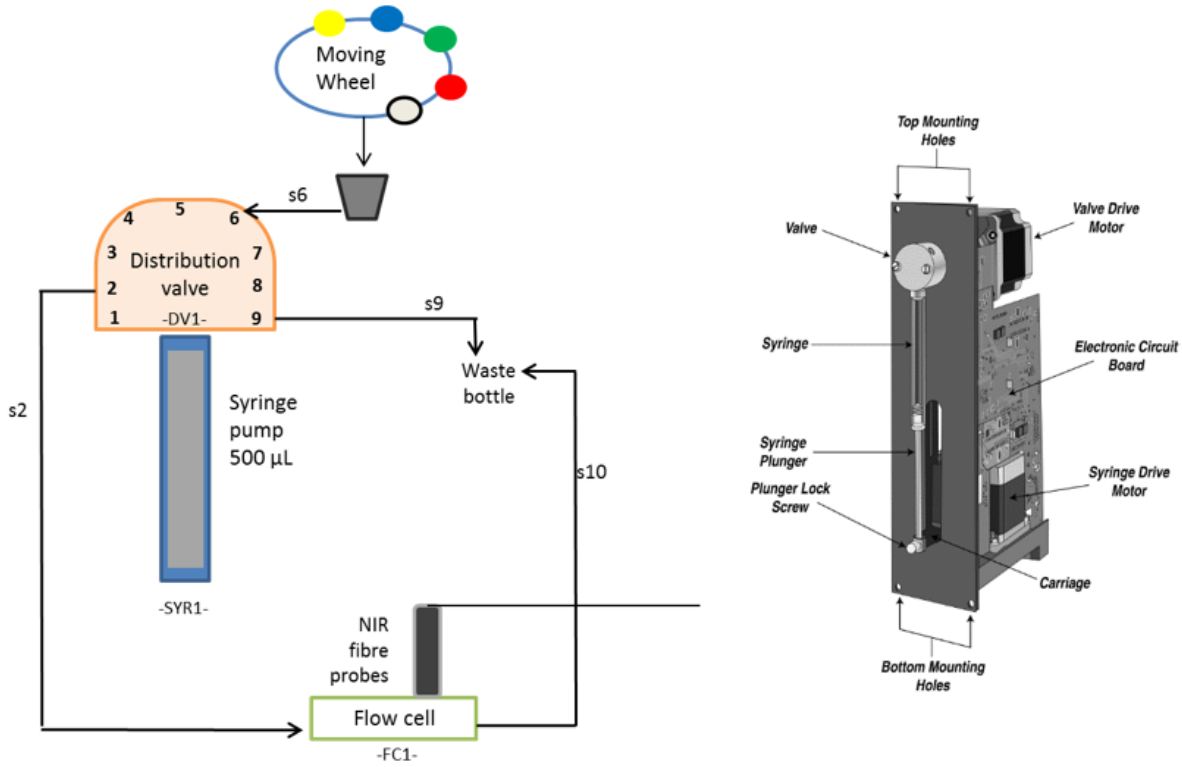

Figure 4.2: Schematic diagram of the AM used (left) and detail of the syringe pump (right).

4.2.1.3 *Flow-cell*

The flow-cell prototype consists of two main metallic plates with two sapphire windows between which the sample is halted for spectral collection as shown in Figure 4.1. The procedure used to clean the flow-cell was to stream water through the cell using a manual syringe, while creating some bubbles; then flushed with ethanol and finally, connected to a tube of clean compressed air which would be gushed through the cell for about 30 sec to ensure no droplets would be trapped in the small measuring gap. To validate the procedure, intercalated samples of water and high concentration of glucose were scanned over the course of some hours to guarantee the spectra of water samples were collected free of glucose residues.

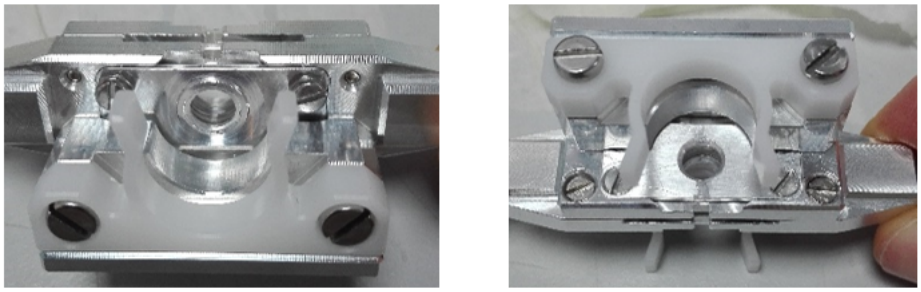

Figure 4.3: Configuration of the "in-house-built" flow cell.

4.2.2 Data Analysis

4.2.2.1 *Software*

SIMCA 13 (Sartorius, Sweden) was used as extensive chemometric analysis tool for multivariate modelling, visualization and interpretation of spectra. Further data manipulation was performed using the software R, through R-Studio.

4.2.2.2 *Spectral Pre-treatments*

Different pre-processing techniques were tested on the spectral datasets. Use of derivatives were one of the most commonly employed methods. The first derivative removed baseline offset variations and the second derivative removed baseline offset differences

and differences in baseline slopes between spectra. Baseline shifts usually result from physical changes as for example the increasing turbidity in the medium, that outcomes from higher cell density, thus resulting in less light reaching the detector. For cases in which chemical information in the spectra is of interest, such as in bioprocesses, derivatives are useful to ensure robustness against these types of perturbations. The Savitzky–Golay (SG) algorithm was applied for smoothing, which avoids noise augmentation from the calculated derivative. SG computes the derivative and applies smoothing in one single step by running a least-squares polynomial fitting (Rinnan et al., 2009)

4.2.2.3 *Figures of merit*

To evaluate the performance of the models, the predicted concentrations of the validation datasets were assessed using various statistical analytical parameters in addition to those already employed in the earlier sections of this thesis. The quality of quantitative concentration predictions was evaluated by the relative percentage error in predicted concentration (%RE), the root mean square error of prediction (RMSEP) and the bias as defined by Eqs.(4.1-4.3), respectively.

Variability is the amount of imprecision. A measure of precision is the coefficient of variation (CV), also known as RSD (relative standard deviation), given by the ratio of standard deviation to the mean spectrum. The advantage of using CV is the fact that it consists of an estimation of the variability independently of the signal intensity

The R-squared of a plot of the standardised residuals was also used to evaluate the goodness-of-fit of the PLS models developed. While the residual is the difference between the observed values and the predicted value, the standardised residual is the raw residual divided by the residual standard deviation.

A well specified model should have randomly distributed residuals. A normal probability plot of the standardised residuals can be obtained in SIMCA both facilitates and validates identified outliers and structure in residuals. In this plot, the points of standardised residuals will lie in a straight line, between -2 and +2, when the original residuals are randomly distributed. Thus, a value of R-squared (the coefficient of

determination of the fitted line to these points) closer to 1 indicates a good residuals' distribution.

$$RE\% = 100 \times \sqrt{\frac{\Sigma_{i=1}^{n}(Y_{obs} - Y_{pred})^2}{\Sigma_{i=1}^{n} Y_{pred}^2}} \tag{4.1}$$

$$RMSEP = \sqrt{\frac{\Sigma_{i=1}^{n}(Y_{obs} - Y_{pred})^2}{n}} \tag{4.2}$$

$$Bias = \frac{\Sigma_{i=1}^{n}(Y_{obs} - Y_{pred})}{n} \tag{4.3}$$

$$\%RSD = \%CV = \frac{SD_{predictions}}{Mean_{predictions}} \times 100 \tag{4.4}$$

4.2.3 Solutions

All chemicals were purchased from Sigma-Aldrich unless otherwise noted. Stock solutions of glucose were prepared from D-(+)-Glucose (G8270-5KG, Sigma-Aldrich) and lactate stock solutions were prepared from (DL-Lactic acid 69775, Fluka, Sigma-Aldrich). To obtain a solution of cells, a commercial dried baker's yeast (*S.cerevisiae*; dried baker's yeast, Tesco, UK) was used. A solution of 30 g/L of washed cells was prepared and stored at 4°C. To prepare this suspension, the appropriate amount of dried yeast was rehydrated and made up to 100 mL of deionised water. After that, this solution was centrifuged at 3000rpm for 10min, the supernatant decanted, the cells were resuspended and then vortexed to ensure good mixing. This washing procedure was repeated two times.

4.2.4 Development of Automated Spectral Acquisition

The developed methodology for spectral collection through this new system is described. The variability involved was quantified and exploratory data analysis allowed for model building and full characterisation of the capabilities of such a measuring system.

4.2.4.1 Software settings for spectral collection

In order to implement the appropriate settings for spectral collection, samples were manually prepared at a first stage and injected in the flow-cell. Spectra were collected through the standard spectrometer software (the interface is shown in Figure 4.4). To initiate spectral collection, a dark spectrum must be collected, which represents the collection of a scan with the light source switched-off. After the light source is switched-on, a reference spectrum of air is collected, for which no sample is inserted in the flow cell and the resultant spectrum is obtained by the following equation:

$$CorrectedSpectrum = \frac{Spectrum - Dark}{Reference - Dark} \tag{4.5}$$

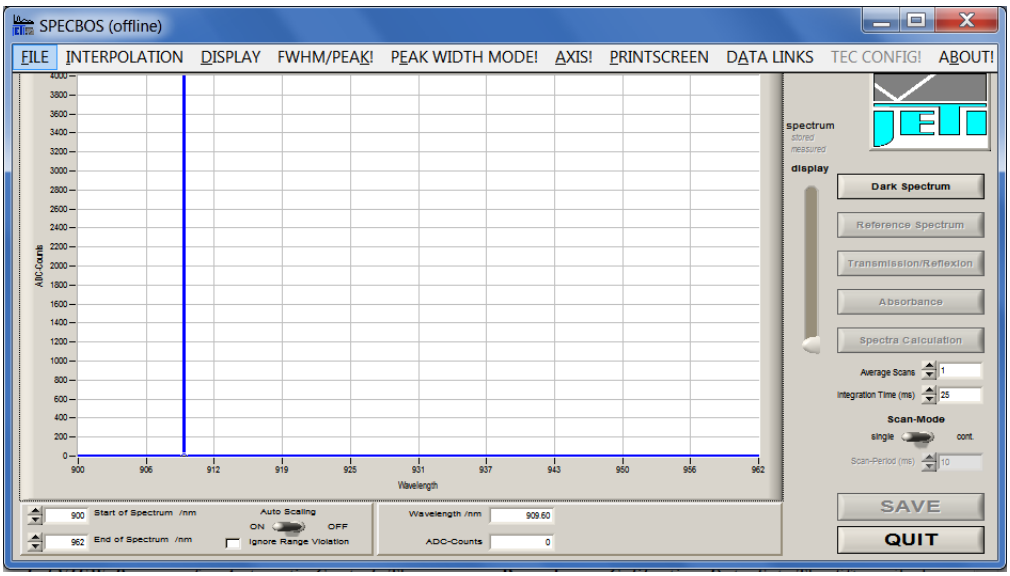

Figure 4.4: Interface of software available for spectral collection freely available (VersaSpec v3.1.4. by JETI).

The selection of the integration time (in ms) is also mandatory, which is the time of light exposure to the pixel and was used to adjust the "ADC counts" (y axis on the plot in Figure 4.4). For best measuring results, it is advised to select an integration time that generates a signal between 2/3 and the full scale of the ADC counts was used. Common integration times for array spectrometers are between 20 and 5000ms. Longer integration times allow for more signal collection, until a point of saturation is

reached.

As shown in Figure 4.5, at 50ms (orange line), and up to 200ms of integration time, there are several regions of the spectra that become saturated. The maximum signal at 25ms (yellow line) is within 2/3 of the full scale of ADC counts at the first region of the spectra. However, at the second part of the spectrum (around a wavelength index of 66), the integration time that satisfied this criterion was 50ms. In addition, the selected integration time should not generate saturated regions in any part of the spectra. Therefore, the developed PLS models were built on the spectra collected at the integration time of 25ms.

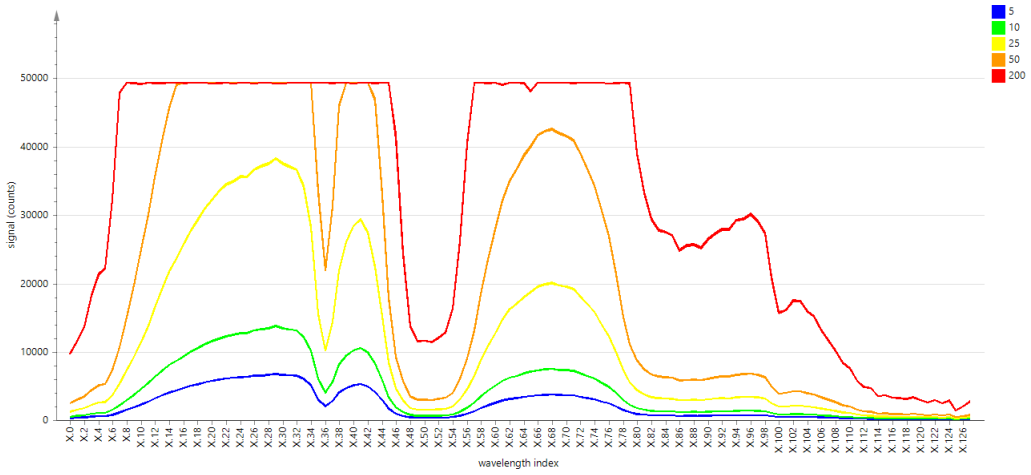

Figure 4.5: Raw spectra of water collected at different integration times: 5ms (blue line), 10ms (green line), 25 ms (yellow line), 50ms (orange line), 200ms (red line).

The coefficients of variation (CV) at each wavelength, given by the ratio of standard deviation to the mean spectrum, were obtained for these raw spectra and were plotted as presented in Figure 4.6. The CV gave an estimation of the variability independently of the signal intensity. The plot in Figure 4.6 signified decreasing variability (blue ¿ green ¿ yellow ¿ red line) with increasing integration time (5ms - blue line, 10ms - green line, 25 ms - yellow, 50ms - orange and 200ms - red line). At the orange line (50ms), some values of CV are zero and the red line (200ms) shows many values of zero, as well. However, as seen in the previous plot, these values correspond to the points of saturation which might mean the standard deviation ($SD = \sqrt{\frac{\Sigma|x-\bar{x}|^2}{n}}$) is zero. The

best scenario is at 25ms (yellow line) as the CV values are the lowest possible without reaching zero.

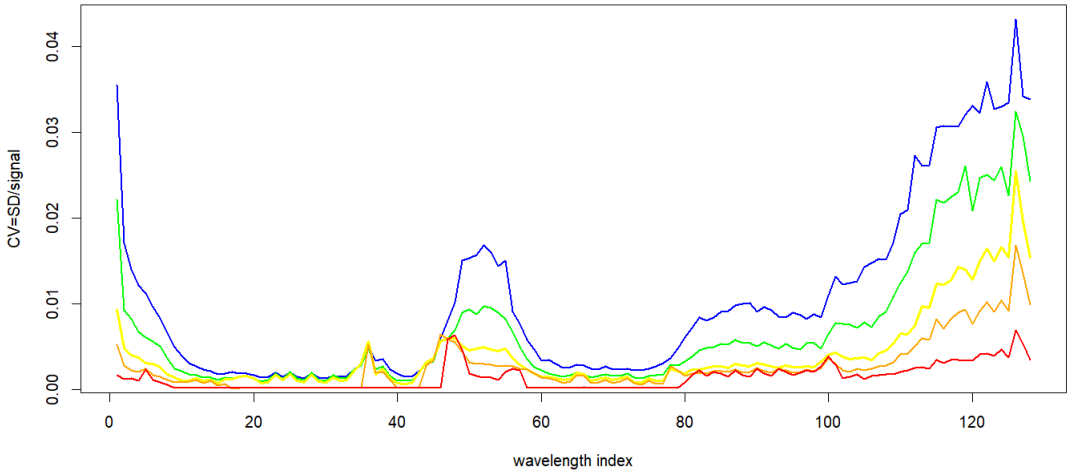

Figure 4.6: Coefficients of variation for raw spectra of water samples for different integration times: 5ms (blue line), 10ms (green line), 25 ms (yellow line), 50ms (orange line), 200ms (red line). The selected integration time was of 25ms.

Once the spectrum is correctly obtained, it can be visualised in the software in terms of transmission/reflection units or absorbance. Data can then be extracted in *.txt* or *.csv* formats.

Even though this approach works for the spectral collection of a few samples, it becomes less practical for when a high number of samples is required. Therefore, a methodology that allows a high number of spectral collection had to be developed, as described below.

4.2.4.2 Automated spectral collection

Given the novel character of the built system, available spectral collection software was deemed insufficient to collect the high number of spectra. A new script, written in *.xml* (eXtensible Markup Language), was therefore developed. A flow diagram of the user-written scripts is shown in Figure 4.7.

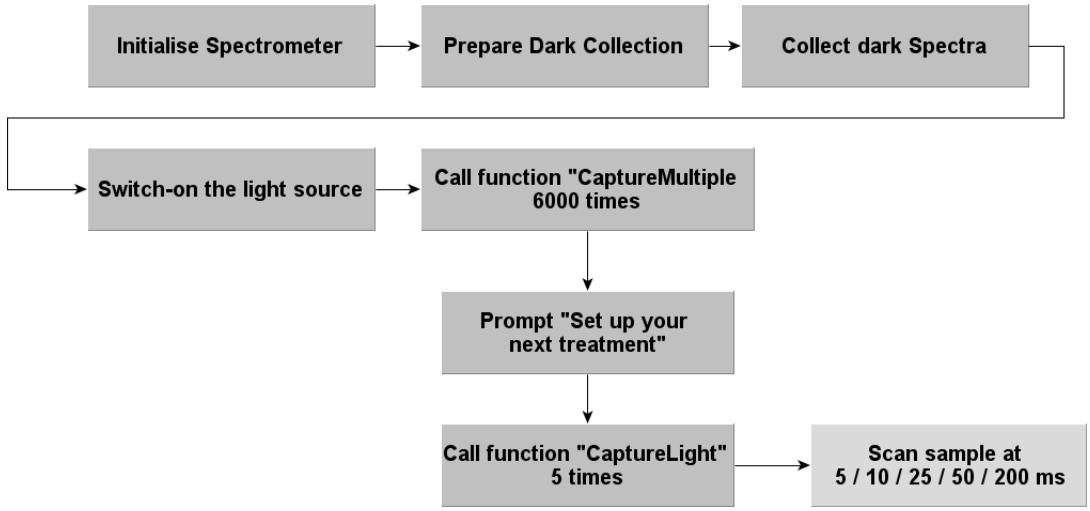

Figure 4.7: Flow diagram of the script written for automated spectral collection.

4.2.4.3 Automated mixture preparation

With the goal of bridging the potential of NIR with the potential of automated liquid handlers, an initial script was developed to use the valve pumps available in the modified AM (Analysis Module) liquid handler (shown in Figure 4.2) that would measure the amount of each stock solution and mixing in the well. A second, optimized approach made use of the syringe pump (also seen in Figure 4.2) for liquid measurement. The step-wise procedure was as follows:

A) AM is initialised; function "InitialiseAM";

B) Spectrometer is initialised; function "InitialiseSpec";

C) With the syringe pump, liquids are measured into the well;

D) To mix the solution, the syringe pump quickly aspirates the well's contents and quickly returns it; this step is performed three times to ensure good mixing;

4.3 Results

After collecting the various samples with the automated liquid handler and the developed scripts, the next step was to test it. The selected datasets were used to test

both the model itself, as well as its ability to predict analytes outside of the calibration dataset. First, the model should be able to accurately dispense the solution and execute the cleaning procedure in an automated manner. Second, the spectrometer and flow-cell should be able to provide valid spectra. This will be tested by building multivariate regression models to analyse variability and compare the different figures of merit.

Except for testing the system, this section will investigate the system's ability to predict analytes outside the calibration dataset.

Calibration models were developed based on four different datasets: binary mixtures ("glucose and lactate" and "glucose and yeast") and ternary mixtures ("glucose, lactate, buffer", "glucose, lactate, fresh medium"). This will assist the ultimate goal of limiting calibration procedures for the end user.

4.3.1 Binary Mixtures

4.3.1.1 Dataset Glucose x Lactate

A binary dataset of glucose and lactate composed of 77 individually samples were prepared by the automated system described below. Three scans were collected per sample. These samples were prepared between the 9th and 10th of February.

Model for Glucose

To quantify the concentration of glucose in this environment, different PLS were developed, using different spectral pre-treatments. Table 4.1 shows the figures of merit for six models. Both RMSEE and RMSEcv were used for model comparison, as well as the R-squared of the standardised residuals. These metrics were defined in section 4.2.2.3.

All models used two latent variables. The two lowest errors of estimation were yielded by models M3 and M5. These models also have an R-squared of the standardised residuals higher than 0.99. The RMSEE was lower at M3 than M5, whilst the RMSEcv was higher. However, the error of prediction (RMSEP) obtained on an external dataset was lower through model M6 and thus this was the final model selected for this binary

Table 4.1: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for glucose concentration (in g/L), and the R-squared of the standardised residuals.

Model	**Pretreatment**	**RMSEE**	**RMSEcv**	**R-squared**
M1	Raw spectra	12.8	13.0	0.927
M2	1der/SG	7.59	7.62	0.985
M3	2der/SG	6.56	6.75	0.996
M4	MSC/1der	8.17	8.25	0.98
M5	MSC/2der	6.67	6.65	0.993
M6	SNV/2der/SG	6.65	6.62	0.993

dataset. The detailed statistics for this model are shown in Table 4.2 and the scores plot is in Figure 4.8.

Table 4.2: Model statistics for the PLS model for glucose concentration (in g/L), based on 77 samples (231 scans).

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.914	0.344	104	0.343	6.65	6.62
2	0.0594	0.630	6.78	0.959		
Cumulative	0.974	0.973	-	0.973		

These results prove the ability of the developed automated system of producing reproducible mixtures between the two analytes, the syringe pump accurately collected the solution from each stock, the cleaning procedure used between samples was efficient. All these samples were randomly prepared and collected without human intervention overnight (from 5pm until 10am). The arrows indicate the increase of glucose (orange arrow) and lactate (green arrow). The position of the mixtures is correctly located in the scores plot, in relation to the pure samples of glucose and lactate, as well as the water sample. However, the PLS model was not able to isolate the effect of glucose in the spectra from the effect of lactate and thus the loadings are a composition of both effects.

Each point from the scores plot represents one full spectra. The preprocessed spectra of pure samples of glucose at 0, 30, 60, 70 and 90 g/L are shown in Figure 4.9. Differences between spectra were not easily detected. The spectra of pure samples of lactate at 0, 0.5, 1.0 and 2.0 M with no preprocessing (top plot) and treated with SNV,

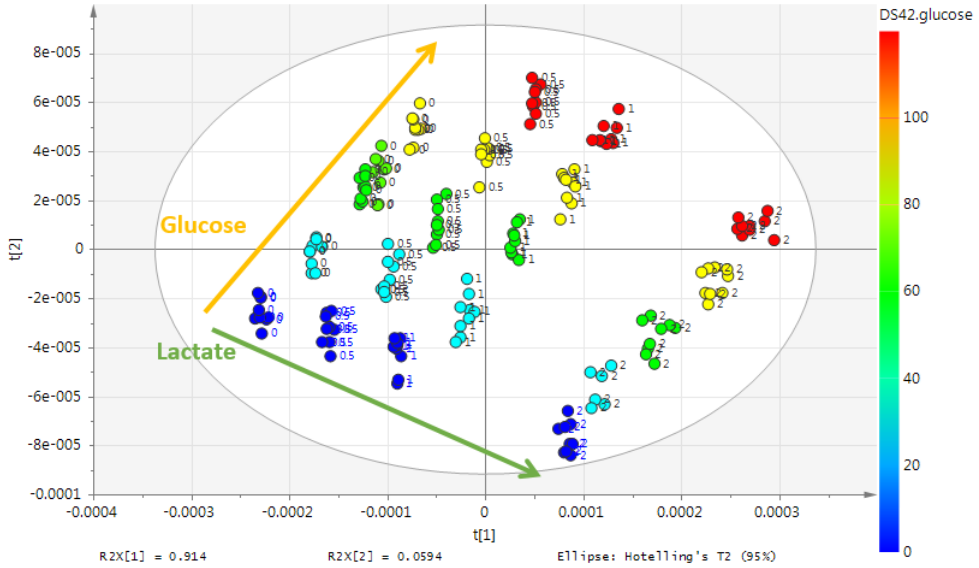

Figure 4.8: Scores plot from the selected PLS model for glucose (M6); scores are labelled by the concentration of lactate (M) and coloured by glucose concentration (in g/L and as shown by the legend scale in the plot). Increasing concentrations of glucose along the orange arrow and increasing concentration of lactate along the green arrow.

second derivative and Savitzky-Golay filter (bottom plot) are shown in Figure 4.10. There are visible differences in the region of 1400 to 1600 nm between pure samples of different concentration of lactate. Figure 4.11 shows three plots for contents of glucose and lactate for the same sample index generated from two components of the PLS model ordered by solutionID with which samples were randomly prepared.

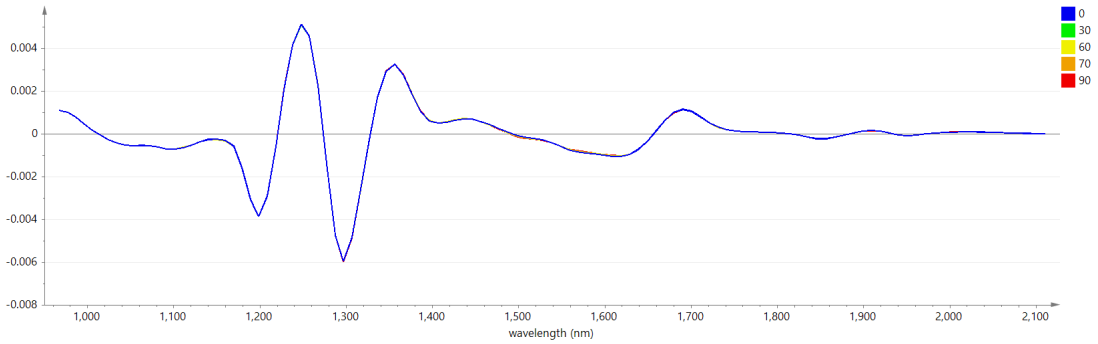

Figure 4.9: Preprocessed spectra of pure samples of glucose (0, 30, 60, 70 and 90 g/L), as used in model M6.

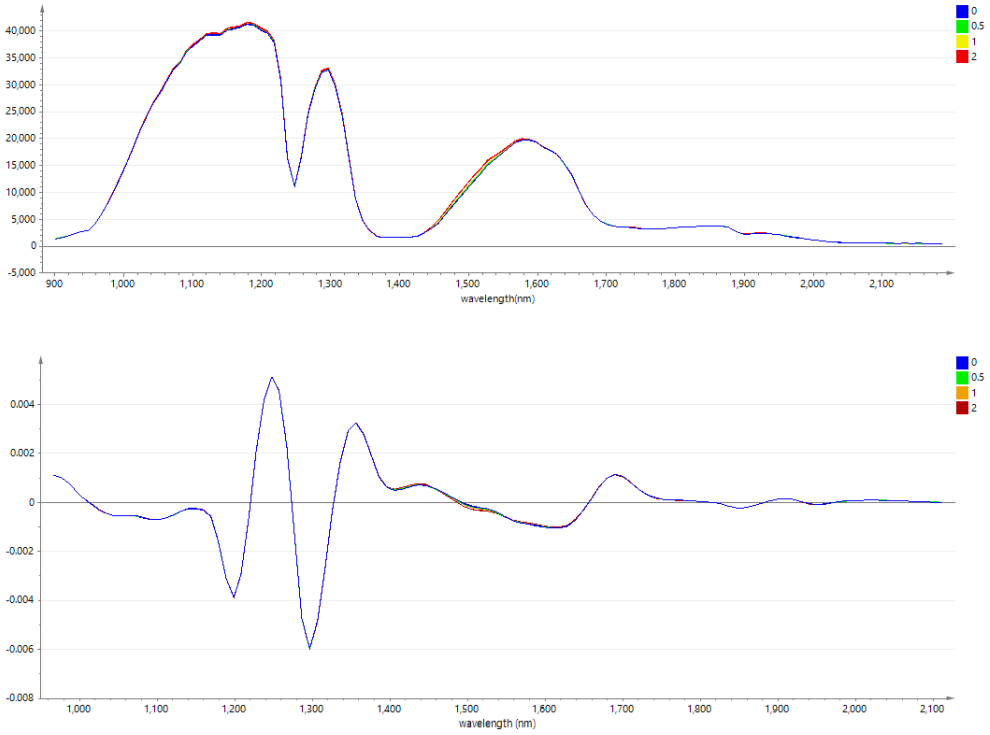

Figure 4.10: Raw spectra of pure samples of lactate (0, 0.5, 1.0 and 2.0 M) (top plot) and pretreated with SNV, second derivative and Savitzky-Golay filter (bottom plot).

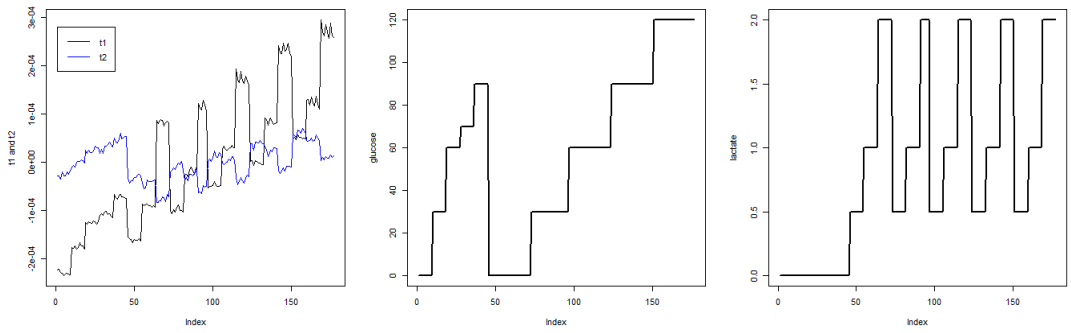

Figure 4.11: Plots ordered by solutionID (x axis): plot A shows the two components of the PLS model(black line is first LV, blue line is second LV), plot B shows the content of glucose and plot C shows the content of lactate. All samples were randomly prepared.

Predictions of Glucose in Different Datasets

In a normal scenario the developed model would be used to predict glucose concentration in similar samples. However, the aim of this section is to challenge the model with analytes outside the calibration dataset. If qualitative information can be obtained at this point, it could indicate that some information can also be obtained from fermentation matrices.

Prediction of a Dataset of pure samples of Glucose and Yeast

Firstly, a binary dataset of glucose and yeast was used and the predictions are plotted in Figure 4.12. Cells have a strong effect on the spectral data as they could cause a change of pathlength by reducing the transmitted light through the sample. Employed pre-processing (SNV) should minimize some of these effects. In fact, a similar error of prediction was obtained for glucose prediction of this dataset (7.9g/L) to the error of estimation (RMSEE) of 6.7g/L. The "observed" values in the plot were not measured. It is assumed that the prepared stock of cells does not change during the experiment and the system is able to mix the suspension and accurately measure a sample. Thus, named "ground truth".

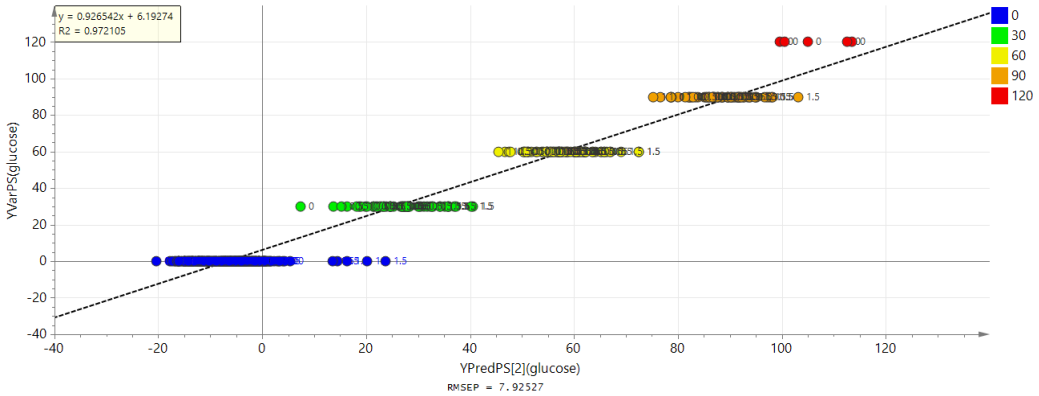

Figure 4.12: Ground truth vs Prediction of glucose in a binary dataset of Glucose and Yeast (up to 1.5g/L). The points are labelled by concentration of yeast (in g/L) and coloured by glucose concentration (in g/L and according to scale shown in the plot).

Prediction of Dataset of pure samples of Glucose and NaCl

The selected model was used to predict pure samples of glucose and pure samples

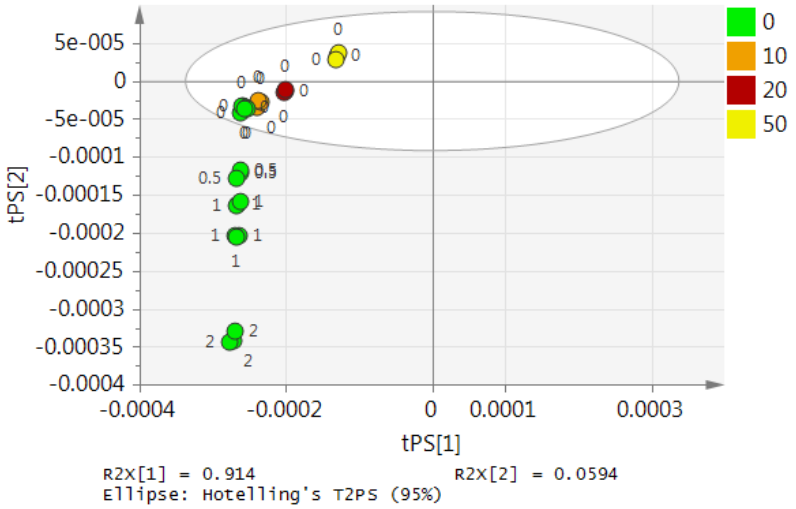

Figure 4.13: Projection of pure samples of glucose (0, 10, 20 and 50 g/L) and NaCl onto the scores plot of the selected model. Scores are coloured according to the glucose concentration and labelled with the percentage of NaCl in the mixture.

of NaCl in water, as shown in Figure 4.13. The spectra of these new samples were projected onto the scores plot of the glucose model (Figure 4.8). The pure samples of glucose were projected into the same position inside the ellipse as in the original plot.

The pure samples of NaCl are outside of the ellipse which means they are classified as outliers by a model that is based on glucose and lactate. The model's capability to correctly identify such exogenous spectra as outliers proves the specificity of the model as well as the competence of the flow-cell with miniaturized spectrometer. Also, a higher concentration of NaCl translated into a lower level of t2. Even though this salt does not have unique NIR absorption bands, it can cause shifts in the water bands along the wavelength axis.

Model for Lactate

For the selected binary dataset, the investigation for a model to predict the concentration of lactate is shown below in Table 4.3. The best results were obtained through model M6 and therefore its detailed statistics are shown in Table 4.4 and the scores plot in Figure 4.14. The RMSEE obtained was 0.076 M, which corresponds to 6.8g/L.

Table 4.3: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for lactate concentration (M), and the R2 of the standardised residuals.

Model	Pretreatment	RMSEE	RMSEcv	R-squared
M1	Raw spectra	0.117	0.118	0.984
M2	1der/SG	0.116	0.117	0.987
M3	2der/SG	0.107	0.107	0.973
M4	MSC/1der	0.105	0.106	0.977
M5	MSC/2der	0.0765	0.0766	0.994
M6	SNV / 2der / SG	0.0762	0.0762	0.994

Table 4.4: Model statistics for the PLS model for lactate concentration (in M6), based on 59 samples (177 scans).

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (M)	RMSEcv (M)
1	0.921	0.833	105	0.832	0.0762	0.0762
2	0.0525	0.157	5.99	0.935		
Cumulative	0.974	0.989	--	0.989		

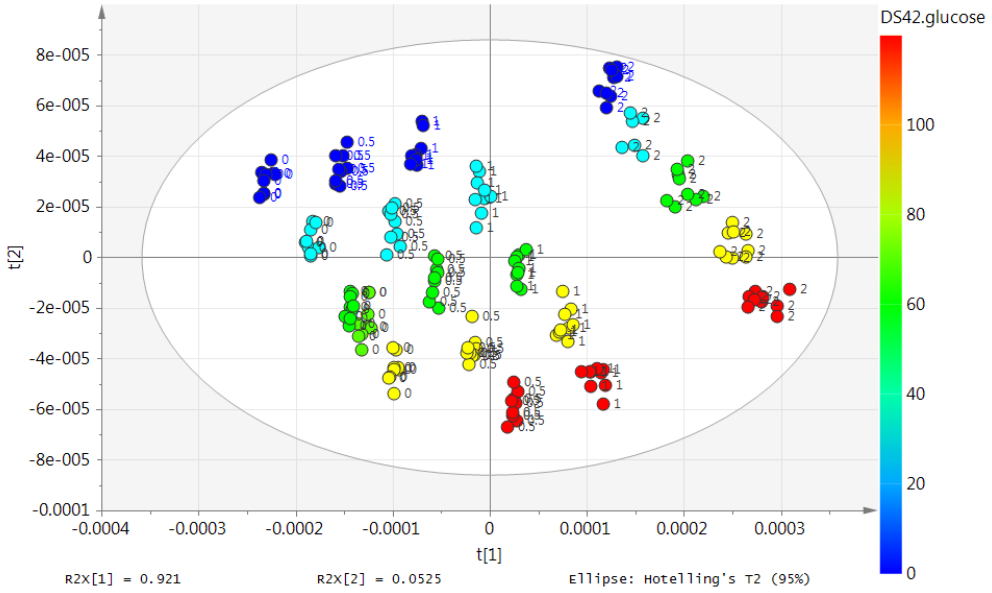

Figure 4.14: Scores plot from the selected PLS model for lactate (M6); scores are labelled with concentration of lactate (M6) and coloured by glucose concentration (in g/L and as shown by the legend scale in the plot).

Predictions of Lactate in Different Datasets

The selected model for Lactate developed in the binary dataset, was used to predict lactate in different settings.

Prediction of Lactate in Binary dataset of Lactate and NaCl

The selected model was used to predict the lactate content in samples from a dataset of lactate and NaCl in water (collected between the 10th and the 12th of February), and the predicted scores projected into the scores space of the model (as shown in Figure 4.14)are shown in Figure 4.15. The fact that outlying the samples with NaCl are located outside the ellipse proves the specificity of the model to describe glucose and lactate, while assuring that the effect being modelled is not uniquely the decrease of water signal. Figure 4.16 shows the observed vs predicted and a RMSEP of 1.3M was obtained for this dataset. Pure samples of lactate (plotted in dark blue) are accurately predicted; the presence of NaCl in the mixture affects the predictions of lactate.

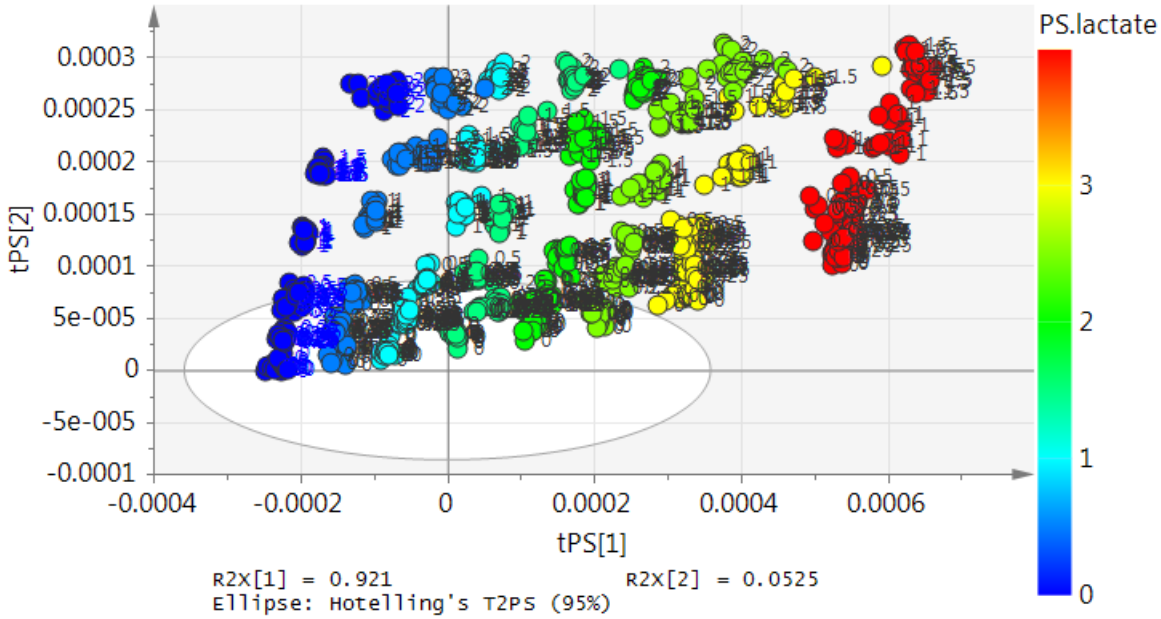

Figure 4.15: Projection in scores plot from Figure 4.14 of mixtures of Lactate and NaCl. The scores are coloured by Lactate concentration (according to the scale shown in the plot) and labelled by NaCl concentration.

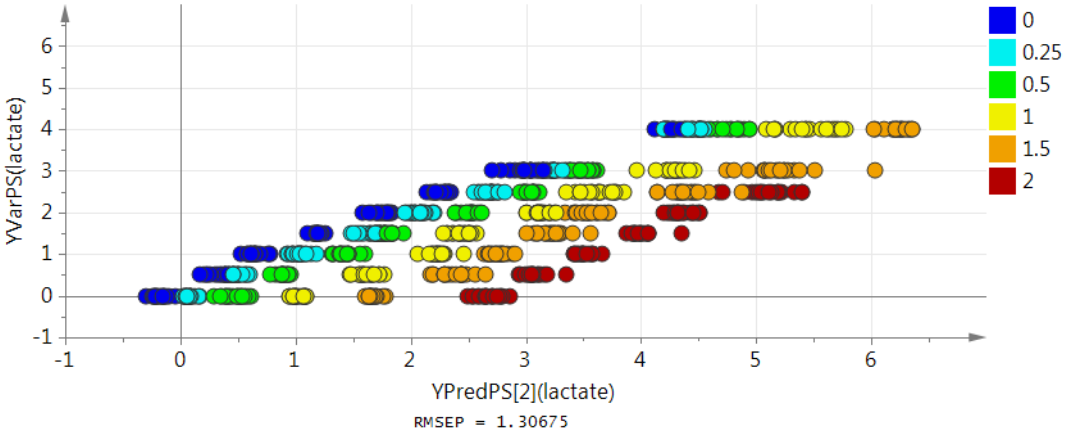

Figure 4.16: Ground truth vs Prediction of lactate in a binary dataset of Lactate and NaCl (up to 1.5M). The points are coloured by NaCl concentration (in M and according to scale shown in the plot).

Prediction of Lactate in Ternary dataset of Lactate, Glucose and NaCl

A ternary dataset of lactate, glucose and NaCl was predicted by the developed model for Lactate. The projected samples are shown in the scores plot (Figure 4.17).

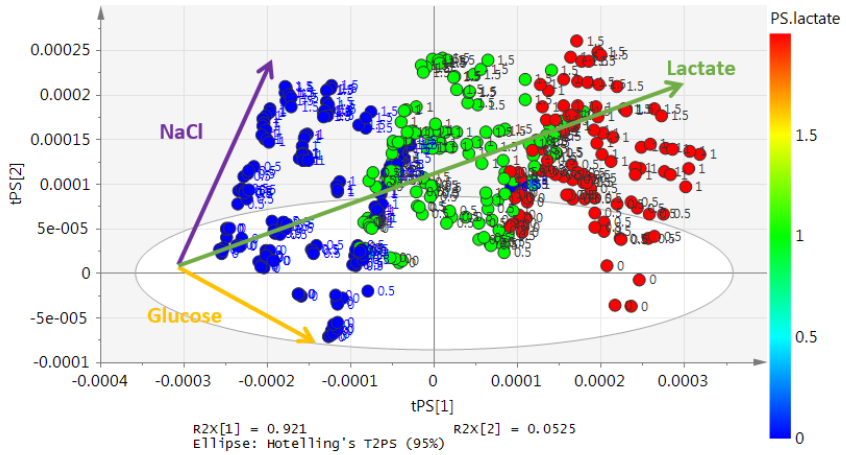

Figure 4.17: Scores plot from the selected PLS model for lactate (M6); scores are labelled with concentration of NaCl salt (M) and coloured by lactate concentration (in M and as shown by the legend scale in the plot).

4.3.1.2 Dataset Glucose x Yeast

Datasets of glucose and yeast were prepared on the dates of 28th February, 1st March and 2nd March. A selection of these samples was used for model development. Five scans were obtained per sample prepared by the automated liquid handler, at 25ms of integration time.

Model for Glucose

For the selected dataset the investigation for a model to predict the concentration of glucose is shown below in Table 4.5. The lowest errors were obtained through model M6, for which a total of 21 samples were used (105 scans). These samples were collected on the 28th February.

The spectral pre-processing techniques ranged from derivatives (1st, 2nd) combined with MSC and SNV. Scatter corrections (SNV and MSC) were used to minimize potential differences in pathlength caused by the presence of yeast in the samples. The best RMSEP were obtained through model M5 obtained after wavelength selection. The detailed statistics of are listed in Table 4.6 and the scores plot in Figure 4.18.

Table 4.5: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for glucose, and the R2 of the standardised residuals.

Model	Pretreatment	RMSEE	RMSEcv	R-squared
M1	Raw spectra	3.51	3.53	0.989
M2	1der/SG	3.77	4.03	0.918
M3	2der/SG	3.93	4.17	0.934
M4	MSC/1der	3.55	3.72	0.918
M5	MSC/2der	3.89	3.91	0.922
M6	SNV/2der/SG	3.08	3.12	0.983

Table 4.6: Model statistics for the PLS model for glucose concentration (in g/L), based on 30 samples (149 scans) from 28th February.

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.774	0.777	27.8	0.771	3.37	3.47
2	0.220	0.211	7.92	0.221		
Cumulative	0.994	0.989	--	0.988		

Predictions of Glucose in Different Datasets

The selected model was used to predict a similar dataset, prepared on a different day and the predictions are plotted (x axis) against the actual values (Y axis) in Figure 4.19. The high R-squared and the error of prediction similar to the RMSEcv demonstrate the robustness of the model to external datasets, without extrapolation.

Model for Yeast

The increase of biomass content increases light scattering and causes a baseline shift in the absorbance spectra (Tamburini et al., 2003). For this study the spectral data was used in spectral counts. Using equation 4.5, spectra of samples of water and yeast were transformed into absorbance (log 1/T) and plotted in Figure 4.20. The expected effect of biomass on the baseline of absorbance spectra was then clarified. These spectra also revealed the noise level after 2000 nm that most likely arise from the optical fibres. Typically, NIR spectra of aqueous solutions reveal two major broad bands due to OH vibrations found at the region between 1450 nm and 1950 nm, as

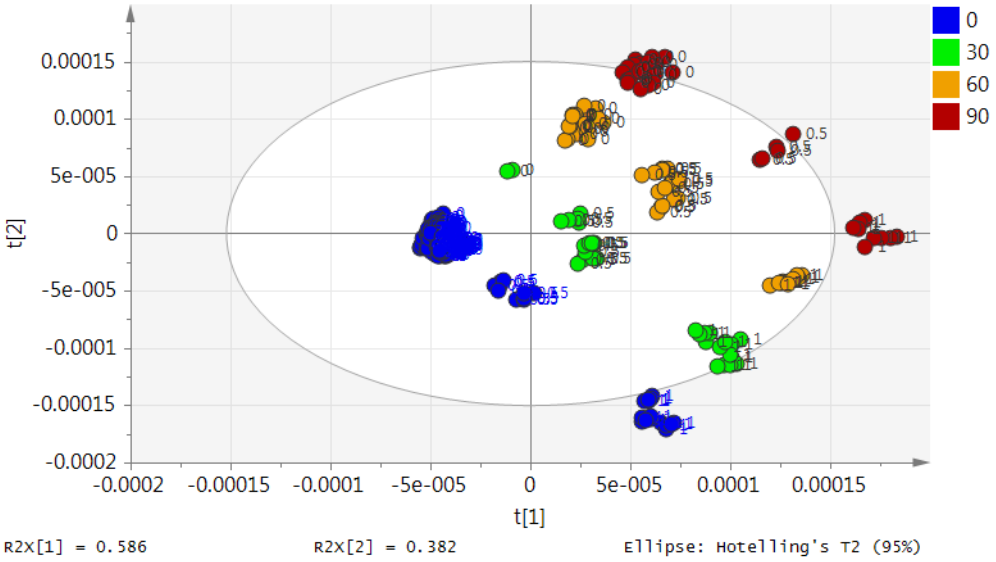

Figure 4.18: Scores plot from the selected PLS model for Glucose; scores are labelled with concentration of yeast (g/L) and coloured by glucose concentration (in g/L) and as shown by the legend on the plot. A total of 11 mixtures and water samples were randomly prepared and scanned over the course of a day.

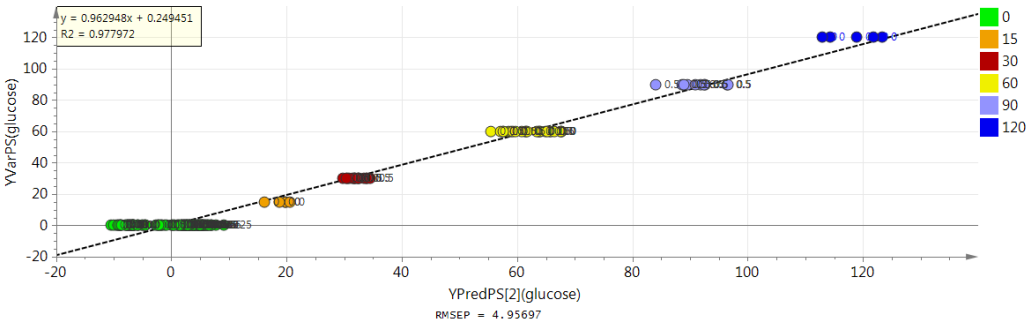

Figure 4.19: Prediction of glucose concentration (g/L) in mixtures of Glucose and Yeast, by the model described on Table 4.6, against the actual values of glucose (Y axis). Points are coloured by glucose concentration (in g/L) and as shown by the legend on the plot. An error of prediction of 5g/L was obtained for this external dataset.

seen in Figure 3.7. The main peak in these spectra should therefore be related to water and perhaps the second bond was not detected by this system. The visible peak in the absorbance spectra centred at about 1650nm corresponds to a region of C-H vibrations, which could be originated by biomass (Vaidyanathan et al., 1999).

For selected samples from the described datasets, the investigation for a model that quantifies yeast is described on Table 4.7. To build a model for quantification

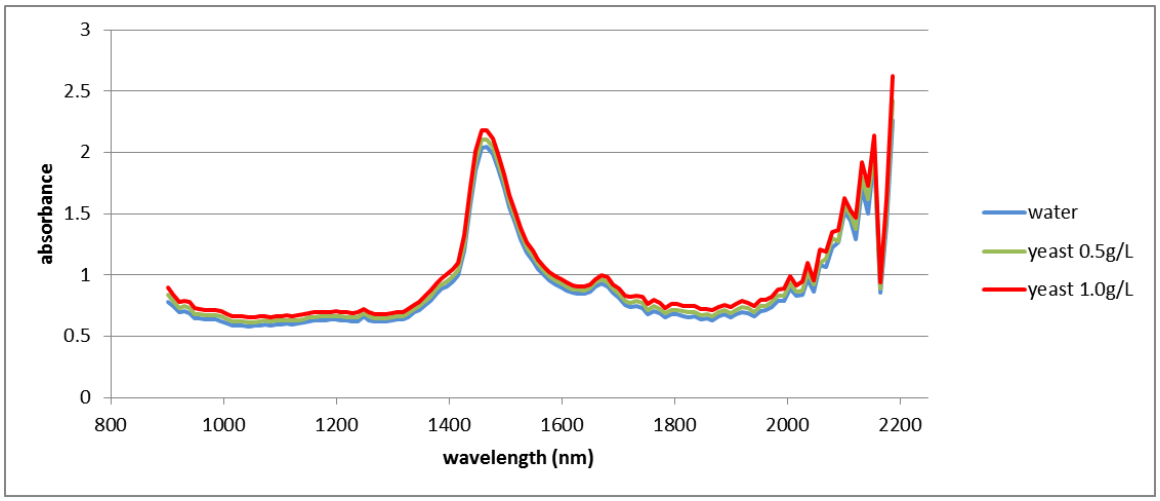

Figure 4.20: Spectra in absorbance units for samples of water (blue line), a mixture of yeast 0.5g/L (green line) and a mixture of yeast 1.0g/L (red line). Solutions of higher concentrations of yeast show higher absorbance values.

of biomass, there are two possible methodologies. One can choose to minimize the scattering effect caused by the cells and apply corrections such as MSC or SNV to capture the chemical variability (which should be mainly protein content) introduced by the cells. Another option is to capitalise on the physical information in the spectra by using the effects on baseline. Therefore, the raw spectra could also be used to track biomass content in the broth. All developed models presented in the table have shown similar values of RMSEEcv of approximately 0.05g/L of yeast. This indicated that the measuring system can detect yeast as low as 0.15g/L. For all models, only one component (rank) was to capture the majority of the variance in the data. The value for each model is also listed in Table 4.7.

As all models yielded very similar results, an external dataset was used to challenge the models and take a decision based on the RMSEP. Table 4.8 lists the error of prediction of yeast content for the same dataset predicted by the different models. For this purpose, samples from the dataset created on the 28th February, with yeast content of up to 1.0g/L and 110 scans (22 samples, each sample scanned 5 times), was used. Models M2 and M3 showed the same value of error of prediction (RMSEP). However, a better correlation between the expected values and the predicted values was obtained through model M2 with an R-squared value of 0.98, against 0.97 obtained for model

Table 4.7: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for yeast (in g/L), and the R-squared of the standardised residuals. Only one component was needed for each model.

Model	Pretreatment	N. Obs. used	RMSEE	RMSEcv	R-squared
M1	Raw spectra	290	0.0505	0.0504	0.968
M2	1der/SG	290	0.0480	0.0479	0.972
M3	2der/SG	290	0.0521	0.0519	0.977
M4	MSC/1der/SG	290	0.0580	0.0580	0.978
M5	MSC/2der/SG	290	0.0472	0.0472	0.980

M3.

Table 4.8: Evaluation of the models for yeast concentration, listed in the previous table, on the prediction of a dataset from 28th February. Both RMSEP and the coefficient of determination (R-squared) are presented for each model. Model M2 performed the best.

Model	Pretreatment	RMSEP	R^2
M1	Raw spectra	0.118	0.972
M2	1der/SG	**0.116**	**0.977**
M3	2der/SG	0.116	0.970
M4	MSC/1der/SG	0.121	0.956
M5	MSC/2der/SG	0.117	0.971

An investigation for selection of specific wavelengths was carried out. However, no improvements were on the predictions were obtained and therefore was not described here in detail. The statistical analysis for the selected PLS model for yeast's concentration (model M2) are listed in Table 4.9. Predictions for mixtures prepared on a different day are plotted in Figure 4.21. All samples were slightly under-predicted in comparison to the expected value, though qualitatively correct.

Table 4.9: Summary of the PLS model for yeast concentration (in g/L), based on 58 samples (290 scans) from 1st and 2nd March.

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.991	0.983	109	0.982	0.048	0.048

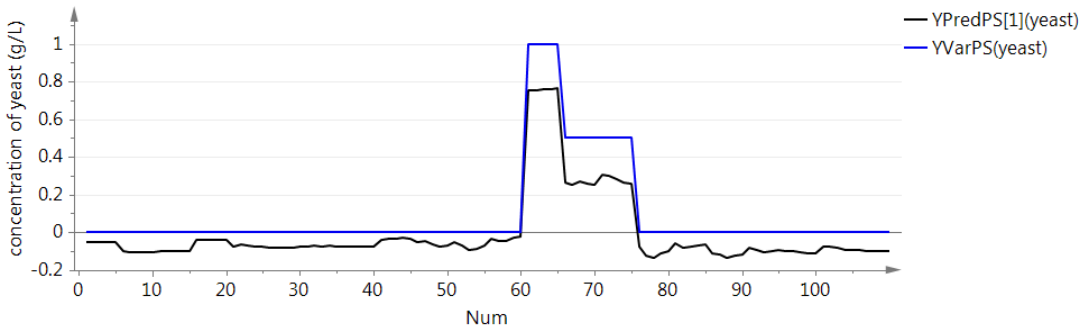

Figure 4.21: Prediction of yeast concentration by model M2 (black line) and expected values of yeast concentration for each sample (blue line). All samples were underpredicted in comparison to the expected value, while the three levels of yeast content were correctly described.

4.3.2 Ternary Mixtures

Mixtures of three analytes were prepared using the automated system developed. Models were developed for glucose and lactate in buffer and fresh media though more ternary mixtures were prepared.

The performance of the models was tested on a similar dataset that was prepared on a different day, with other ternary mixtures and binary mixtures. This allowed for quantification of the effect of time periods and various conditions on the spectroscopic measurements and stability of the system.

4.3.2.1 Dataset Glucose x Lactate x Buffer

Model for Glucose

A PLS model for glucose in the presence of lactate and buffer was developed and the details are listed in Table 4.10. This dataset was prepared through means of automation from the 17th to the 18th of February.

Table 4.10: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for glucose, and the R-squared of the standardised residuals, for the dataset of "Glucose x Lactate x Buffer".

N calibration set	222
Pre-processing	SNV and 2nd derivative (11,1) SG-filter 5
Selected wavelengths	All available after pre-processing: 968-2111.2 nm
Number of LVs	3
R^2(X)	LV1: 61.8% LV2: 19.4% LV3: 15.3% Cumulative: 96.6%
R^2(Y)	LV1: 46.3% LV2: 42.2% LV3: 9.36% Cumulative: 97.9%
Eigenvalue	LV1: 70.5 LV2: 22.5 LV3: 17.5
Q^2	LV1: 45.9% LV2: 78.1% LV3: 81.3% Cumulative: 97.8%
R^2 for standardised residuals	0.995
RMSEE (g/L)	6.27
RMSEcv (g/L)	6.56

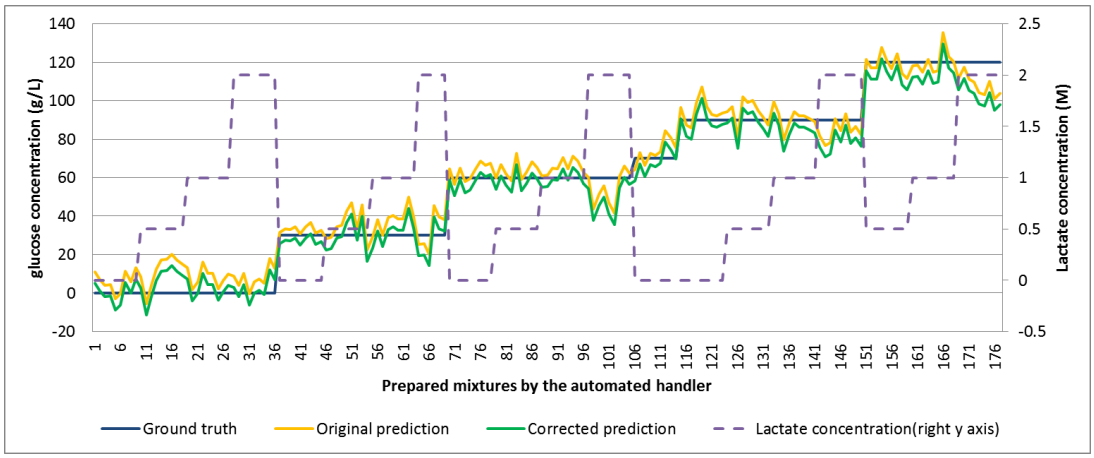

Figure 4.22: Prediction of glucose in binary dataset by current model. Original prediction through the model (orange line), ground truth content for each prepared mixture (blue line) and prediction of glucose concentration after correction of the baseline (green line). The content on lactate for each mixture is plotted by the purple dashed line and the reading is on the right hand side of the plot. The RMSEP for glucose concentration was 8.53g/L before correction and 8.40g/L after correction.

The described model was used to predict glucose concentration in a dataset of glucose and lactate in water. Figure 4.22 plots the prediction of glucose (orange line), lactate concentration in each sample (in dashed purple line, on the secondary axis). The different levels of glucose were accurately predicted, independently of the lactate levels.

A RMSEP of 8.53g/L was obtained, whilst a value of 6.6g/L of RMSEcv was obtained for the calibration samples in buffer. A correction was applied to the predicted values of glucose.

To correct for potential differences between days, and because no reference spectrum was used to correct each spectrum, a "baseline correction" was applied to the predictions themselves. For that, the mean predicted value of the samples with no glucose (0g/L) was subtracted to each obtained prediction. The result is indicated by the green line in Figure 4.22.

Model for Lactate

A PLS model for lactate in glucose and buffer was developed for the same dataset and the details are listed in Table 4.11. The plot of observed *vs* predicted for a different

day is shown in Figure 4.23.

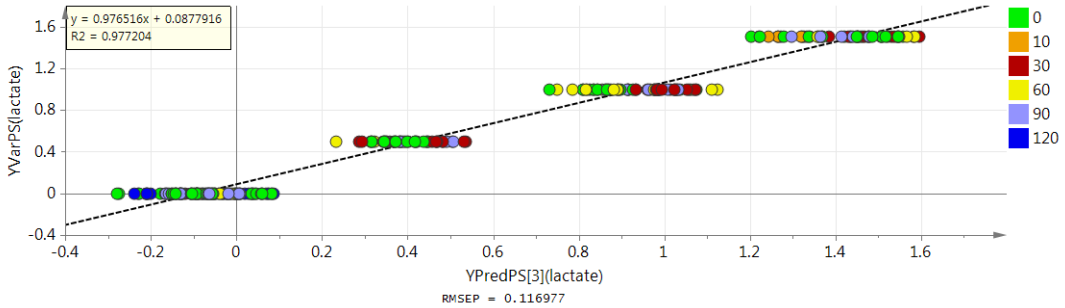

Figure 4.23: Observed vs. predicted for lactate concentration of a different dataset prepared from the 24th to the 25th of February. The RMSEP obtained was of 0.12M. Scores are coloured by concentration of glucose as labelled in the plot.

Table 4.11: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for lactate, and the R-squared of the standardised residuals, for the dataset of "Glucose x Lactate x Buffer".

N calibration set	336
Pre-processing	SNV and 2nd derivative (11,1) SG-filter 5
Selected wavelengths	All available after pre-processing: 968-2111.2 nm
Number of LVs	3
$R^2(X)$	LV1: 0.657 LV2: 0.165 LV3: 0.144 Cumulative: 0.966
$R^2(Y)$	LV1: 0.691 LV2: 0.239 LV3: 0.0553 Cumulative: 0.985
Eigenvalue	LV1: 74.9 LV2: 18.8 LV3: 16.4
Q^2	LV1: 0.689 LV2: 0.769 LV3: 0.789 Cumulative: 0.985
R^2 for standardised residuals	0.990
RMSEE (M)	0.0695
RMSEcv (M)	0.0721

Table 4.12: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for glucose, and the R-squared of the standardised residuals, for the dataset of "Glucose x Lactate x Fresh Medium".

N calibration set	222
Pre-processing	SNV and 2nd derivative (11,1) SG-filter 5
Selected wavelengths	All available after pre-processing: 968-2111.2 nm
Number of LVs	3
R^2(X)	LV1: 69.9% LV2: 18.8% LV3: 8.04%
R^2(Y)	LV1: 47.3% LV2: 41.7% LV3: 8.17%
Eigenvalue	LV1: 79.7 LV2: 21.7 LV3: 9.17
Q^2	LV1: 46.5% LV2: 89.0% LV3: 74.1%
R^2 for standardised residuals	0.977
RMSEE (g/L)	7.40
RMSEcv (g/L)	7.52

4.3.2.2 Dataset Glucose x Lactate x Fresh Medium

A ternary mixture of glucose, lactate and fresh medium was prepared between 16 and 17th of February. The results for model development for glucose and lactate in this setting are described below.

Model for Glucose

A PLS model for glucose in lactate and fresh medium was developed and the details are listed in Table 4.12. The model was tested for the prediction of a similar dataset, though prepared on different days.

The binary dataset of glucose x lactate was used to quantify the influence of fresh

medium in the spectra and an RMSEP of 17.5g/L was obtained (as seen in the plot of "observed vs predicted" shown in Figure 4.24). All samples were underestimated by this model, although linearity was observed at R-squared of 0.98. The samples with zero content of glucose were predicted at an average of -15.48g/L. A correction was proposed for which the predicted average of these samples was used to correct the prediction of other samples (pure samples and binary mixtures in the dataset). The three different lines of glucose concentration are plotted in Figure 4.25: ground truth (or expected) in dark blue, the original prediction in yellow and the corrected prediction in green. Through this method a new RMSEP of 6.09 g/L was obtained, which matches the expected values. The glucose model for binary mixtures yielded a RMSEEcv of 6.65 g/L (as seen in Table 4.2) while the model for glucose based on the ternary mixtures was 7.4g/L (as seen in Table 4.12).

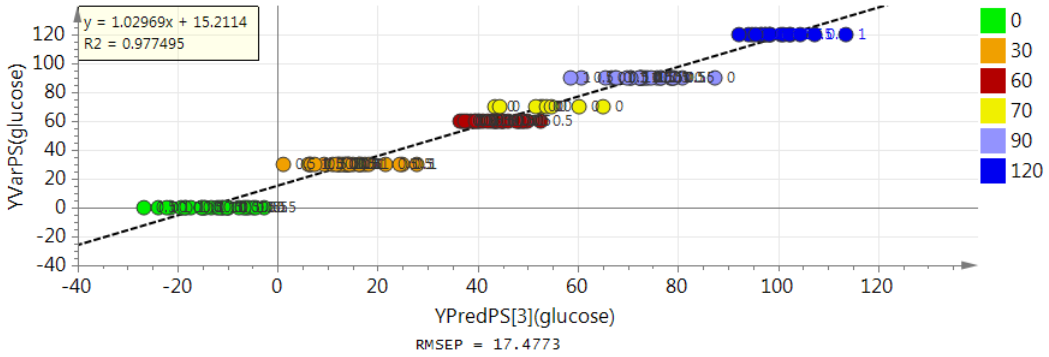

Figure 4.24: Linearity of the prediction of the binary dataset "glucose x lactate" through the model based on ternary mixtures of "glucose x lactate x medium". The R-squared is satisfactory but a shift of 15.2g/L is detected.

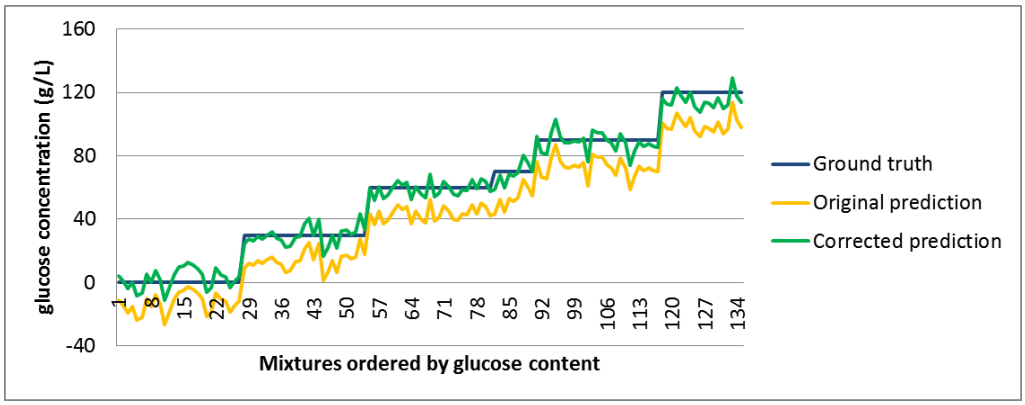

Figure 4.25: Prediction of binary dataset. Original prediction through the model (orange line), ground truth content for each prepared mixture (blue line) and prediction of glucose concentration after correction of the baseline (green line). A RMSEP of 6.09g/L is then obtained.

Model for Lactate

A model for lactate prediction in the dataset of "Glucose x Lactate x Fresh Medium" was developed and the summary of the statistical analysis is shown in Table 4.13. As done in the previous section, a binary dataset of glucose and lactate in water was predicted by this model. The concentration of lactate was over-predicted by the PLS model based on ternary mixtures. The predicted average of samples without lactate was used to correct the other samples and the results are shown in Figure 4.26. A RMSEP of 0.106M was obtained after correction.

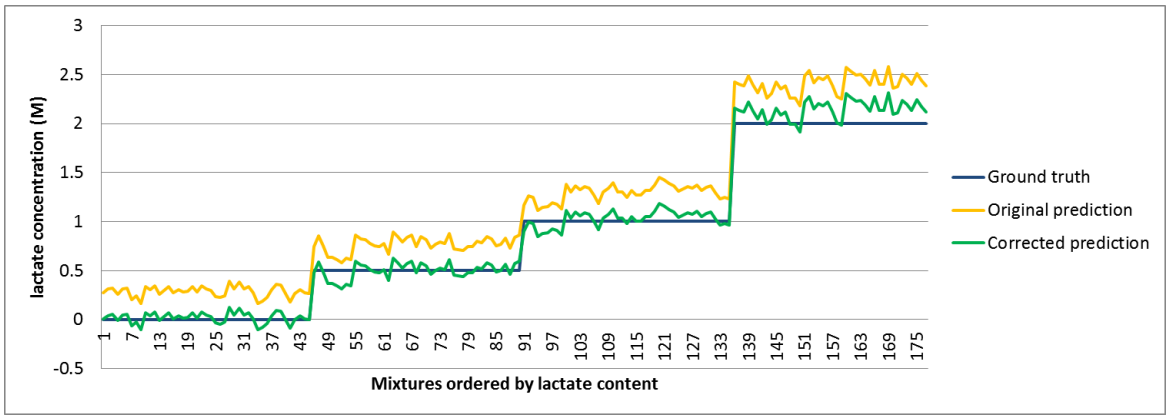

Figure 4.26: Original prediction through the model (orange line), ground truth content for each prepared mixture (blue line) and prediction of lactate concentration after correction of the baseline (green line). A value of RMSEP of 0.106M was then obtained.

Table 4.13: Collection of the errors of estimation (RMSEE/RMSEcv) for the PLS models developed for lactate, and the R-squared of the standardised residuals, for the dataset of "Glucose x Lactate x Fresh Medium".

N calibration set	222
Pre-processing	SNV and 2nd derivative (11,1) SG-filter 5
Selected wavelengths	All available after pre-processing: 968-2111.2 nm
Number of LVs	3
R^2(X)	LV1: 73.2% LV2: 15.9% LV3: 7.66% Cumulative: 96.8%
R^2(Y)	LV1: 67.9% LV2: 26.3% LV3: 4.27% Cumulative: 98.4%
Eigenvalue	LV1: 83.5 LV2: 18.1 LV3: 8.74
Q^2	LV1: 67.7% LV2: 81.5% LV3: 72.6% Cumulative: 98.4%
R^2 for standardised residuals	0.997
RMSEE (g/L)	0.0712
RMSEcv (g/L)	0.0736

4.4 Discussion

The application of automation into calibration development for spectroscopy can be highly beneficial. The procedures that in the previous chapter took about three days' work were now outperformed by automation simply with an overnight run of mixtures' preparation.

In addition, manual mixture preparation is prone to variations in how the samples were measured, dispensed into a preparation vessel and mixed. The presented PCA scores plot have demonstrated the ability of the developed system for replicability since the scores from the mixtures were placed in the expected position relatively to the scores originated from the pure samples. This indicated that each required stock was accurately aspirated with the syringe pump. The suggested methodology for mixing the sample was efficient as well as the cleaning procedure between measurements.

The results also exhibited the capability of the miniaturized spectrometer used in integration with the flow-cell. However, the errors of prediction were not satisfactory (as seen in Table 4.14) as ideally should be below 1g/L.

Table 4.14: Summary of errors of estimation for glucose models developed in the different datasets.

Dataset	**RMSEE**	**RMSEcv**
Glucose x Lactate	6.65	6.62
Glucose x Yeast	3.08	3.12
Glucose x Lactate x Buffer	6.27	6.56
Glucose x Lactate x NaCI Salt	15.1	15.4
Glucose x Lactate x Medium	7.40	7.50

There are various reasons why these results were less positive than perhaps anticipated. First, the spectral collection might have to be further optimized. An integration time of 25ms was selected but there might be the need to obtain more averages for a single spectrum. Usually one single spectrum already results from co-added scans (e.g., 32, 128) (Riley et al., 2001; Crowley et al., 2005; Sampaio et al., 2014), which should improve the signal-to-noise ratio. Also, the pathlength used in the flow-cell was approximately 0.5mm. In comparison, the settings used in the previous chapter consisted

of 32co-added scans per spectra and an actual measuring gap of 1mm was used in the immersion probe (section 3.2.5.2). Given the strong absorption of water, the use of NIR spectroscopy to characterise biological environments could be limited (Jensen et al., 2003). The penetration depth is reduced by the water absorption, thereby, reducing the signal-to-noise ratio (SNR). Also, for the fact that the spectra were used as "counts", instead of corrected absorbance (Eq. 4.5), it means that there was no correction applied for potential differences involving the background, temperature, state of the instrument.

Despite the relatively high error of predictions, the predicted trends were valid. This gives useful qualitative information that can be used when there is no need for tight control of the analyte levels. In addition, the models for glucose and yeast in a binary dataset both showed relatively low errors of estimation. These were therefore selected as the models to be tested further with cultivation samples in the next chapter.

4.5 Conclusions

The aim of this chapter was to develop a capable NIR-based measuring system that can then be used by R&D scientists. Traditionally, the methodology to obtain a working model based on NIR is labour-intensive. One of main challenges for the implementation of spectroscopy in early stages of process development is the fact that the process itself is not yet well defined, which complicates the required design space.

While developing such a measuring system, this chapter also investigated potential calibration strategies that could minimize the efforts required by the end-user to obtain valid predictions of the analytes of interest. Lastly, this chapter highlighted the advantages of combining spectroscopy with automation by demonstrating the quick, repeatable outcome typical of automation.

To achieve these goals, a miniaturized NIR spectrometer in combination with an automated liquid-handler module (named AM, which consists of a syringe pump, pump valves, dispensing wheel and a vessel) and a prototype flow-cell were developed.

To operate this new setup, this chapter also presented a guideline for the set-up and manipulation of this automated system. The necessary procedures, here developed in a user-written script, would then be integrated in a more user-friendly software for the end user.

The chapter went on to demonstrate the power of automation and the AM specifically, to collect a large amount of mixtures. Without human intervention and therefore human error, mixtures were dispensed to the flow cell, spectra were collected, the liquid was discarded to waste and a cleaning procedure was executed. Such dataset can easily be prepared overnight, and a valid model can be made available for the fermentation samples in the following day.

The consequent exploratory data analysis used binary and ternary mixtures as datasets to calibrate predictive models. Individual PLS models for Glucose and Lactate were developed (and tested) in both settings and their predictive ability tested. Ideally, the obtained models would be able to give some valid information on more challenging matrices, such as a sample from a fermentation (which was then tested in the next

chapter).

This relatively simple analysis provided a wealth of data. It was shown that initial results were highly replicable, with low variation between them. In addition, within the same datasets, predictions were relatively accurate. It proved harder, however, to extrapolate models to samples outside the calibration dataset.

Various reasons were presented in the discussion as to why the errors of estimation were high. Nonetheless, the predicted trends were valid, providing useful qualitative information that can limit the calibration efforts of the end-user of such an automated system. In the end, a binary mixture dataset was selected as optimal calibration dataset that will be used to predict fermentation samples in the next chapter.

Chapter 5

Realisation of NIR in Automated Bioreactor Systems

5.1 Introduction

Bioprocess development is a time, cost and labour-intensive task, which aims to understand the interactions between process parameters and product quality. To determine the best performing microbial strain and optimal cultivation conditions, a large number of small-scale experiments has to be executed. This number is ever increasing with advanced methods in genetic engineering and molecular biology being able to generate thousands of different clones. With pressure to cut costs and reduce the time-to-market of a biotechnological product, microbial cultivations have to be performed ever faster. Bioreactor miniaturization can unlock the benefits of high-throughput screening and optimization of fermentation parameters (Rowland-Jones et al., 2017; Kusterer et al., 2008; Funke et al., 2010; Velez-Suberbie et al., 2017). Through the means of automation, single-use bioreactor systems can deal with volumes in a highly accurate, reproducible, and sterile manner (Xu et al., 2017). A fully automated bioprocess permits round the clock exploitation, high levels of precision and is an effective tool for detection and further evaluation of events. Socially, the effect of automation is also relevant as it allows less interesting tasks to be performed more efficiently and reliably by machines, whilst the scientist can be involved in more useful work.

The capabilities of process analytical technology, and specifically spectroscopic methods, for monitoring relevant analytes of bioprocesses have been discussed in the literature review and reinforced throughout the previous chapters. Challenges lie in living organisms being one of the raw materials, the complex matrix background and low concentrations of the metabolite of interest. In the particular case of miniature bioreactors, the low volume of the vessels is also a constraint, for example, as it limits the volume of samples to be collected to build a calibration dataset through the use of offline measurements. Overall, as already mentioned in chapter 2, the challenges of PAT implementation in Ambr® systems are threefold: calibration, integration and cost.

Calibration is the first step involved in the development of a model and, likewise, for NIR model. With this purpose, known data are used to determine the relation-

ship between the observed spectral variation and the corresponding levels of relevant components, i.e. analytes, in the medium inside the bioreactor. In later stages of the bioprocess development, the process is largely consistent and reproducible and therefore, calibration methodology for spectroscopic measurement can be relatively straightforward. By contrast, at the early stages of bioprocess development this is not the case. Many experiments need to be conducted as the bioprocess itself is not yet defined. Therefore, novel approaches that do not rely on already established processes are essential, solutions that function across a range of conditions without extensive calibration methods.

Depending on the configuration used, the level of difficulty in the integration of a spectroscopic instrument together with the Ambr® system in the laminar flow hood (where it is placed to allow aseptic operations), might vary but it certainly poses a challenge.

The implementation of spectroscopy instruments in the Ambr® would increase its price and would have to be justified. At large scale, the culture is valuable and spectroscopic approaches can be economically viable.

Some of these issues were addressed by Musmann et al. (2016) and Rowland-Jones et al. (2017). Musmann et al. (2016) have reviewed the applicability of spectroscopic techniques for miniature bioreactors and concluded that infrared and Raman spectroscopy are preferable in view of their greater range of applications in combination with their suitability for high-throughput systems. Additionally, the work from Rowland-Jones et al. (2017) took this further and three spectroscopic methods (NIR, Raman and 2D-fluorescence) were experimentally tested to evaluate their ability to monitor multiple analytes under suitable operational constraints required for mini-bioreactor cultures. The main findings of this study were that:

- 2D-fluorescence measured ammonium concentration more accurately than Raman and NIR.
- Raman spectroscopy was more robust at measuring lactate and glucose concentrations in cell culture supernatant (RMSEcv 1.11 and 0.92 g/L, respectively) than the other two techniques. The findings suggest that Raman spectroscopy is more

suited for this application than NIR and 2D-fluorescence. The implementation of Raman spectroscopy increases at-line measuring capabilities, enabling daily monitoring of key cell culture components within miniature bioreactor cultures.

Even though Raman spectroscopy yielded better results than NIR (RMSEcv of 2.33 g/L for lactate and 1.53 g/L for glucose), in this study, this was not always the case. In the study by Trunfio et al. (2017), these three spectroscopic methods were also compared and NIR was found to perform the best. The fact is that spectroscopic measurements are not yet a reality at the early stages of bioprocess development. Even though, there is an increase on the intention of biopharmaceutical companies to implement PAT that is 9.3% more in 2017 than in 2016 (Langer 2017). One of the main reasons for the postponement of this process is the insufficient people in-house to manage the implementation (Langer, 2017).

As stated in the introduction, this thesis aims to bridge the capabilities of automated liquid handlers with miniature bioreactors and NIR spectroscopy, unlocking the use of PAT tools on a daily-routine and therefore facilitating the access to information for the end-use. The concretisation of spectroscopy in the early stages of bioprocess development using automated single-use miniaturized bioreactors and PAT would represent a breakthrough in bioprocess development, improving productivity and shortening development cycles, ultimately resulting in cheaper drugs reaching the clinic faster with potentially improved safety and efficacy.

The previous chapter has shown the development of offline models for the key biochemical parameters (glucose and biomass concentration) by making use of an automated liquid-handler to feed the sample to the NIR instrument.

This chapter would discuss the feasibility of the implementation of an NIR instrument in automated bioreactors system such as the Ambr® system, using three different methodologies: 1) pre-calibrated models (presented in chapter 4) used in the prediction of samples collected from cultivations in the Ambr®; 2) models developed on current standard solutions used for calibration development; and 3) models calibrated with samples collected from the Ambr. After presenting the outcome of those methodologies, the discussion takes stock of important lessons that can be drawn for future

implementation in the form of a roadmap for the integration of the NIR in the Ambr bioreactors.

5.2 Materials and Methods

The end goal was to find the best methodology to obtain real time information on the analytes in the medium of fermentation runs in a system such as the Ambr®15. For this, fermentations of baker's yeast (*S.cerevisiae*) were performed in the Ambr®15f system and samples were collected. Given the existing correlation between glucose and other analytes in the medium, spikes of high concentration of glucose were used to break these correlations. These prepared mixtures were then scanned in the NIR set-up previously described and offline readings of glucose were also obtained.

Minibioreactor system – Ambr® 15f

All cultivations were performed in an Ambr®15f system (shown in 5.1), from Sartorius Stedim Biotech (Royston, UK). The Ambr®15f consists of 24 single-use minibioreactors (each with a working volume of 10-15mL) for fermentation processes that are arranged in two sets of 12 (culture station 1 and culture station 2). Each culture station (CS) has independent controlled background air flow, temperature sensor and stirring plate for agitation control. Each vessel (shown in Figure 5.2) has its own heater for fine temperature control and it is equipped with an agitator with one Rushton-like impeller (11 mm diameter), a pH sensor spot and a DO sensor spot at the bottom of the vessel, a sparge tube and a fluid supply tube (for acid and base addition, for example). The Ambr®15f system is programmed to take pH and DO readings every 12s, and these readings can be used to control pH and DO. The pH spots are able to measure between 6 and 8. A liquid handler (LH) is used to extract culture samples and to perform automated liquid additions using disposable pipette tips. To maintain aseptic operations, the system is placed inside a biological safety cabinet.

Figure 5.1: The Ambr®15 fermentation system (on the left) and the disposable-bioreactor vessel (on the right) which mimics a classical lab scale reactor.

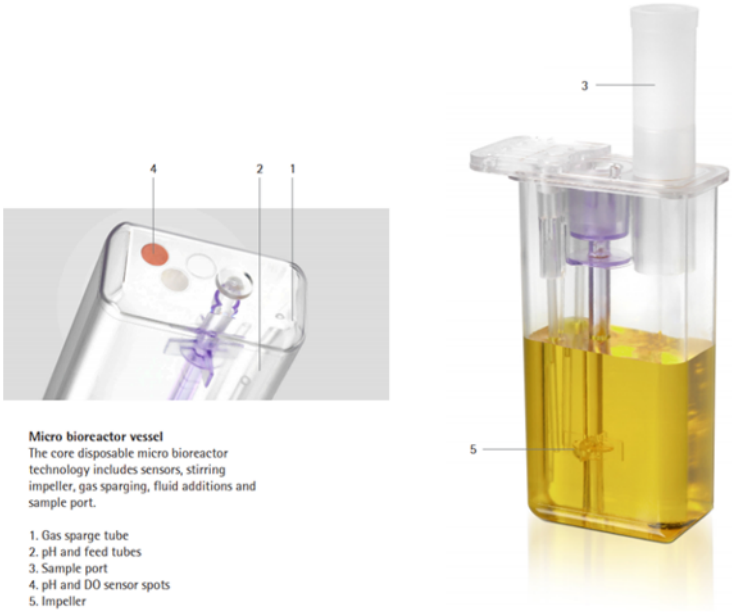

Figure 5.2: Details of the Ambr®15f vessel: gas sparge tube (1), pH and feed tubes (2), sample port (3), pH and DO sensor spots (4) and impeller (5).

Bioprocess and Sampling Methodology

To facilitate the manipulation of the organism and since the end goal was to demonstrate the quantification of glucose concentration, commercial dried baker's yeast (*S.cerevisiae*; dried baker's yeast, Tesco, UK) was used. A solution of 30 g/L of washed cells was prepared and stored at 4°C. To prepare this suspension, the appropriate amount of dried yeast was rehydrated and made up to 100 mL of deionised water. After that, these solutions were centrifuged at 3000 rpm for 10min, the supernatant decanted, the cells were resuspended and then vortexed to ensure good mixing. This washing procedure was repeated two times.

A basic *S.cerevisiae* medium was used, based on the procedures by Finn et al. (2006), consisting of $(NH4)_2SO_4$, 15.0 g; KH_2PO_4, 8.0 g; $MgSO_4$, 3.0 g; $ZnSO_4$, 0.4g per litre. This solution was autoclaved, and the pH adjusted to 6.5 with ammonium hydroxide. All chemicals were purchased from Sigma-Aldrich unless otherwise noted. A D-(+)-Glucose (G8270-5KG, Sigma Aldrich) stock solution of 400g/L was filtered sterilised and used according to the desired concentration for each run.

Process conditions

The standard process setpoints were: temperature at 30°C, aeration of 1.5vvm (15mL/min), with no control of pH or DO implemented. The stirring speed started at 1000 rpm and was increased to 1500 rpm, or to a maximum of 2000 rpm, in those cases where the DO decreased below 20%. The initial working volume was 15mL for every vessel and 10% of inoculum was used.

Sampling Methodology

Each sample collected from each vessel was scanned in the NIR straight after collection. The sampling methodology is described on Figure 5.3. Two vessels were sampled per time point and a maximum of three sampling time points were performed per run. A manual sample of 2mL was collected per vessel. This sample was then split into

two: 1mL was used with no further processing and 1mL was centrifuged and its supernatant collected. Both full sample and supernatant were used to prepare the five different mixtures described in Table 5.1: "raw", "gluc1", "gluc2", "gluc3" and "diluted". These mixtures were prepared with the goal of systematically changing the trajectory of glucose consumption, avoiding correlations between glucose and yeast.

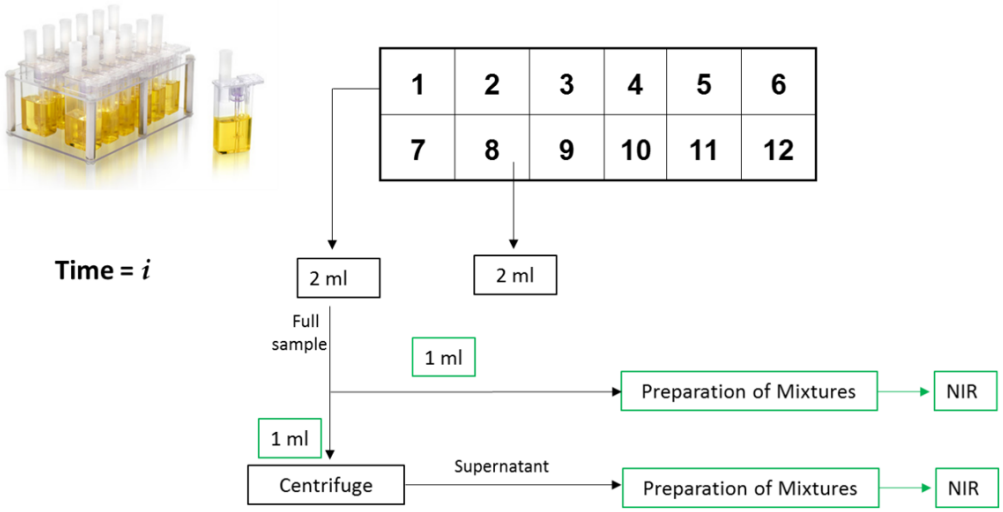

Figure 5.3: Methodology followed for each sample (2mL) collected from one vessel from the Ambr®15f plate. A volume of 1mL is centrifuged and the supernatant collected and used for preparation of mixtures which are then scanned in the NIR instrument. A volume 1mL of the collected full sample is used for preparation of mixtures which are also scanned in the NIR.

Table 5.1: Composition of five mixtures prepared for both full sample from one vessel and supernatant. A volume of 1mL was prepared for each mixture. The volumes (in mL) of glucose, water and sample from the running cultivation are listed. A stock solution of 200g/L of glucose (previously filter-sterilised) was used for these treatments

Mixture name	**Raw**	**Gluc1**	**Gluc2**	**Gluc3**	**Diluted**
Volume of glucose stock 200g/L (ml)	0.00	0.75	0.90	0.95	0.00
Concentration of glucose originated from stock 200g/L (g/L)	0.00	150	180	190	0.00
Volume of Sample (ml)	1.00	0.250	0.100	0.05	0.05
Water (ml)	0.00	0.00	0.00	0.00	0.95

Ground truth Glucose

Glucose concentration of the samples collected from the bioreactors were obtained by strips that were designed to measure blood glucose levels (Accu-Chek Aviva, Roche Diagnostics), due to its practicality and low price.

The "ground truth" concentration of glucose in each volume of prepared mixture was then calculated by adding the content of glucose stock to the appropriate content sample concentration (which was determined through blood-glucose strips):

$$Conc.Gluc_{GroundTruth} = \frac{V_{GlucoseStock} \times Conc.Gluc_{stock} + V_{Sample} \times Conc.Gluc_{glucosestrips}}{V_{TotalMixture}} \tag{5.1}$$

Mixtures "gluc1", "gluc2" and "gluc3" included 150g/L, 180g/L and 190g/L of glucose, respectively. These high spikes of glucose increase the concentration to "acceptable" levels of glucose; levels that have been previously detected by the developed prototype (spectrometer and flow-cell set-up).

The glucose concentration in the original sample can then be calculated as described in the table below (Table 5.2).

NIR pre-processing and Software

Spectra of both the whole matrix and filtrate samples were acquired with the spectrometer previously described in Chapter 4. A total of five scans were collected per

Table 5.2: Calculation of glucose concentration in the original sample collected from the bioreactor (either in the form of full sample or supernatant), based on the prediction for each prepared mixture.

Mixture name	**Raw**	**Gluc1**	**Gluc2**	**Gluc3**	**Diluted**
Glucose prediction in original sample(g/L)	$Gluc_{pred}$	$\frac{Gluc_{pred}-150}{0.25}$	$\frac{Gluc_{pred}-180}{0.1}$	$\frac{Gluc_{pred}-190}{0.05}$	$\frac{Gluc_{pred}}{0.05}$

sample, over the entire range available (901 – 2186 nm). All NIR data was analysed with SIMCA 13 and SIMCA 14, as well as R software.

Figures of merit

To evaluate the performance of the models, the predicted concentrations of test set samples were assessed using different statistics. The quality of concentration quantitative predictions was evaluated by the relative percentage error in predicted concentration (%RE), the root mean square error of prediction (RMSEP) and the bias as defined by Eqs.(5.2-5.4), respectively.

$$RE\% = 100 \times \sqrt{\frac{\Sigma_{i=1}^{n}(Y_{obs}-Y_{pred})^2}{\Sigma_{i=1}^{n}Y_{pred}^2}} \tag{5.2}$$

$$RMSEP = \sqrt{\frac{\Sigma_{i=1}^{n}(Y_{obs}-Y_{pred})^2}{n}} \tag{5.3}$$

$$Bias = \frac{\Sigma_{i=1}^{n}(Y_{obs}-Y_{pred})}{n} \tag{5.4}$$

5.3 Results and Discussion

With the goal of determining the best approach for model development for NIR in a setting such as the Ambr® systems, three different approaches were tested.

5.3.1 Summary of Datasets

To obtain real samples, several fermentations were performed in the Ambr® bioreactors, throughout 4 different days. The specific conditions used in each vessel are described below. Only one culture station (CS) was used per day, even though the system has two available.

5.3.1.1 Run A - 28 April 2017

A scheme of the conditions used in each bioreactor of the CS during this run is shown in Figure 5.4. Vessels 1 to 6 (top row of CS) were inoculated with 3g/L of yeast and vessels 7 to 12 were inoculated with 6g/L of yeast. The first set of six vessels on the left (v1, v2, v3 and v7, v8 and v9) had 15g/L of initial glucose in the medium and the six vessels on the right (v4, v5, v6, v10, v11 and v12) had 30g/L of initial glucose in the medium.

A sample was collected from a total of four vessels (v1, v3 and v7 and v9), three hours posterior to inoculation. For each vessel, the full sample as well as the supernatant were scanned in the NIR system. From the mixtures described in Table 5.1, only the "gluc3" was prepared, at this point.

The readings of Dissolved Oxygen (%DO) over time for each vessel, recorded by the Ambr, are shown in Figure 5.5. The agitation speed (in rpm) is also plotted in the same figure.

All vessels were inoculated between 12h15 and 12h30. Three hours after, four vessels were sampled: v1, v3, v7 and v9. After eight hours, a final reading of glucose was collected offline.

From the readings of %DO, it is seen that from the moment of inoculation, vessels v1 to v5 show lower %DO as they have more cells than v6 to v12. Vessel 2 showed

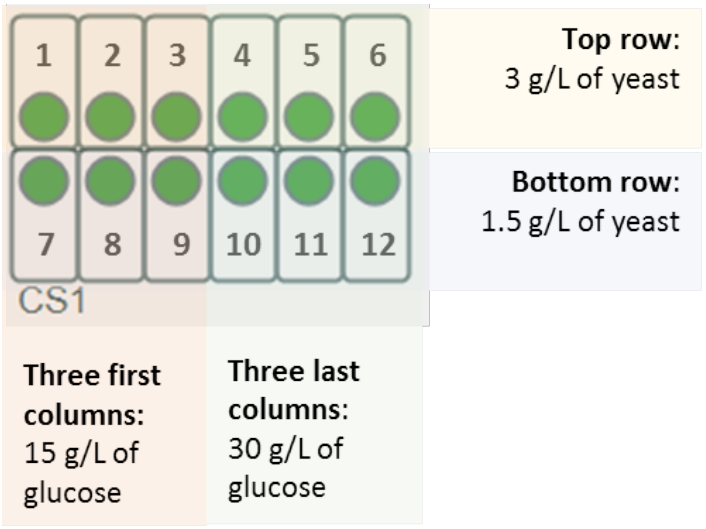

Figure 5.4: Conditions of all vessels of CS1: vessels of the top row (vessel 1 to vessel 6) were inoculated with 3g/L of yeast and the vessels at the bottom row (vessel 7 to vessel 12) were inoculated with 1.5 g/L. For each of these conditions, two different concentration levels of glucose were tested: 15g/L (vessels 1, 2, 3, 7, 8, 9) and 30g/L (vessels 4, 5, 6, 10, 11, 12).

a different behaviour from its replicates and therefore was disregarded. Depletion of glucose in vessels that were inoculated with higher concentration of yeast (v3, v4, v5 and v6) occurred at approximately 16h30, which led to an increase on the %DO. However, the %DO was still 80% which might also indicate some growth. Crabtree positive yeasts, such as *S. cerevisiae*, can go into alcoholic fermentation at high rates of sugar uptake (when high concentrations of sugar are used), even at high contents of oxygen (Weusthuis et al., 1994). Alcoholic fermentation is not desired because it inhibits growth and reduces the biomass yield and thus the production of a potential protein. To avoid this effect, fed-batch mode at a low rate of sugar can be used.

Based on the offline readings of glucose (plotted in Figure 5.6), its consumption rate was 2.8 g.L-1.h-1 in vessels with 3g/L yeast and 15g/L initial glucose (marked with circles in the plot); and 1.3 g.L-1.h-1 in vessels that were inoculated with 1.5g/L yeast (marked with triangles) which contained 15g/L of initial glucose. No sample was collected from the vessels containing 30g/L of initial glucose, at t=3h. However, using the last glucose reading, from one vessel with 3g/L of initial yeast, the approximate consumption was 2.3 g.L-1.h-1 (plotted with squares).

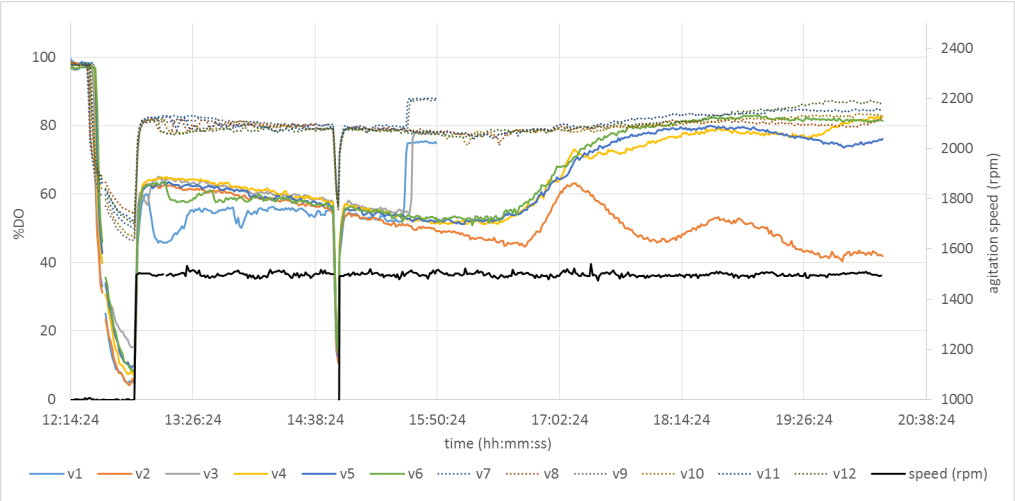

Figure 5.5: Run A (from 28th April): %DO (dissolved oxygen) for each vessel (v1 to v12) and agitation speed (black line, secondary axis) over time. All vessels were inoculated between 12h15 and 12h30, and there was a single sample point collected 3 hours after; vessels 1, 3, 7 and 9 were sampled. The lines in the plot are from the top row vessels (which were inoculated with 3 g/L yeast) and the pointed lines are from vessels on the bottom row (inoculated with 1.5 g/L yeast).

From this run, a total of 16 mixtures were generated for the NIR analysis: four vessels were sampled, the samples separated into full sample and supernatant. For each of these, both "raw" and mixture "gluc3" were prepared to be scanned.

5.3.1.2 Run B - 02 May 2017

The conditions of each bioreactor in the culture station are described in Figure 5.7, all vessels were inoculated with 3 g/L of yeast solution, the top row vessels contained medium with 15g/L of initial glucose and the medium in the vessels at the bottom row of the CS contained 30g/L of glucose initially.

The readings over time for each vessel of this run, recorded by the Ambr, are shown in Figure 5.8 and the glucose offline readings for the collected samples are plotted in Figure 5.9.

All vessels were inoculated at 11h30. Vessels v1 and v7 were sampled right after. At 15h00, v2 and v8 were sampled. At 18h00, vessels v3 and v9 were sampled.

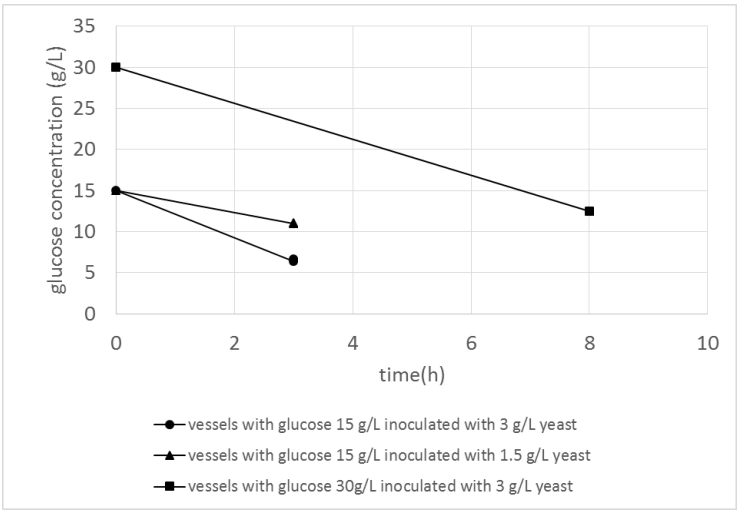

Figure 5.6: Glucose offline readings for samples collected from the vessels containing 15g/L of initial glucose in the medium and 3g/L of yeast (circles), samples from vessels containing 15g/L of initial glucose and 1.5g/L of yeast (triangles), samples from vessels containing 30g/L of initial glucose and 3g/L of yeast (squares).

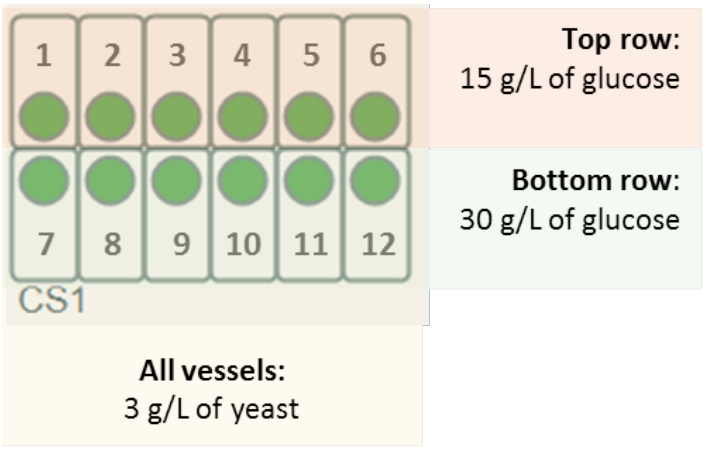

Figure 5.7: Run conditions from vessels at CS1: all vessels were inoculated with 3g/L of yeast and the medium in the vessels at the top row contained 15g/L of glucose and the medium in the vessels at the bottom row of the CS contained 30g/L of glucose.

Based on the offline readings of glucose (plotted in Figure 5.8), the consumption of glucose was 3.6 g.L-1.h-1 (on the vessels with 15g/L of initial glucose) and 4.0 g.L-1.h-1 (on the vessels with 30 g/L of initial glucose).

In comparison to the vessels with the same yeast content from run A, for which the glucose consumption rate was 2.8 g.L-1.h-1 (15g/L initial glucose) and 2.3 g.L-1.h-1 (30g/L initial glucose), the values are higher for run B. Given the procedure used for biomass, it is possible that vessels of run B were inoculated with slightly higher content of cells than vessels from run A.

From run B, a total of 60 mixtures were generated and scanned in the NIR setup at the same time: six vessels were sampled, each collected sample separated into full sample and supernatant. Each of these was then used to prepare the 5 mixtures described in Table 5.1 to be scanned.

5.3.1.3 Run C – date 03 May 2017

The conditions used in each bioreactor are described in Figure5.10 all vessels had fresh medium with 30g/L of initial glucose; the vessels of the top row were inoculated with 6 g/L of yeast and the vessels from the bottom row 3g/L of yeast.

The readings over time for each vessel, collected through the Ambr®, are shown in Figure 5.11 and the glucose offline readings for the collected samples are plotted in Figure 5.12.

Because the vessels inoculated with 6g/L (v1 to v6, marked with lines in the plot) had low %DO, the oxygen could become a limiting factor of the cells and therefore the agitation was increased from 1500 to 1900 rpm.

Run C generated a total of 60 mixtures: six vessels were sampled, each collected sample separated into full sample and supernatant. Each of these was then used to prepare the 5 mixtures described in Table 5.1

5.3.1.4 Run D – date 04 May 2017

For this run the conditions of Run C were used again, as shown in Figure 5.10. The dissolved oxygen and speed, for the 12 vessels, throughout the new run, are plotted in

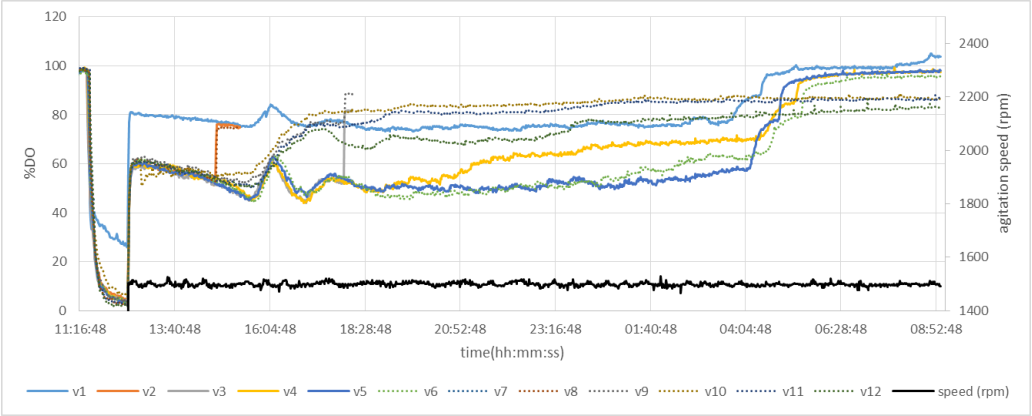

Figure 5.8: Readings over time of Dissolved Oxygen (%DO) (main Y axis) and agitation speed (rpm) (secondary Y axis), for Run B. Full lines are from the vessels in top row vessels of the CS - containing 15g/L of initial glucose in the medium - and pointed lines are from vessels in the bottom row, which contained 30g/L of initial glucose. Vessels were inoculated at 11h30 and samples were collected straight after from vessels v1 and v7; at 15h00 from vessels v2 and v8; and at 18h00 from vessel v3 and vessel v9. Agitation was constant at 1500rpm (in the secondary axis).

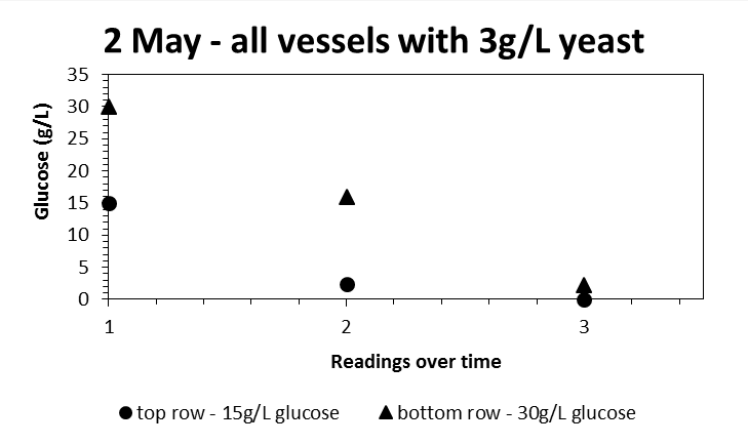

Figure 5.9: Glucose offline readings for samples collected from the vessels containing 15g/L of initial glucose in the medium (circles) and for samples from the vessels containing 30g/L of initial glucose (triangles). All vessels were inoculated with the same amount of yeast (3g/L). Using the first two points, the glucose consumption rate at the top row vessels was 6.9g glucose/L.h and at the bottom row was 6.7g glucose/L.h.

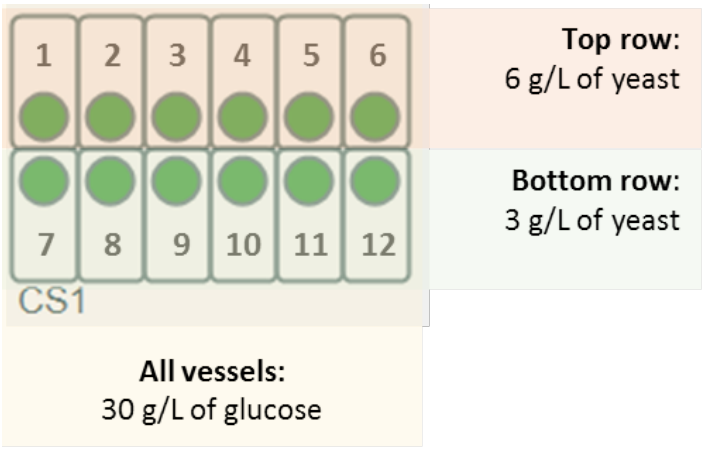

Figure 5.10: Run conditions from vessels at CS1: all vessels has the same medium with initial glucose concentration of 30g/L; the top row vessels were inoculated with 6g/L of yeast and the vessels at the bottom row of the CS were inoculated with 3g/L of yeast.

Figure 5.13. Bioreactors 2 (orange line) and 5 (dark blue line) were not selected to be sampled because they showed a different behaviour from the expected, which is seen for bioreactors 1 and 4. Vessel 6 (dotted green) might have been inoculated with 6g/L of yeast, instead of the planned 3 g/L, since the values of DO match these conditions.

As seen before, the vessels with high content of yeast might have moved into alcoholic fermentation, or consumption of another by-product. The levels of %DO indicate that glucose finished at approximately 14h30, for bioreactors v1, v2, v5, v6.

Using the first reading of v4 and the first reading of v1 the consumption of glucose was 7.3 g.L-1.h-1 for vessels with 6g/L of yeast. Using the first reading of v7 and the first reading of v9 the consumption of glucose for bioreactors with 3g/L of yeast was 4.1 g.L-1.h-1. These values are in agreement with the ones from the previous run.

Run D generated a total of 80 mixtures: eight vessels were sampled, each one split into full sample and supernatant which were then used to prepare the 5 mixtures described in Table 5.1.

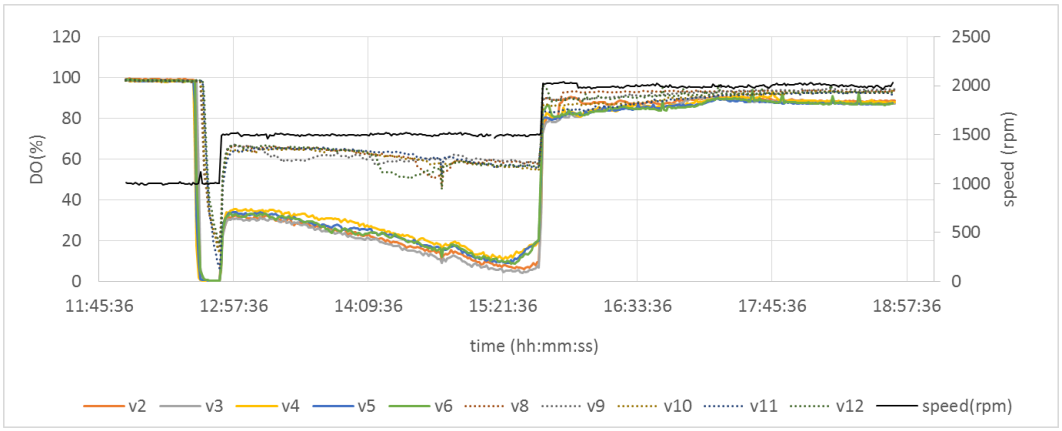

Figure 5.11: Dissolved Oxygen (%DO) and agitation speed (rpm) in the second axis, from Run C. Coloured lines are from the vessels in top row (inoculated with 6g/L yeast) and pointed lines are from vessels in the bottom row (inoculated with 3g/L yeast). Inoculation was at 12h40, samples were collected straight after from v1 and v7 (not shown in the plot), then from v2 and v8 at 15h30 and after, from v3 and v9 at 18h55.

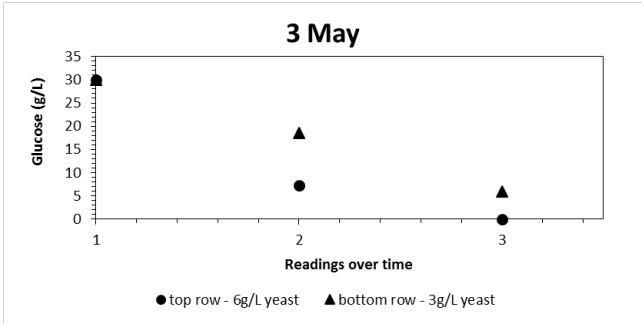

Figure 5.12: Reference glucose readings of samples collected from the top row vessels (inoculated with 6g/L of yeast) marked with circles and from the bottom row vessels (inoculated with 3g/L of yeast) marked with triangles. The same medium was used in all vessels and with an initial concentration of glucose of 30g/L. Using the first two points, glucose consumption rate was 7.6 (g glucose)/L.h for the circles and 4.0 (g glucose)/L.h (using the first two readings) for the triangles.

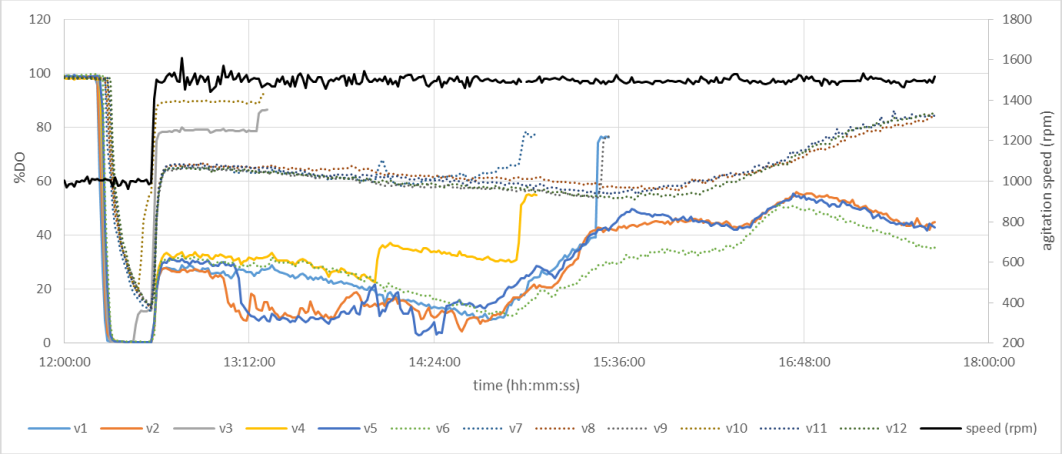

Figure 5.13: Dissolved Oxygen (%DO) in the main y-axis and agitation speed (rpm) in the secondary y-axis, from Run D. Full lines correspond to the vessels in top row of the culture station (inoculated with 6g/L yeast) and pointed lines are from vessels in the bottom row (inoculated with 3g/L yeast). Inoculation was at 12h20, samples were collected straight after from vessels v3 and v10, then at 13h15 from the same vessels, then at 14h00 (and again at 15h00) from v4 and v7 and finally, at 15h30, from vessels v1 and v9.

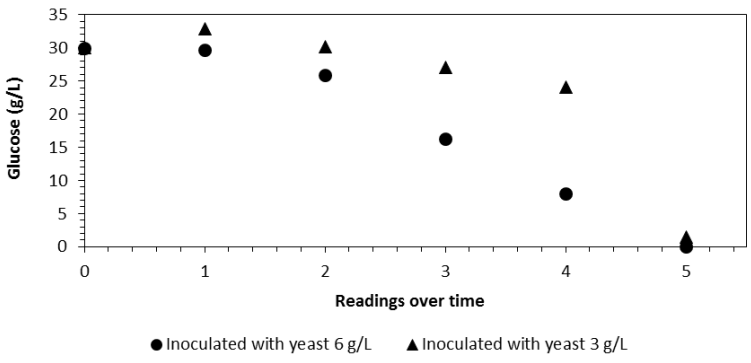

Figure 5.14: Reference glucose readings of samples collected from the top row vessels (inoculated with 6g/L of yeast) marked with triangles and from the bottom row vessels (inoculated with 3g/L of yeast) marked with circles. The glucose consumption rate was 4.1 g.L-1.h-1 and 7.3 g.L-1.h-1, respectively. The same medium was used in all vessels and with an initial concentration of glucose of 30g/L.

5.3.2 Testing the Pre-calibrated Model

Two models (developed in the previous chapter), based on samples prepared through the use of liquid-handler, were now challenged to predict the concentrations of glucose and yeast in the real cultivation samples from the Ambr®systems.

5.3.2.1 Glucose Quantification

The model developed for glucose described in the previous chapter is summarised in Table 5.3. The model developed for glucose was tested on the prediction of supernatant and full samples and their mixtures. The results of predictions are summarised in Table 5.4 and Table 5.5. The samples with a prediction higher than 300g/L were considered outliers and not accounted for in these tables.

Table 5.3: Model statistics for the PLS model for glucose concentration (in g/L), based on 30 samples (149 scans) from Run A.

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.774	0.777	27.8	0.771	3.37	3.47
2	0.220	0.211	7.92	0.221		
Cumulative	0.994	0.989	--	--		

Table 5.4: Evaluation of the pre-calibrated model for glucose prediction of supernatant samples (and prepared mixtures) from the different cultivation runs.

	Run A	Run B	Run C
Glucose conc. range (g/L)	6.36 - 190.5	0.117 - 191.5	0.0 - 191.5
Prediction glucose conc. range (g/L)	11.5 - 200.2	-25.6 - 216.5	-41.9 - 220.4
RMSEP (g/L)	8.17	18.0	24.6
%RE	6.99	12.8	18.1
Bias (g/L)	-5.86	-11.2	-9.12
N samples (and number of scans)	8 (40)	4 (20)	5 (25)

From the results in these tables it is evident that the pre-calibrated model is able

Table 5.5: Evaluation of the pre-calibrated model for glucose prediction of the whole samples (and prepared mixtures) from the different cultivation runs.

	Run A	Run B	Run C
Glucose conc. range (g/L)	6.36 - 190.6	0.117 -190.8	0.365 - 191.5
Prediction glucose conc. range (g/L)	-209.0 - 192.2	-189.0 – 209.5	-590.0 – 285.0
RMSEP (g/L)	94.5	112.0	220.8
%RE	74.3	71.8	170.7
Bias (g/L)	39.3	51.2	68.6
N samples (and number of scans)	9 (45)	13 (65)	17 (85)

to give some information about the supernatants but the predictions of whole samples yielded relative errors (RE) higher than 70%.

5.3.2.2 Yeast Quantification

The pre-calibrated model developed for yeast described in the previous chapter is summarised in Table 5.6.

Table 5.6: Model statistics for the PLS model for yeast concentration (in g/L), based on 23 samples (115 scans) from Run A.

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.959	0.962	104	0.962	0.0375	0.0381
2	0.0355	0.0164	3.83	0.408		
Cumulative	0.994	0.979	--	0.978		

In Figure 5.15 the prediction of yeast concentration against ground truth are plotted. The samples were randomly scanned in the last day of cultivations.

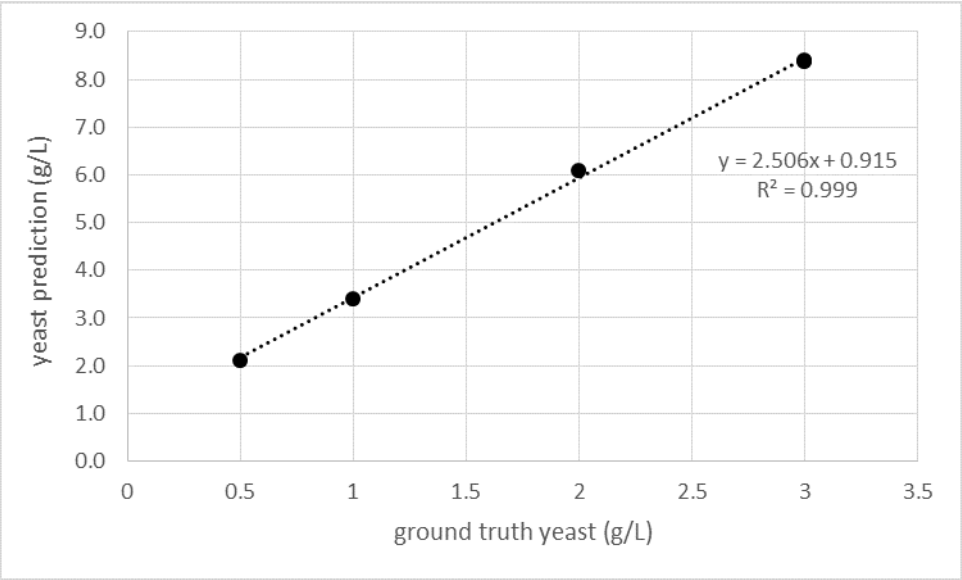

Figure 5.15: Yeast prediction (based on model described in Table 5.6) against ground truth, for four mixtures of yeast in water.

5.3.3 Model Development on Updated Standards

With the goal of achieving a solution for samples collected from Ambr®cultivations, new models were developed based on samples that were scanned throughout the days of cultivations. This approach allows to have the exact same stocks of glucose and yeast represented in the calibration dataset. A dataset composed by 3 standards of yeast (0.5, 1.0 and 2.0g/L), water samples (23 samples), fresh medium and pure glucose solutions (24 samples) – 250 scans in total - was used for model development. The scores plot of a PCA model for these data points is shown in Figure 5.16.

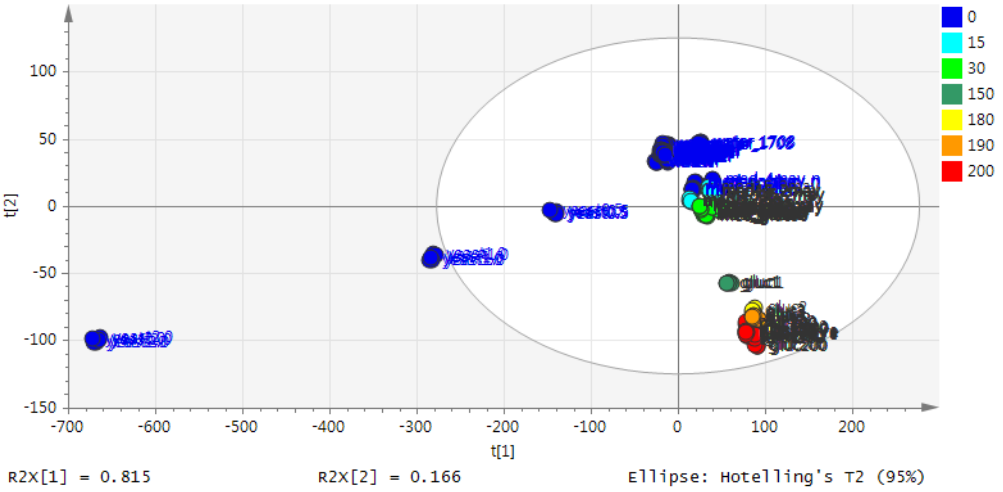

Figure 5.16: Scores plot from PCA developed based on standards scanned during the cultivation period. Water samples (centred in PC1=0), yeast samples (0.5, 1.0 and 2.0 g/L, on the left-hand side of the 1PC), fresh medium without glucose, fresh medium with 15g/L of glucose, fresh medium with 30g/L of glucose, pure glucose samples (150, 180, 190 and 200 g/L).

5.3.3.1 PLS Model for Yeast

The statistics of the model for the PLS developed for yeast concentration are listed in Table 5.7. Posterior to PLS model development, the model was tested on the prediction of samples for which the yeast content was unknown. Thus, the predictions were plotted per vessel or per time point and are shown below for the different runs.

Table 5.7: Model statistics for the PLS model that predicts yeast concentration in g/L.

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.812	0.928	87.7	0.928	0.0373	0.0372
2	0.169	0.0576	18.2	0.805		
Cumulative	0.981	0.986	-	0.986		

Predictions of Yeast

To compare the predictions of yeast with the results from the pre-calibrated model (which are summarised in Table 5.5), a similar table is shown below (Table 5.8). This table was built based on samples with known concentration of yeast. The analysis of

Table 5.8: Evaluation of the model for yeast prediction of full samples (and prepared mixtures) from the different cultivation runs, developed on standards scanned throughout the same days. Only samples with known concentration of yeast were used.

	Run B	Run C	Run D
Ground truth of yeast conc. range (g/L)	0.15 – 3.0	0.15 – 6.0	3.0 – 6.0
Prediction yeast conc. range (g/L)	0.0955 – 3.92	0.125 – 6.71	3.72 – 5.27
Range(predicted) / range(ground truth)	1.34	1.13	1.18
RMSEP (g/L)	0.546	0.435	2.23
%RE	27.4	17.8	43.0
Bias (g/L)	-0.39	-0.25	0.74
N samples (number of scans)	9 (44)	8 (40)	3 (15)

samples with unknown concentration of yeast had to be qualitative and is discussed in the plots below.

The errors of prediction for yeast for unknown mixtures from runs B and C were below 0.55g/L (Table 5.8). The three samples from run D showed higher error of prediction but it can be due to different stock of yeast used. The prediction of each prepared mixture was used to estimate the content of the original sample and each appropriate calculation is described in Table 5.9.

Table 5.9: Calculation of yeast content in the original sample based on the prediction of the prepared mixtures.

Mixture name	***Raw***	***Gluc1***	***Gluc2***	***Gluc3***	***Diluted***
Prediction in mixture (g/L)	a	b	c	d	e
Estimation of content in original sample	a	$\frac{b}{0.25}$	$\frac{c}{0.1}$	$\frac{d}{0.05}$	$\frac{e}{0.05}$

Figure 5.17 shows the prediction of yeast for the raw samples collected during run A, as well as the estimated content calculated from the prediction of the mixture "gluc3".

No offline reading was obtained for yeast content on these samples. However, given that vessels 1 and 3 were initially inoculated with 3g/L of yeast and vessels 7 and 9 with 0.5g/L, they should have a higher amount of yeast at the time of sampling (3 hours after). Therefore, the predictions of raw sample for v1 and v3 seem to be more accurate than the calculation from the mixture of "gluc3".

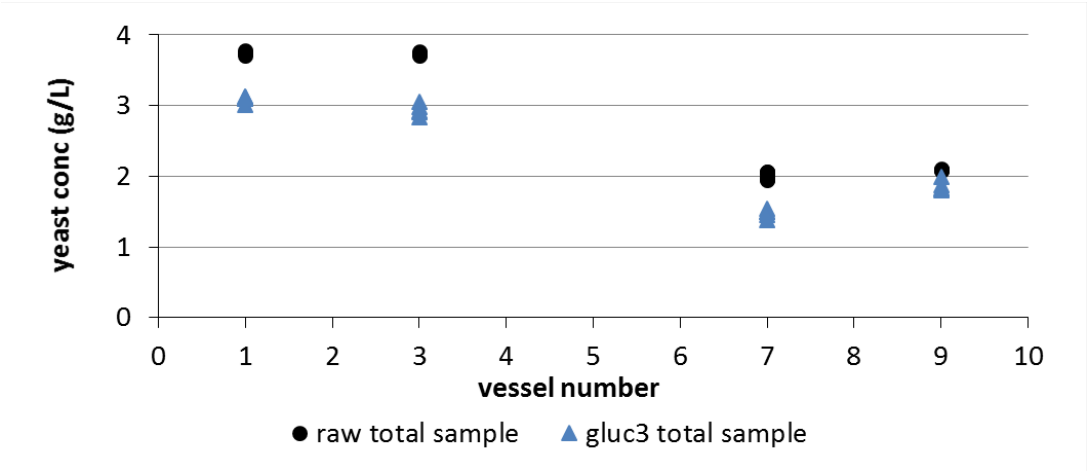

Figure 5.17: Prediction of yeast content in full samples: raw (circles) and in mixture "gluc3" (triangles), for vessel 1, 3, 7 and 9, from run A (28th April). Vessels 1 and 3 are duplicates, as well as vessel 7 and vessel 9.

For run B, samples were collected in two different time points: straight after inoculation (timepoint 1) and 3 hours later (timepoint 2). The predictions are plotted in Figure 5.18.

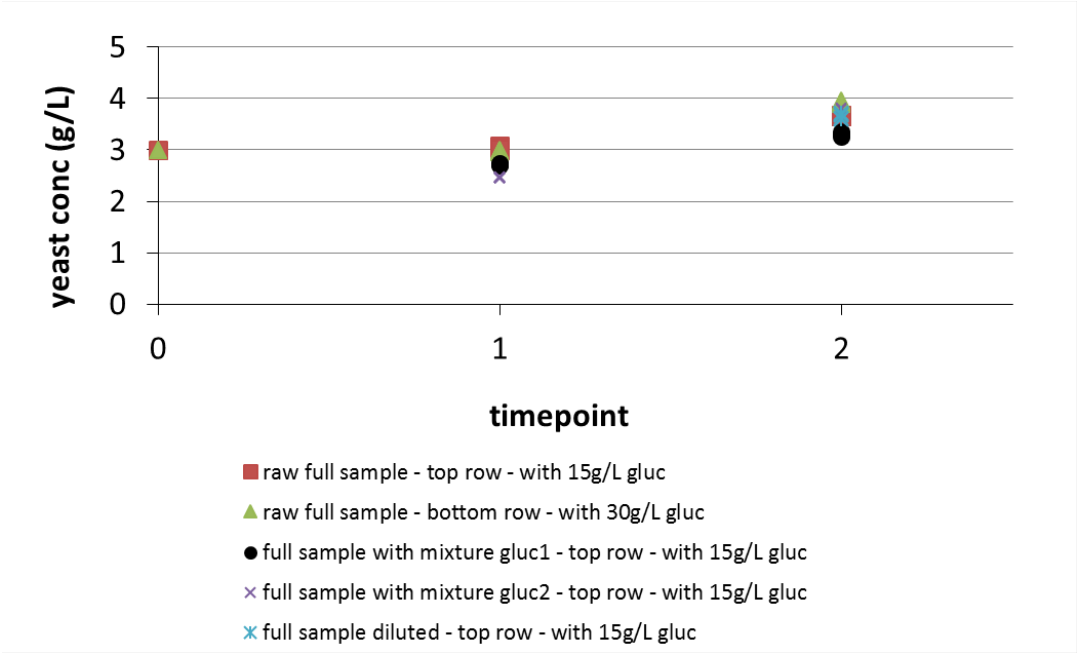

Figure 5.18: Prediction of yeast content in full samples from run B (2nd May). Point zero in the plot does not result from model prediction.

For run C, the estimations of the yeast content in the original sample for each scanned mixture (raw sample, in mixture "gluc1", in mixture "gluc2", in "gluc3" or diluted) are plotted in Figure 5.19 and Figure 5.20. The first plot shows the results from the top row of the CS, which was inoculated with 6g/L of yeast and the second plot shows the results from the bottom row of the CS, inoculated with 3g/L of yeast. Two points of sample collection: after inoculation (v1 and v3) and 3 hours after (v2 and v8).

The average prediction for the samples from vessel 1 (using all the mixtures) is 6.3 ± 1.3 and from vessel 2 is 6.9 ± 1.5. Thus, a growth of 0.6 g/L of yeast was detected between the two time points.

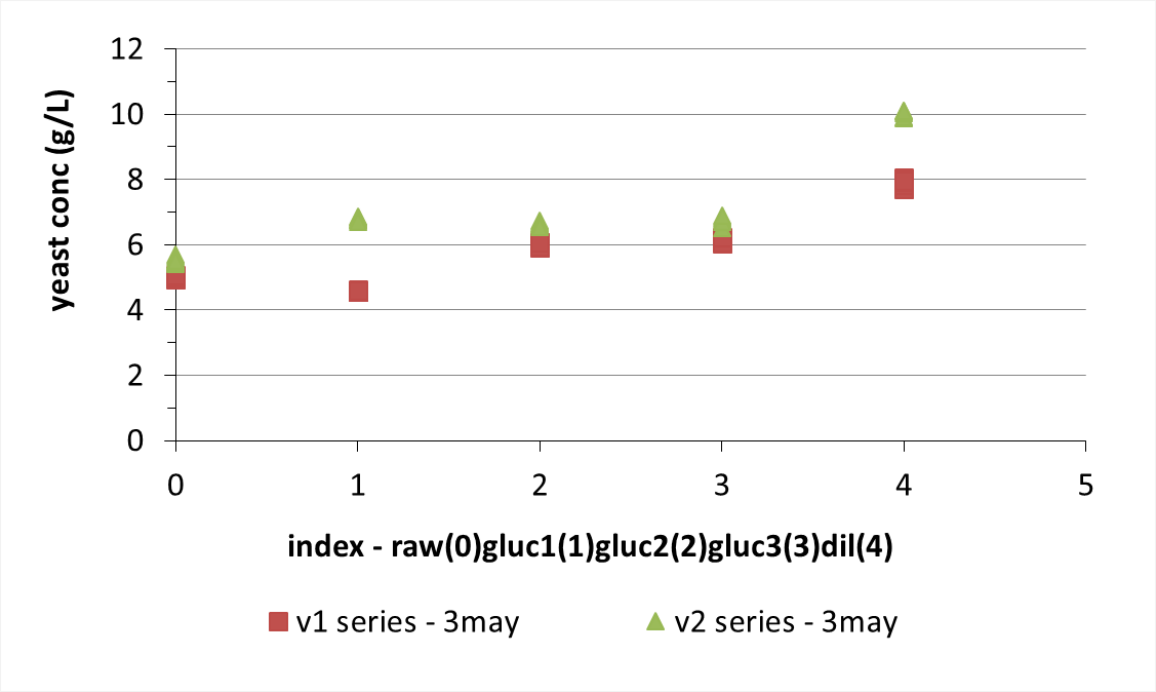

Figure 5.19: Prediction of yeast concentration for the two samples collected from the vessels inoculated with 6g/L of yeast at run C: vessel 1 (v1, red squares series) was sampled straight after inoculation and vessel 2 (v2, green triangles series) was sampled 3 hours after inoculation. The x axis shows an index for the type of mixture scanned: raw samples (0), in mixture "gluc1" (1), in mixture "gluc2" (2), in mixture "gluc3" (3), and diluted (4). The prediction was corrected for the amount of original sample in each mixture as described in Table 5.9.

The prediction of yeast concentration from the samples collected from the bottom

row (inoculated with 6 g/L) estimated from the different mixtures is plotted in Figure 5.20. The sample from vessel number 7 (purple diamond series) was obtained straight after inoculation and the one from vessel 8 (black circle series) was obtained 3 hours after. The mixture gluc3 for the sample from v7 is lower than expected and for the sample from v3 is higher than expected. Thus, only considering the raw sample from v7 and from v8, the growth of yeast was 0.33g/L.

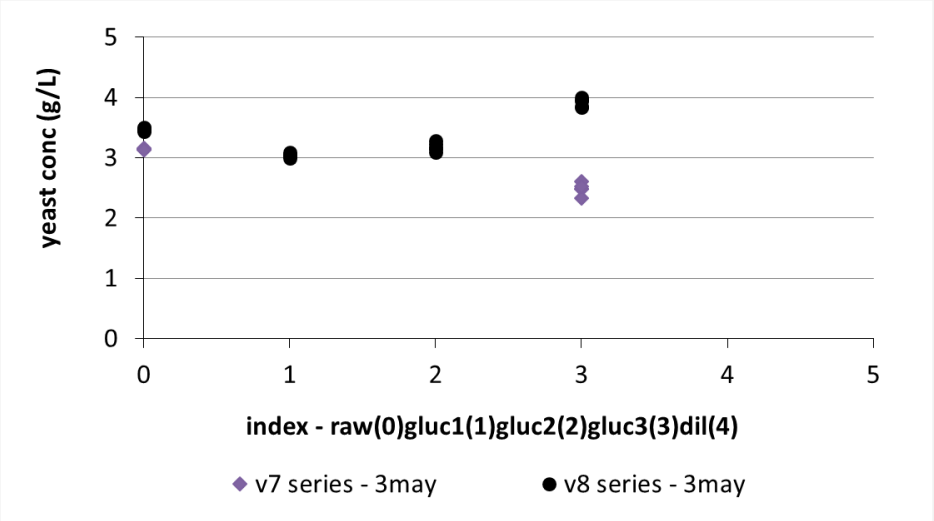

Figure 5.20: Estimation of yeast concentration for the two samples collected from the vessels inoculated with 3g/L of yeast at run C: vessel 7 (v7, purple diamond series) was sampled straight after inoculation and vessel 8 (v8, black circle series) was sampled 3 hours after inoculation. The x axis shows an index for the type of mixture scanned: raw sample (0), in mixture "gluc1" (1), in mixture "gluc2" (2), in mixture "gluc3" (3), and diluted (4). The prediction was corrected for the amount of original sample in each mixture as described in Table 5.9.

5.3.3.2 PLS Model for Glucose

The PLS model developed for glucose, based on the described dataset, is summarised in Table 5.10. The details of the calibration and validation datasets are described in Table 5.11. Predictions of glucose for run D are plotted in Figure 5.21. As expected, in vessels inoculated with higher cell densities glucose was consumed faster, which means that valid trends were obtained. However, the plotted values were adjusted to the initial concentration of 30g/L. Because high error of prediction were originally obtained, a correction step is proposed below. For run D (4 May), the predicted trends are plotted in Figure 5.22.

Table 5.10: Model statistics for the PLS model that predicts glucose concentration.

Latent variables	$R^2(X)$	$R^2(Y)$	Eigenvalue	Q^2	RMSEE (g/L)	RMSEcv (g/L)
1	0.704	0.529	76	0.521	18.2	18.2
2	0.277	0.412	29.9	0.94		
Cumulative	0.981	0.941	--	--		

Table 5.11: Performance of the developed model for glucose prediction of the calibration dataset and the validation dataset.

	Calibration dataset	Validation dataset
Glucose conc. range (g/L)	0 200	0 200
Prediction conc. range (g/L)	-22.6 - 203.9	-66.3 - 235.5
RMSEE or RMSEP (g/L)	18.1	24.7
%RE	21.0	17.0
Bias (g/L)	1.13×10^{-5}	-0.17

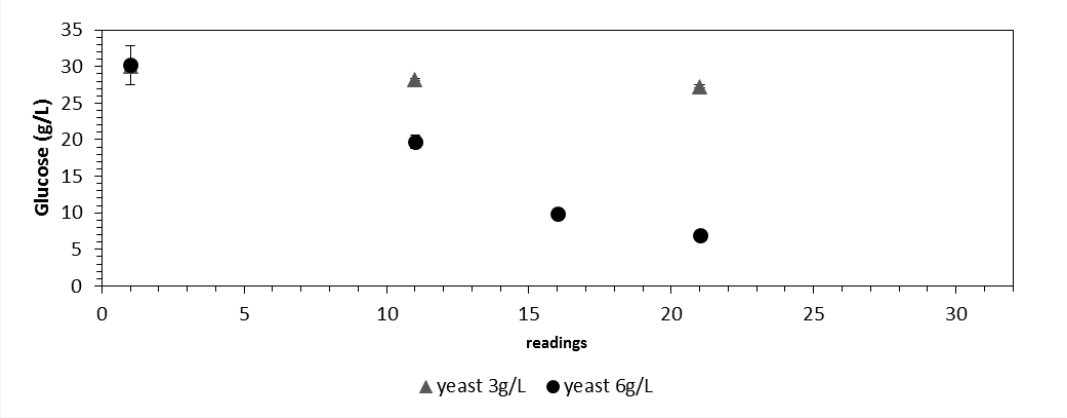

Figure 5.21: Glucose prediction (adjusted) for run D, on vessels with 3g/L of yeast (triangles) and 6g/L of yeast (circles)

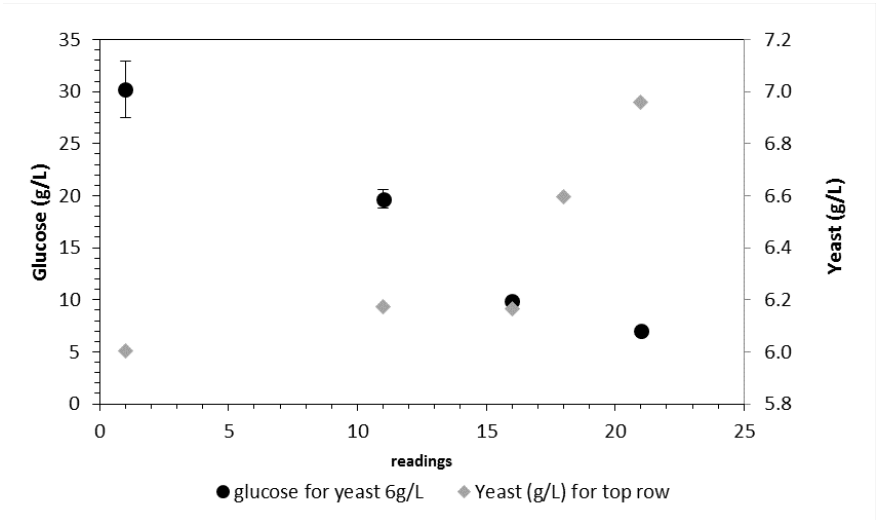

Figure 5.22: Prediction of concentrations of glucose and yeast, over the course of run D.

5.3.3.3 Correction for Glucose Prediction

A methodology is proposed to correct the concentration of glucose predicted by the PLS model, as described in Figure 5.23. Firstly, the NIR spectra of samples with known content of yeast and glucose (i.e. mixtures of full samples) is used to calibrate a PLS model for yeast and to calibrate a PLS model for glucose. With these predictions, a new corrected prediction of glucose concentration is obtained.

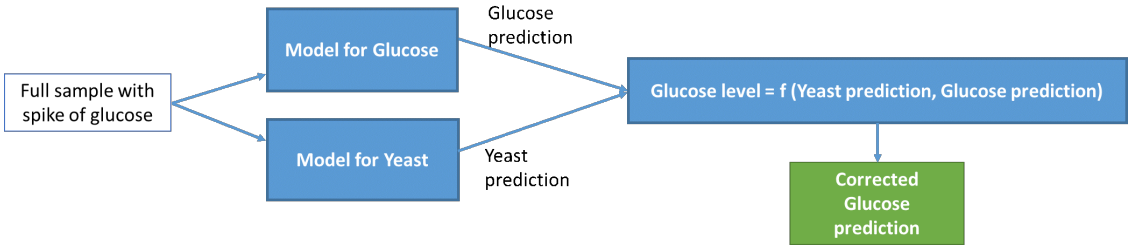

Figure 5.23: Methodology for correction of glucose prediction.

The correlation between the error of prediction of glucose (x axis) and the prediction of yeast for each sample is shown in Figure 5.24. A linear regression was built based on the samples in black and tested on the red samples. The statistics of this model are summarised in Table 5.12.

Table 5.12: Summary of regression built to obtain a new corrected prediction of glucose based on the prediction of glucose and the prediction of yeast. The interaction term is not significant and a general R-squared of 0.90 was obtained.

	Estimation	Standard error	t-value	p-value
Prediction of glucose	0.368	0.0725	5.08	8.82e-06 ***
Prediction of yeast	-51.7	15.3	-3.39	0.00157 **
Interaction term	0.0211	0.0742	0.284	0.778
Intercept	131.5	13.4	9.80	2.66e-12 ***

The regression was built with 9 samples (a total of 45 points) and was tested with 13 samples (total of 65 points). The prediction of the calibration and the validation datasets are shown in Figure 5.25 and Figure 5.26, respectively.

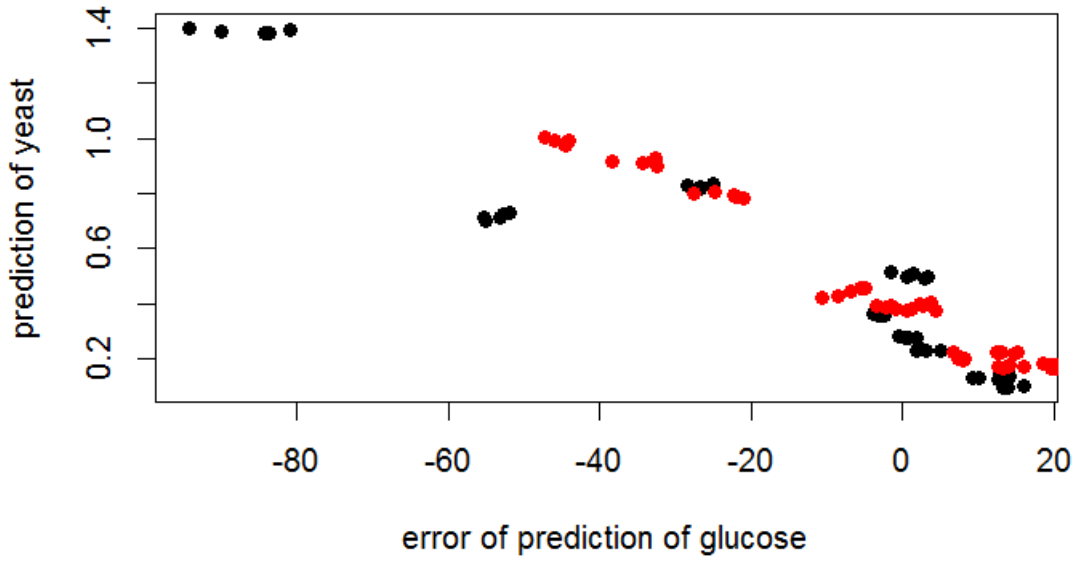

Figure 5.24: Prediction of yeast against error of prediction of glucose. Samples used for calibration (black dots) and used for validation (red dots). Selection of samples with more than 100g/L of glucose and less than 5g/L of yeast.

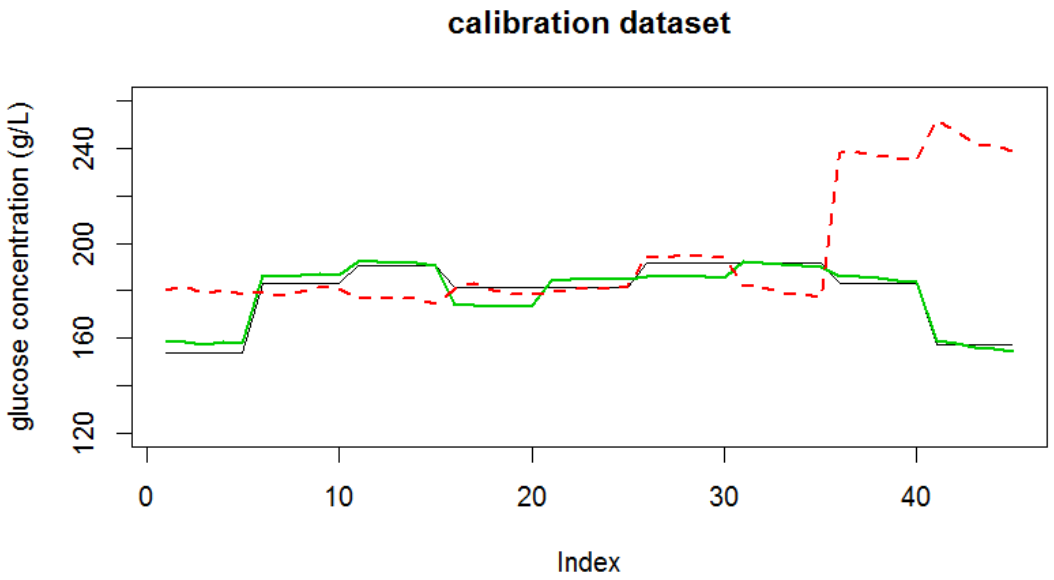

Figure 5.25: Glucose concentration for the samples used for the regression: ground truth (black line), original prediction from the PLS model (red dashed line) and new corrected prediction of glucose (green line).

Figure 5.26: A total of 13 samples (5 scans each) was used for validation. Ground truth (black line), original prediction from the PLS model (red dashed line) and new corrected prediction of glucose (green line).

Therefore, the following correction can be used on the glucose prediction:

$$NewGlucoseValue = 0.368 \times PredGlucose - 51.7 \times PredYeast \tag{5.5}$$

Table 5.13: Performance of the regression model for calibration and validation datasets, before and after correction.

	Calibration dataset		**Validation dataset**	
	Before correction	After correction	Before correction	After correction
Glucose conc. range (g/L)	153.8 – 191.5		151.8 – 190.9	
Original prediction glucose conc. range (g/L)	174.7 – 251.7		164.5 – 257.9	
Prediction yeast conc. range (g/L)	0.0955 – 1.40		0.0751 – 2.09	
RMSEP (g/L)	35.7	**4.04**	35.8	**7.94**
%RE	19.8	2.25	19.9	4.40
Bias (g/L)	-15.4	-3.9×10^{-14}	-7.48	4.46

5.3.4 Model Development on Cultivation Samples

A PLS model was calibrated on a total of 59 scans from samples of supernatants and spiked levels from run B (12 individual mixtures). Figure 5.27 shows the scores plot for this model and Table 5.14 summarises its statistics. A RMSEE of 4.35g/L was obtained, and the loadings plot for the selected wavelengths is shown in Figure 5.28.

Table 5.14: Model statistics for the PLS developed on mixtures from supernatant samples collected during run B.

Latent variables	**R^2(X)**	**R^2(Y)**	**Eigenvalue**	**Q^2**	**RMSEE (g/L)**	**RMSEcv (g/L)**	**R^2 (Normaliz. Residuals)**
1	0.823	0.984	26.3	0.985	4.35	4.52	0.995
2	0.109	0.0114	3.50	0.785			
Cumulative	0.933	0.997	-	0.997			

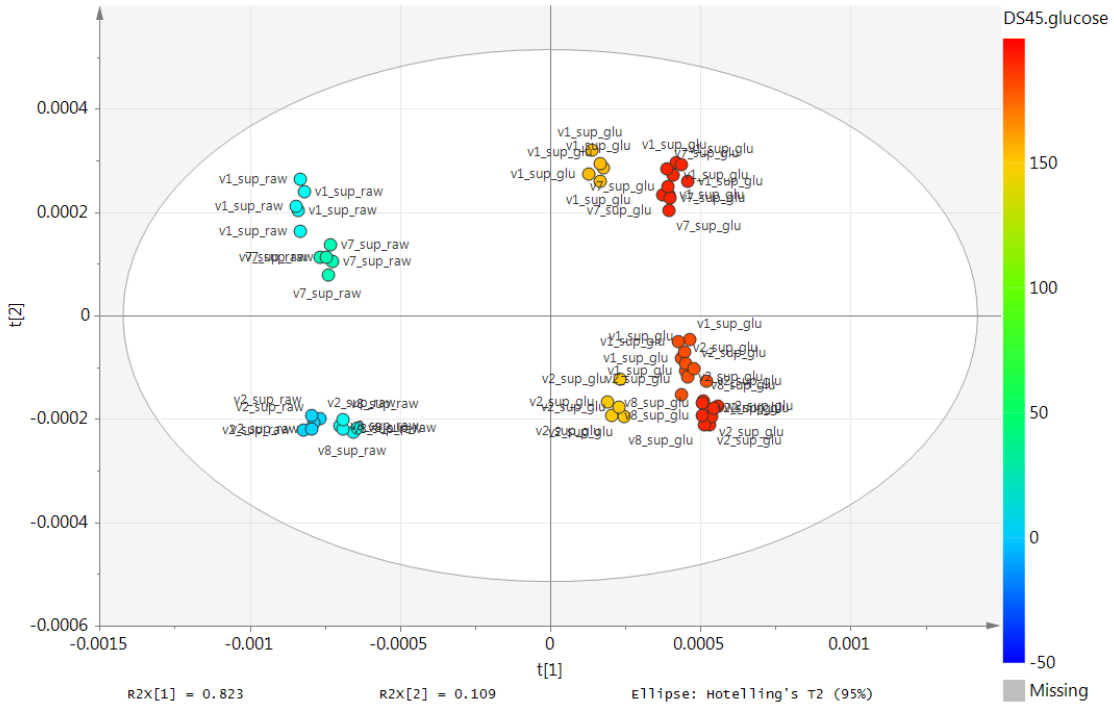

Figure 5.27: Scores plot of the PLS developed on mixtures from supernatant samples collected during run B.

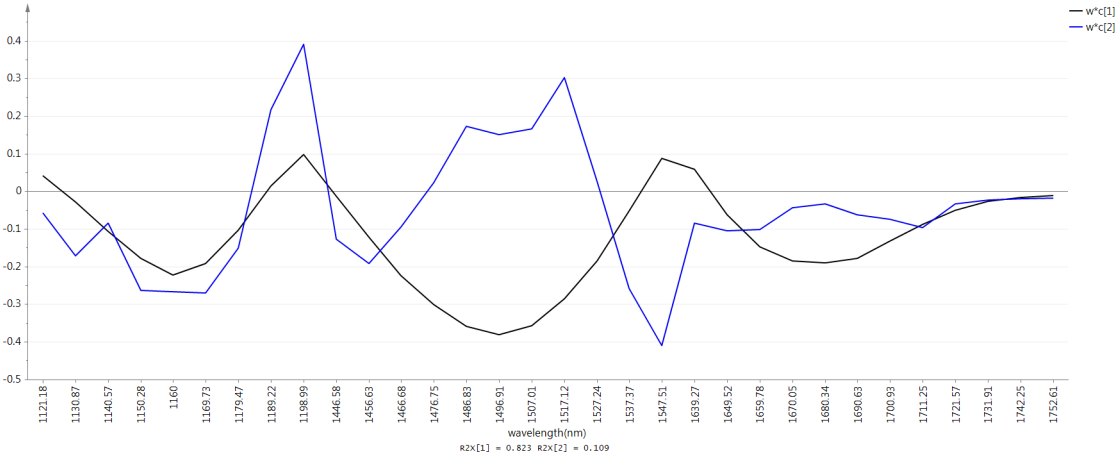

Figure 5.28: Loadings of the PLS developed on mixtures from supernatant samples collected during run B. Latent variable 1 (black line) and latent variable 2 (blue line).

Validation Runs

External validation of the model was done by using mixtures prepared in different days. The RMSEP and bias are listed in Table 5.15 and Table 5.16.

Table 5.15: Summary of statistics of predictions of mixtures prepared with supernatant from runA (raw supernatant and spiked "Gluc3").

	Raw samples	**Gluc3**
RMSEP (g/L)	6.79 (v1, v3, v7)	5.16 (v1, v3)
Bias (g/L)	**0.15** (v1, v3, v7)	1.96 (v1) -6.15 (v3)

Table 5.16: Summary of statistics of predictions of mixtures prepared with supernatant from runC (raw supernatant and spiked "Gluc1", "Gluc2" and "Gluc3").

	Raw samples	**Gluc 1**	**Gluc 2**	**Gluc3**
RMSEP (g/L)	10.6 (v1, v7, v2, v8) 6.86 (v1, v7, v2)	19.71 (for v2 and v8)	16.04 (for v1, v2, v8)	16.41 (for v1, v7, v2, v8, v9) 7.38 (for v8, v2, v7)
Bias	-0.626 (v1, v7, v2, v8) 5.14 (v1, v7, v2)	-7.73	-16.6	7.19 (for v1, v7, v2, v8, v9) -4.13 (for v8, v2, v7)

5.3.5 A Roadmap to Final Implementation

While Chapter 4 provided the groundwork for building a working automated liquid handler that implemented spectroscopic tools, Chapter 5 used it to get readings of cultivation samples. At the same time, both chapters described shortcomings and lessons for the future. With these lessons in mind, it is possible to give a description of a prototypical automated bioreactor system that incorporates spectroscopic tools.

For the studied system to be used daily by scientists in Research and Development, the first step would be to assemble the tested automated analysis system (AM, liquid handler, flow cell and spectrometer) in the Ambr® system. This could be challenging for lack of space in the current set-up. Alternatively, a larger laminar flow hood would have to be used. A proposal for the set-up is shown in Figure 5.29. Ideally, an appropriate filter for cells would also be integrated, allowing for manipulation of solely the supernatant, for which the errors of prediction of the analyte to be determined are lower. It is known that high cell densities might hinder spectral readings.

Specific software is required to control the Ambr® and the modified AM, to obtain and record data from the spectrometer, MODDE (from Sartorius, used for Design of Experiments) and eventually a PAT Management Software (such as SIPAT or SynTQ). A PLS model can be developed in SIMCA (Sartorius) or The Unscrambler (CAMO) and loaded into the Management Software. The collected spectra could be utilised and the predictions can be sent to an eventual screen (HMI, Human Machine Interface).

Ideally the developed models can be used for feedback control. Based on the prediction, made available by the PAT Management Software, the automation layer can take the adequate action.

As for the calibration strategy used, there are two possible scenarios. Scenario 1 focus on a situation where qualitative information would suffice. On the other hand, scenario 2 applies to a situation where more accurate readings are required. This thesis has shown that a PAT specialist would be required to perform scenario 1 as an appropriate model has to be designed and executed.

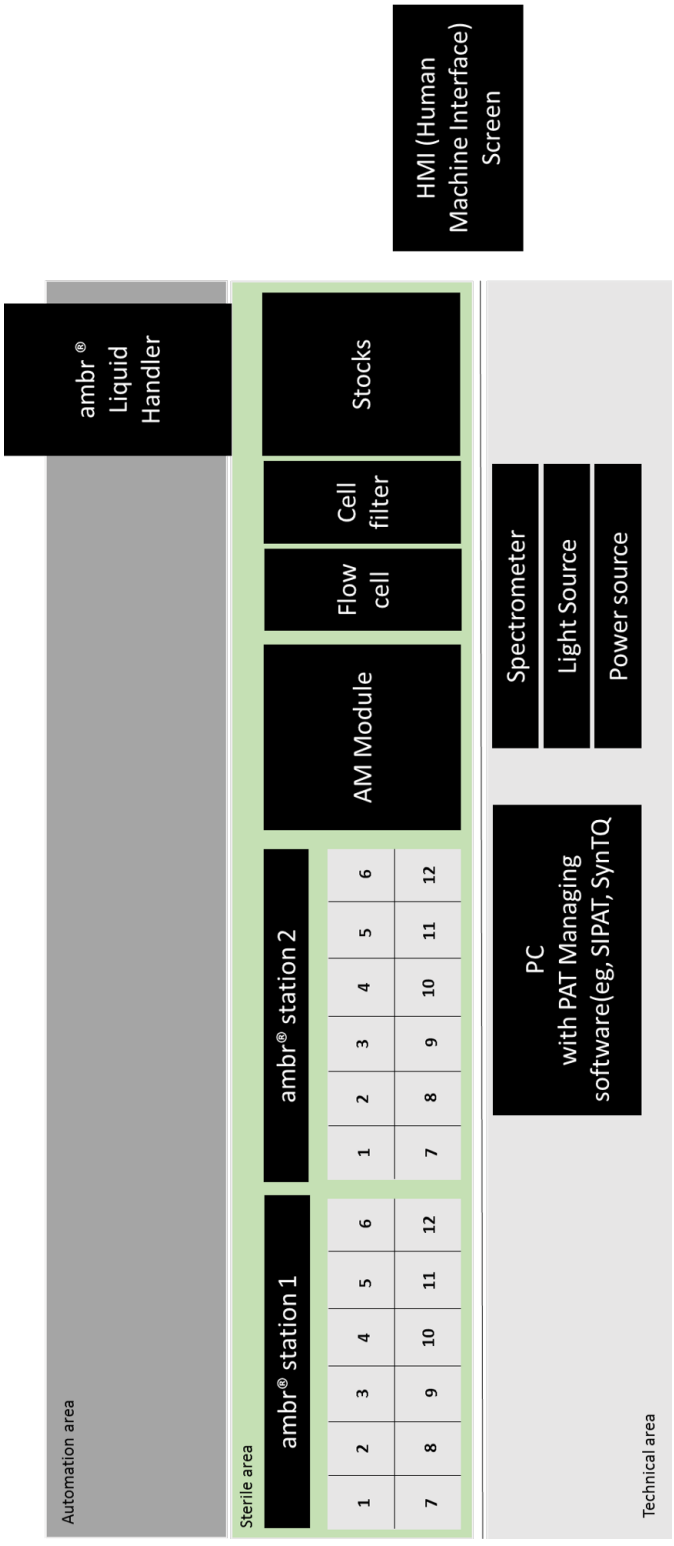

Figure 5.29: Potential configuration used for integration of spectroscopy in the Ambr.

5.3.5.1 Scenario 1 - a qualitative approach

For the first scenario, a calibration dataset based on synthetic standards can be developed in the AM (analysis module). The end-user would define the design space upfront, i.e, select two main analytes and the range of concentrations that will be used. The sterilised stocks of the two analytes, fresh media and a stock of cells, would have to be made available in the sterile area. The automation would then prepare the samples. During the fermentations, samples can be collected from the vessel, filtered and the spectral reading obtained in the flow-cell, results in a prediction for which the trend is still valid.

5.3.5.2 Scenario 2 - a quantitative approach

For the second scenario, the Ambr® should be used to run fermentations and then the supernatants obtained. This can be prepared to be done overnight, for example. The Ambr® system would be prepared with the same fresh medium as for the real runs, and a stock of sterilised glucose prepared and available in the same hood. Then, different concentrations of microorganism would be inoculated in each vessel. A possible approach would be:

1 - System starts with the built-in model for yeast and glucose.

2 - Standards of air are collected to obtain the absorbance spectrum

3 – Depending on the desired design space, standards of the desired analytes glucose, lactate in buffer at the correct temperature are used to correct the model from the previous status to the desired conditions (e.g. spectrometer status and temperature).

4 – Some of the samples can be collected and the concentration of glucose or lactate measured offline for confirmation.

5 – Prepare mixtures with the liquid-handler to extend the calibration dataset.

6 - During the calibration run, the monitor (HMI) can display trends from the in-built model. For all following fermentations, levels of glucose from the adjusted calibrations and displayed in real-time.

7 – If the developed models are accurate, feedback control can take place and the

levels are adjusted appropriately.

5.4 Conclusions

This thesis aims to gain productivity and efficiency improvements in the early stages of bioprocess development. To that extent, the previous chapter developed a prototype of an automated NIR system that can save time and reduced human error significantly. This chapter applied that automated system to real cultivations using the Ambr® bioreactors.

The real cultivation samples from the Ambr® systems were analysed through three separate methods. First, two models that were developed in the previous chapter were now challenged to predict the concentrations of glucose and yeast in real cultivations. Second, rather than using standards that were scanned months before, standards of glucose and yeast (prepared with the same raw materials and in the same time frame as the cultivation runs) were used for model development. These results were obtained at-line and using the spent medium collected from the samples after centrifugation. Third, rather than using pre-calibrated models based on simple mixtures of pure analytes, a new model was developed based on samples collected from the actual cultivation.

Method one, that pre-calibrated a model based on different standards, yielded errors that were outside of an acceptable scope. When using current standards, the results were significantly better, particularly for yeast (0.50 g/L). Following a correction for the glucose prediction, the errors of prediction were also highly reduced (4 g/L). The best predictions, however, were obtained by the third method, where a new model was developed on samples collected from the actual cultivation. Predictions of samples of supernatants varied between 5.16 and 16 g/L, as seen in Tables 5.15 and 5.16.

Various reasons can be thought of why initial predictions are outside of an acceptable margin of accuracy. The spectra were used as "counts", instead of corrected absorbance (as per Eq. 4.5) which means that there was no correction applied for potential differences in e.g. background, temperature, state of the instrument. Perhaps a model transfer such as PDS (piecewise direct standardisation) would have been able to correct for such differences.

To improve these results, an approach was proposed in which the predictions of the

two individual PLS models (for yeast and glucose concentration) are used to correct the prediction of glucose concentration. Second, the concentration of glucose in the samples collected from the bioreactors could have been measured through, for example, a YSI analyser (YSI2700 Biochemistry Analyzer, YSI Life Sciences) or HPLC. This would yield a better reference for model development. The advantage of the used strips was the practicality of obtaining the result at the moment of samples collection, it only requires about 20µL of sample which makes them suitable for the small volume collected from each minibioreactor. The reference concentrations can therefore be subjected to some error.

Regardless of these limitations, model development on the actual cultivations used for the predictions has the highest chance of success. The downside might be that a PAT specialist or a chemometrician may have to assist in the model development. In other cases, however, the pre-calibrated models can still supply valid information and save significant time, in the least by providing qualitative information on the trends of the analytes during cultivation.

This chapter's final contribution comes in the form of a roadmap for the implementation of spectroscopy and automation into the Ambr bioreactors. Two possible scenarios were suggested: qualitative and quantitative, with the latter unlocking the potential of feedback control of analytes. This would represent a major step into a fully self-controlled bioreactor system leading to further productivity improvements.

Chapter 6

Conclusions and Latest Developments

This project set out to develop novel approaches for the integration of NIR measurements in early process development of biopharmaceuticals. With this purpose, automated minibioreactors were used (as they expand the possibilities of design of experiments and subsequently increase knowledge of the process at early stages), available NIR instruments were used (more affordable than other spectroscopic systems), new methodologies were developed. This culminated in the first-ever prototype that combines an automated liquid-handler with spectroscopy. Several multivariate methods were constructed for the quantification of key analyte concentrations and an approach to obtain better predictions was also proposed.

Throughout this thesis several attempts of using models to predict samples outside the calibration range, were made. Even though this is an uncommon approach, the goal was to evaluate the qualitative information obtained with this method and its applicability in a fermentation environment.

After introducing the field with the Literature Review in Chapter 2, Chapter 3 shows a first approach to data, focusing on sources of variability involved in a fermentation matrix, while using different NIR instruments. Both glucose and lactate were modelled due to their importance in fermentations. The highest errors of cross validation (RMSEcv) were 0.839 g/L for lactate (obtained on spent media) and 2.45 g/L for glucose (obtained in the presence of cells). This is largely due to increased complexity of the background and increased variance in the spectra, which in turn also reduces the contribution from the analyte of interest. Models predicting glucose concentrations yielded no satisfactory results, besides the prediction from the samples in water. The stronger effects of lactate on spectral features at the level of concentrations used might be one of the reasons.

Chapter 4 described work developed during the 6 months internship at Sartorius Stedim and the development of a system from scratch. For this purpose, an existing automated liquid handler was adapted together with a prototype flow-cell and a prototype diode-array spectrometer. The ability of this system to accurately prepare and measure mixtures of key analytes in fermentation was assessed through an extensive ex-

ploratory data analysis on binary and ternary mixtures. This relatively simple analysis provided a large wealth of data. It was shown that initial results were highly replicable, with low variation between them. In addition, within the same datasets, predictions were relatively accurate. It proved harder, however, to extrapolate models from one sample to another one.

Various reasons for this were explained in that chapter, relating to sample selection, the spectrometer, the spectra and the pathlength. However, despite the relatively high error of predictions, the predicted trends were valid. This gives useful qualitative information that can be used when there is no need for tight control of the analyte levels. This information is valuable for the end-user, and was applied in the following chapter to predict cultivation samples.

In Chapter 5, yeast fermentations conducted in the Ambr® bioreactors were described and the predictive ability of the models developed in Chapter 4 was tested. These pre-calibrated models on glucose and yeast standards failed on the prediction of complex fermentation matrices. Therefore, a different calibration dataset was selected. For these new models, not only glucose and yeast's fresh standards were used, but also samples of fresh medium and water samples, which resulted in lower errors of prediction of real fermentation samples. Particularly, the yeast model yielded good results, with a RMSEP of approximately 0.5g/L for two different runs (B and C).

Because the model for glucose prediction was still not applicable, an extra correction step was applied. A reduction of approximately 5 times of the error of prediction was obtained for a test dataset by applying the proposed correction, that makes use of both predictions from the PLS model for glucose and the PLS model for yeast.

Further to this, a final model for glucose on the mixtures of supernatant samples was developed. Rather than using pre-calibrated models based on simple mixtures of pure analytes, a new model was developed based on samples collected from the actual cultivation. This yielded the best results, suggesting that pre-calibration may only be desirable in specific cases. In more complicated cases, however, a PAT specialist or chemometrician may still have to intervene to build a model from the cultivation samples at hand.

Here too, there were various lessons drawn on how to potentially improve the readings. First, additional corrections might be required to match datasets collected at different time points. Second, a better method for reference analysis of glucose could have been used. Third, an approach was proposed in which the predictions of the two individual PLS models (for yeast and glucose concentration) are used to correct the prediction of glucose concentration.

In the end, these lessons culminated in a roadmap for the implementation of the developed automated spectroscopic system in the Ambr bioreactors. While this thesis focused on NIR spectroscopy, the project made parallel strides using Raman, as discussed in the next section.

6.1 Latest Developments

This thesis set out to investigate the use of spectroscopic methods in the early stages of developments. Starting in 2013, it was part of a larger project within Sartorius to expedite the biopharmaceutical development and benefit from the capabilities of spectroscopy. As the previous chapter concluded, the prospects of NIR being the main spectroscopic tool to fulfil that, might seem limited. However, a lot has happened in the seven years that have passed since then. Most importantly, Sartorius have recently (February, 2020) released a new commercial tool that finally unlocks the potential of spectroscopy into the early stages of bioprocess development. Rather than using NIR to do so, this tool, named BioPAT Spectro which is displayed in Figures 6.1 and 6.2, uses Raman technology from Kaiser Optical Systems or Tornado Spectral Systems to do so. In the context of the insights provided by this thesis, this chapter takes a closer look at that and other recent developments in the field of Bioprocess Monitoring.

In the beginning of this project, Raman spectroscopy was not as developed as it is now. The high cost of the instrument, possible fluorescence of some components of the medium, the possible length of the optical fibres, were some examples of limitations to this technique. However, manufacturers of Raman instruments have optimized their solutions for bioprocess applications and developed specific probes for the application in

Figure 6.1: The Ambr® system with BioPat Spectro using Raman from Kaiser Optical Systems

Figure 6.2: The commercial version of the AM - Analysis Module.

flow-cells (such as bIO-PRO from Kaiser[1] or MarqMetrix[2]), or in single-use applications (such as ProCellics from RESOLUTION Spectra Systems[3]).

In addition, Chapter 2 already mentioned the advantage of Raman being insensitive to water. Furthermore, given the issues of NIR in dealing with water as shown by this thesis, Raman's water-insensitivity has proven to be more of an advantage than initially anticipated. For these reasons, Raman was eventually chosen to be the best option for process development as well as commercial manufacturing. As shown in this thesis, BioPAT Spectro can allow for easier and faster model building, high-throughput process development with spectroscopy. With such tool, all spectral and process data can be used in a SIMCA-ready file for model building. Some of the cited advantages are:

- Ambr® derived (Raman) models are more robust due to the use of all process data, a large DoE design space, and automated spiking of Ambr® samples with analyte stock solutions;
- Ambr® can use SIMCA® models to predict analyte concentrations and execute process control in real time.

Note, however, that none of this is to say that NIR will no longer be employed in the future. Instead, as shown in this thesis, NIR might still yield useful results under certain conditions. In addition, a recent study by (Li et al., 2018) showed that while Raman spectroscopy led to slightly better estimations for mAb concentration, NIR spectroscopy showed a higher signal-to-noise ratio. As they concluded, "NIR spectroscopy remains a potential candidate, as further improvements of its real-time prediction ability could be expected from investigation of other chemometric methods, such as non-linear regression methods, and by using additional spectral data" (Li et al., 2018, pp.212). Trunfio et al. (2017) also concluded that NIR generated lower prediction errors than MIR and Raman. Finally, even when Raman outperforms NIR, Kozma et al. (2017) has shown that NIR can still provide valid trends, which is confirmed in this thesis.

[1] https://kosi.com/products/kaiser-raman-probes/bio-pro-probe-ks-785nm
[2] https://www.marqmetrix.com/products/flowcell/
[3] https://resolutionspectra.com/procellics-first-line-real-time-bioprocess-raman-analyzer/

All in all, NIR and Raman are complementary since they provide information about different analytes. Therefore, to have both techniques measuring the same sample would also be a possibility. As spectroscopy is non-destructive, this is feasible, and using the MVDA techniques from SIMCA one could get the best of both worlds.

Besides the actual spectroscopic tools, supporting tools for the BioPat Spectro solution are also arising, such as new filters or automated analysers for reference tests. For example, Nova Biomedical released the BioProfile FLEX2 Online autosampler. This technology is based on MicroSensor Card technology with optical measurement and freezing point osmometry and it can test for Gluc, Lac, Gln, Glu, NH_4^+, Na^+, K^+, Ca^{++}, pH, PCO_2, total cell density, viable cell density, viability, cell diameter, and osmolality. When an OPC-compatible control system is used, this tool can provide real-time analysis and feedback control of all measured parameters.

Developments like these can thrive in the current era of Industry 4.0 where automation is taken to ever higher levels. It involves cyber systems that consist of smart factories where the units of manufacturing industry interact, share information, and make adaptive decisions without human intervention. In combination with Quality by Design approaches allow for significant reductions in the development of drugs.

With the current onset of the Coronavirus (Covid-19) pandemic, the need for Quality by Design in order to allow a fast, cheap development of drugs is vital. The current average time span between preclinical studies and the approval of a new product is 12 years (Morgan et al., 2011). Likewise, the price is a challenge as the life cycle of a new pharmaceutical product can be estimated with different assumptions but the average ranges from US$161 million (2009) to US$2.87 billion (2013) (Morgan et al., 2011). Concurrently, big pharmaceutical companies are losing their main sources of income as patent of important products expire, which creates pressure for more optimized processes, while biosimilars join the market (Steinwandter et al., 2019). A quality by design approach, as well as the implementation of Industry 4.0, would allow for the much needed reduction of time-to-market of a new drug. With the onset of a global pandemic looming, the world is in desperate need for these developments to come sooner, rather than later.

Bibliography

Abu-Absi, N. R., Kenty, B. M., Cuellar, M. E., Borys, M. C., Sakhamuri, S., Strachan, D. J., Hausladen, M. C. and Li, Z. J. (2011), 'Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe', *Biotechnology and Bioengineering* **108**(5), 1215–1221.

Acha, V., Meurens, M., Naveau, H. and Agathos, S. N. (2000), 'ATR-FTIR sensor development for continuous on-line monitoring of chlorinated aliphatic hydrocarbons in a fixed-bed bioreactor.', *Biotechnology and Bioengineering* **68**(5), 473–87.

Åkesson, M., Karlsson, E. N., Hagander, P., Axelsson, J. P. and Tocaj, A. (1999), 'On-Line Detection of Acetate Formation in *Escherichia coli* Cultures Using Feed Transients', *Biotechnology* **64**(5), 590–598.

Arnold, S. A., Crowley, J., Vaidyanathan, S., Matheson, L., Mohan, P., Hall, J. W., Harvey, L. M. and McNeil, B. (2000), 'At-line monitoring of a submerged filamentous bacterial cultivation using near-infrared spectroscopy', *Enzyme and Microbial Technology* **27**(9), 691–697.

Arnold, S. A., Crowley, J., Woods, N., Harvey, L. M. and McNeil, B. (2003), 'In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation.', *Biotechnology and Bioengineering* **84**, 13–9.

Arnold, S. A., Gaensakoo, R., Harvey, L. M. and McNeil, B. (2002), 'Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch *Escherichia coli* process', *Biotechnology and Bioengineering* **80**(4), 405–413.

Bakeev, K. (2010), *Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries*, first edn, Blackwell Publishing.

Bech Jensen, E. and Carlsen, S. (1990), 'Production of recombinant human growth hormone in *Escherichia coli*: expression of different precursors and physiological effects of glucose, acetate, and salts', *Biotechnology and Bioengineering* **36**, 1–11.

Berlec, A. and Strukelj, B. (2013), 'Current state and recent advances in biopharmaceutical production in *Escherichia coli*, yeasts and mammalian cells.', *Journal of Industrial Microbiology & Biotechnology* **40**(3-4), 257–74.

Berry, B. N., Dobrowsky, T. M., Timson, R. C., Kshirsagar, R., Ryll, T. and Wiltberger, K. (2016), 'Quick Generation of Raman Spectroscopy Based In-Process Glucose Control to Influence Biopharmaceutical Protein Product Quality During Mammalian Cell Culture', *Biotechnology Progress* **32**(1), 224 – 234.

BioPlan (2017), 14th Annual Report and Survey of Biopharmaceutical Manufacturing - A Study of Biotherapeutic Developers and Contract Manufacturing Organizations, Technical report, BioPlan Associates, Inc.

Blank, T. B., Sum, S. T., Brown, S. D. and Monfre, S. L. (1996), 'Transfer of Near-Infrared Multivariate Calibrations without Standards.', *Analytical chemistry* **68**(17), 2987–2995.

Brereton, R. G. (2003), *Chemometrics: data analysis for the laboratory and chemical plant*, Wiley.

Buckley, K. and Ryder, A. G. (2017), 'Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing : A Short Review', *Applied Spectroscopy* **71**(6), 1085–1116.

Cavinato, A. G., Mayes, D. M., Ge, Z. and Callis, J. B. (1990), 'Noninvasive method for monitoring ethanol in fermentation processes using fiber-optic near-infrared spectroscopy', *Analytical Chemistry* **62**(18), 1977–1982.

Cereghino, J. L. and Cregg, J. M. (2000), 'Heterologous protein expression in the methylotrophic yeast Pichia pastoris', *FEMS Microbiology Reviews* **24**(1), 45–66.

Cervera, A. E., Petersen, N., Lantz, A. E., Larsen, A. and Gernaey, K. V. (2009), 'Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation.', *Biotechnology Progress* **25**(6), 1561–81.

Chang, H. N., Jung, K., Choi, J.-D.-R., Lee, J. C. and Woo, H.-C. (2014), 'Multi-stage continuous high cell density culture systems: a review.', *Biotechnology Advances* **32**(2), 514–25.

Chen, J., Arnold, M. a. and Small, G. W. (2004), 'Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions', *Analytical Chemistry* **76**(18), 5405–5413.

Chen, M., Khare, S., Huang, B., Zhang, H., Lau, E. and Feng, E. (2013), 'Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application', *Industrial and Engineering Chemistry Research* **52**(23), 7886–7895.

Chen, Z.-P., Zhong, L.-J., Nordon, A., Littlejohn, D., Holden, M., Fazenda, M., Harvey, L., McNeil, B., Faulkner, J. and Morris, J. (2011), 'Calibration of multiplexed fiber-optic spectroscopy.', *Analytical Chemistry* **83**(7), 2655–9.

Chiruvolu, V., Cregg, J. and Meagher, M. (1997), 'Recombinant protein production in an alcohol oxidase-defective strain of *Pichia pastoris* in fedbatch fermentations', *Enzyme and Microbial Technology* **21**, 277–283.

Choi, J. H., Keum, K. C. and Lee, S. Y. (2006), 'Production of recombinant proteins by high cell density culture of *Escherichia coli*', *Chemical Engineering Science* **61**(3), 876–885.

Classen, J., Aupert, F., Reardon, K. F., Solle, D. and Scheper, T. (2017), 'Spectroscopic sensors for in-line bioprocess monitoring in research and pharmaceutical industrial application', *Analytical and Bioanalytical Chemistry* **409**(3), 651–666.

Clavaud, M., Roggo, Y., Von Daeniken, R., Liebler, A. and Schwabe, J.-O. (2013), 'Chemometrics and in-line near infrared spectroscopic monitoring of a biopharmaceutical chinese hamster ovary cell culture: prediction of multiple cultivation variables', *Talanta* **111**, 28–38.

Coggins, E. F., Gross, A., Johnston, R., Lambalot, R., Otero, J. M., Rader, R. A., Roberts, R. L., Repetto, R., Ultee, M. E., Vogel, J. D., Wajid, A. and Wheelwright, S. M. (2012), 9th Annual Report and Survey of Biopharmaceutical Manufacturing Capacity and Production, Technical Report April, BioPlan Associates, Inc.

Corro-Herrera, V. A., Gomez-Rodriguez, J., Hayward-Jones, P. M., Barradas-Dermitz, D. M. and Aguilar-Uscanga, M. G. (2016), 'In-Situ Monitoring of Saccharomyces Cerevisiae ITV01 Bioethanol Process Using Near-Infrared Spectroscopy NIRS and Chemometrics', *American Institute of Chemical Engineers* **32**(2), 510–517.

Cos, O., Ramón, R., Montesinos, J. L. and Valero, F. (2006), 'Operational strategies, monitoring and control of heterologous protein production in the methylotrophic yeast *Pichia pastoris* under different promoters: a review.', *Microbial Cell Factories* **5**, 17.

Crowley, J., Arnold, S. A., Wood, N., Harvey, L. M. and McNeil, B. (2005), 'Monitoring a high cell density recombinant Pichia pastoris fed-batch bioprocess using transmission and reflectance near infrared spectroscopy', *Enzyme and Microbial Technology* **36**(5-6), 621–628.

Datta, P., Linhardt, R. J. and Sharfstein, S. T. (2013), 'An 'omics approach towards CHO cell engineering.', *Biotechnology and Bioengineering* **110**(5), 1255–71.

De Schutter, K., Lin, Y.-C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouzé, P., Van de Peer, Y. and Callewaert, N. (2009), 'Genome sequence of the recombinant protein production host *Pichia pastoris*.', *Nature Biotechnology* **27**(6), 561–6.

Deloitte (2018), 2018 Global life sciences outlook - Innovating life sciences in the fourth industrial revolution: Embrace, build, grow, Technical report.

Doak, D. and Phillips, J. (1999), 'In situ monitoring of an *Escherichia coli* fermentation using a diamond composition ATR probe and mid-infrared spectroscopy', *Biotechnology Progress* **15**(3), 529–39.

Eiteman, M. A. and Altman, E. (2006), 'Overcoming acetate in *Escherichia coli* recombinant protein fermentations.', *Trends in Biotechnology* **24**(11), 530–6.

Esbensen, K., Guyot, D., Westad, F. and Houmoller, L. (2002), *Multivariate Data Analysis: In Practice : an Introduction to Multivariate Data Analysis and Experimental Design*, CAMO.

Esmonde-White, K. A., Cuellar, M., Uerpmann, C., Lenain, B. and Lewis, I. R. (2017), 'Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing', *Analytical and Bioanalytical Chemistry* pp. 637–649.

Fazenda, M. L., Dias, J. M., Harvey, L. M., Nordon, A., Edraba-Ebel, R., LittleJohn, D. and McNeil, B. (2013), 'Towards better understanding of an industrial cell factory: investigating the feasibility of real-time metabolic flux analysis in Pichia pastoris', *Microbial Cell Factories* **12**(51), 1–14.

FDA (2004), 'PAT Guidance for Industry—A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance'.

FDA (2012), 'Guidance for Industry - Q11 Development and Substances', *ICH* .

Fearn, T., Riccioli, C., Garrido-Varo, A. and Guerrero-Ginel, J. E. (2009), 'On the geometry of SNV and MSC', *Chemometrics and Intelligent Laboratory Systems* **96**(1), 22–26.

Finn, B., Harvey, L. M. and McNeil, B. (2006), 'Near-infrared spectroscopic monitoring of biomass, glucose, ethanol and protein content in a high cell density baker's yeast fed-batch bioprocess.', *Yeast (Chichester, England)* **23**(7), 507–17.

Funke, M., Buchenauer, A., Mokwa, W., Kluge, S., Hein, L., Müller, C., Kensy, F. and Büchs, J. (2010), 'Bioprocess control in microscale: scalable fermentations in disposable and user-friendly microfluidic systems.', *Microbial cell factories* **9**(1), 86.

Ge, Z., Cavinato, A. G. and Callis, J. B. (1994), 'Noninvasive Spectroscopy for Monitoring Cell Density in a Fermentation Process', *Analytical Chemistry* **66**(8), 1354–1362.

Hall, J. W., McNeil, B., Rollins, M. J., Draper, I., Thompson, B. G. and Macaloney, G. (1996), 'Near-Infrared Spectroscopic Determination of Acetate, Ammonium, Biomass, and Glycerol in an Industrial <I>Escherichia coli</I> Fermentation', *Applied Spectroscopy* **50**(1), 102–108.

Hamilton, S. R., Davidson, R. C., Sethuraman, N., Nett, J. H., Jiang, Y., Rios, S., Bobrowicz, P., Stadheim, T. a., Li, H., Choi, B.-K., Hopkins, D., Wischnewski, H., Roser, J., Mitchell, T., Strawbridge, R. R., Hoopes, J., Wildt, S. and Gerngross, T. U. (2006), 'Humanization of yeast to produce complex terminally sialylated glycoproteins.', *Science (New York, N.Y.)* **313**(5792), 1441–3.

Henriques, J. G., Buziol, S., Stocker, E., Voogd, A. and Menezes, J. C. (2009), Monitoring Mammalian Cell Cultivations for Monoclonal Antibody Production Using Near-Infrared Spectroscopy, *in* 'Optical Sensor Systems in Biotechnology', Vol. 116, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 29–72.

Hong, M. S., Severson, K. A., Jiang, M., Lu, A. E., Love, J. C. and Braatz, R. D. (2018), 'Challenges and opportunities in biopharmaceutical manufacturing control', *Computers and Chemical Engineering* **110**, 106–114.

Huang, J., Romero-Torres, S. and Moshgbar, M. (2010), 'Practical considerations in data pre-treatment for NIR and Raman spectroscopy', *American Pharmaceutical Review* **13**, 116–127.

Iuchi, S. and Weiner, L. (1996), 'Cellular and molecular physiology of *Escherichia coli* in the adaptation to aerobic environments', *Journal of Biochemistry* **120**(6), 1055–1063.

Jamrógiewicz, M. (2012), 'Application of the near-infrared spectroscopy in the pharmaceutical technology.', *Journal of Pharmaceutical and Biomedical Analysis* **66**, 1–10.

Jayapal, K. P., Wlaschin, K. F., Hu, W.-S. and Yap, M. G. S. (2007), Recombinant protein therapeutics from CHO cells-20 years and counting, Technical report, Society for Biological Engineers' (SBE) - CHO Consortium.

Jensen, P. S., Bak, J. and Andersson-Engels, S. (2003), 'Influence of Temperature on Water and Aqueous Glucose Absorption Spectra in the Near- and Mid-Infrared Regions at Physiologically Relevant Temperatures', *Applied Spectroscopy* **57**(1), 28–36.

Jenzsch, M., Bell, C., Buziol, S., Kepert, F., Wegele, H. and Hakemeyer, C. (2017), Trends in Process Analytical Technology : Present State in Bioprocessing, *in* 'Adv Biochem Eng Biotechnol', Springer International Publishing AG 2017.

Jestel, N. L. (2005), *Process Raman Spectroscopy*, 1st edn, Blackwell Publishing Ltd, Oxford, UK.

Jin, J.-W., Chen, Z.-P., Li, L.-M., Steponavicius, R., Thennadil, S. N., Yang, J. and Yu, R.-q. (2012), 'Quantitative Spectroscopic Analysis of Heterogeneous Mixtures: The Correction of Multiplicative Effects Caused by Variations in Physical Properties of Samples', *Analytical Chemistry* **84**(1), 320–326.

Julien, C. (2006), 'Production of humanlike recombinant proteins in Pichia pastoris', *BioProcess International* pp. 22–31.

Jungo, C., Marison, I. and von Stockar, U. (2007), 'Mixed feeds of glycerol and methanol can improve the performance of Pichia pastoris cultures: A quantitative study based on concentration gradients in transient continuous cultures.', *Journal of Biotechnology* **128**(4), 824–37.

Karbalaei, M., Rezaee, S. A. and Farsiani, H. (2020), 'Pichia pastoris: A highly successful expression system for optimal synthesis of heterologous proteins', *J Cell Physiol.* **235**, 5867–5881.

Kim, J. Y., Kim, Y.-G. and Lee, G. M. (2012), 'CHO cells in biotechnology for pro-

duction of recombinant proteins: current state and further potential.', *Applied Microbiology and Biotechnology* **93**(3), 917–30.

Kneipp, K., Kneipp, H., Itzkan, I., Dasari, R. R. and Feld, M. S. (1999), 'Ultrasensitive chemical analysis by Raman spectroscopy.', *Chemical Reviews* **99**(10), 2957–2975.

Kozma, B., Hirsch, E., Gergely, S., Párta, L., Pataki, H. and Salgó, A. (2017), 'On-line prediction of the glucose concentration of CHO cell cultivations by NIR and Raman spectroscopy: Comparative scalability test with a shake flask model system', *Journal of Pharmaceutical and Biomedical Analysis* **145**, 346–355.

Kusterer, A., Krause, C., Kaufmann, K., Arnold, M. and Weuster-botz, D. (2008), 'Fully automated single-use stirred-tank bioreactors for parallel microbial cultivations', *Bioprocess and biosystems Engineering* **31**, 207–215.

Langer, E. S. (2017), 'Current Trends in Analytical Testing - Improved analytical methods continues to be a current need for nearly every aspect of biopharma manufacturing'.

Lee, S. Y. (1996), 'High cell-density culture of *Escherichia coli*.', *Trends in Biotechnology* **14**(3), 98–105.

Li, M., Ebel, B., Chauchard, F., Guédon, E. and Marc, A. (2018), 'Parallel comparison of in situ Raman and NIR spectroscopies to simultaneously measure multiple variables toward real-time monitoring of CHO cell bioreactor cultures', *Biochemical Engineering Journal* **137**, 205–213.

Lourenço, N. D., Lopes, J. A., Almeida, C. F., Sarraguça, M. C. and Pinheiro, H. M. (2012), 'Bioreactor monitoring with spectroscopy and chemometrics: a review.', *Analytical and Bioanalytical Chemistry* **404**(4), 1211–37.

Luypaert, J., Massart, D. L. and Vander Heyden, Y. (2007), 'Near-infrared spectroscopy applications in pharmaceutical analysis.', *Talanta* **72**(3), 865–83.

Macaloney, G., Hall, J., Rollins, M., Draper, I., Anderson, K., Preston, J., Thompson, B. and McNeil, B. (1997), 'The utility and performance of near-infra red spectroscopy

in simultaneous monitoring of multiple components in a high cell density recombinant *Escherichia coli* production process', *Bioprocess Engineering* **17**(3), 157–167.

Macaloney, G., Hall, J., Rollins, M., Draper, I., Thompson, B. and McNeil, B. (1994), 'Monitoring biomass and glycerol in an *Escherichia coli* fermentation using near-infrared spectroscopy', *Biotechnology Techniques* **8**(4), 281–286.

Macauley-Patrick, S., Fazenda, M. L., McNeil, B. and Harvey, L. (2005), 'Heterologous protein production using the *Pichia pastoris* expression system', *Yeast* **22**(4), 249–70.

Makarova, E., Han, Y., Ringel, M. and Telpis, V. (2019), 'Digitization, automation, and online testing: The future of pharma quality control', *McKinsey* .

Massart, D., Vandeginste, B., Deming, S., Michotte, Y. and Kaufman, L. (1988), 'Chemometrics: A textbook', *Elsevier* **2**, 500.

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C. and Greyson, D. (2011), 'The cost of drug development : A systematic review', *Health policy* **100**(1), 4–17.

Morton, C. L. and Potter, P. M. (2000), 'Comparison of Escherichia coli, Saccharomyces cerevisiae, Pichia pastoris, Spodoptera frugiperda, and COS7 Cells for Recombinant Gene Expression', *Molecular Biotechnology* **16**, 193–202.

Musmann, C., Joeris, K. and Markert, S. (2016), 'Spectroscopic methods and their applicability for high-throughput characterization of mammalian cell cultures in automated cell culture systems', *Engineering in Life Sciences* **16**, 405–416.

Naes, T., Isakson, T., Fearn, T. and Davies, T. (2002), *A user-friendly guide to multivariate calibration and classification*, Vol. 17, NIR Publications, Chichester.

Nakano, K., Rischke, M., Sato, S. and Märkl, H. (1997), 'Influence of acetic acid on the growth of *Escherichia coli* K12 during high-cell-density cultivation in a dialysis reactor', *Applied Microbiology and Biotechnology* **48**(5), 597–601.

Niu, H., Jost, L., Pirlot, N., Sassi, H., Daukandt, M., Rodriguez, C. and Fickers, P. (2013), 'A quantitative study of methanol/sorbitol co-feeding process of a Pichia pastoris Mut+/pAOX1-lacZ strain.', *Microbial Cell Factories* **12**(1), 33.

Ogata, K., Nishikawa, H. and Ohsugi, M. (1969), 'A yeast capable of utilizing methanol', *Agricultural and Biological Chemistry* **33**(10).

Pan, J., Rhee, J. and Lebeault, J. (1987), 'Physiological constraints in increasing biomass concentration of *Escherichia coli* B in fed-batch culture', *Biotechnology Letters* **9**(2), 89–94.

Pons, M.-N., Le Bonté, S. and Potier, O. (2004), 'Spectral analysis and fingerprinting for biomedia characterisation.', *Journal of Biotechnology* **113**(1-3), 211–30.

Potvin, G., Ahmad, A. and Zhang, Z. (2012), 'Bioprocess engineering aspects of heterologous protein production in Pichia pastoris: A review', *Biochemical Engineering Journal* **64**, 91–105.

Reh, G. (2020), '2020 Global Life Sciences Outlook', *Deloitte Insights* .

Rhiel, M., Cohen, M. B., Murhammer, D. W. and Arnold, M. a. (2002), 'Nondestructive near-infrared spectroscopic measurement of multiple analytes in undiluted samples of serum-based cell culture media.', *Biotechnology and Bioengineering* **77**(1), 73–82.

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

Riley, M. R., Arnold, M. A. and Murhammer, D. W. (1998), 'Matrix-Enhanced Calibration Procedure for Multivariate Calibration Models with Near-Infrared Spectra', *Applied Spectroscopy* **52**(10), 1339–1347.

Riley, M. R., Crider, H. M., Nite, M. E., Garcia, R. a., Woo, J. and Wegge, R. M. (2001), 'Simultaneous measurement of 19 components in serum-containing animal cell culture media by fourier transform near-infrared spectroscopy.', *Biotechnology Progress* **17**(2), 376–8.

Riley, M. R., Rhiel, M., Zhou, X., Arnold, M. A. and Murhammer, D. W. (1997),

'Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy.', *Biotechnology and Bioengineering* **55**(1), 11–5.

Rinnan, Å., Van Den Berg, F. and Engelsen, S. B. (2009), 'Review of the most common pre-processing techniques for near-infrared spectra', *TrAC Trends in Analytical Chemistry* **28**(10), 1201–1222.

Rodrigues, L. O., Vieira, L., Cardoso, J. P. and Menezes, J. C. (2008), 'The use of NIR as a multi-parametric in situ monitoring technique in filamentous fermentation systems.', *Talanta* **75**(5), 1356–61.

Roussel, S. A., Igne, B., Funk, D. B. and Hurburgh, C. R. (2011), 'Noise robustness comparison for near infrared prediction models', *Journal of Near Infrared Spectroscopy* **36**(27 January 2011), 23–36.

Rowland-Jones, R. C., van den Berg, F., Racher, A. J., Martin, E. B. and Jaques, C. (2017), 'Comparison of Spectroscopy Technologies for Improved Monitoring of Cell Culture Processes in Miniature Bioreactors', *Biotechnology Progress* **33**(2), 337–346.

Roychoudhury, P., Harvey, L. M. and McNeil, B. (2006), 'The potential of mid infrared spectroscopy (MIRS) for real time bioprocess monitoring.', *Analytica Chimica Acta* **571**(2), 159–66.

Roychoudhury, P., O'Kennedy, R., McNeil, B. and Harvey, L. M. (2007), 'Multiplexing fibre optic near infrared (NIR) spectroscopy as an emerging technology to monitor industrial bioprocesses.', *Analytica Chimica Acta* **590**(1), 110–7.

Sampaio, P. N., Sales, K. C., Rosa, F. O., Lopes, M. B. and Calado, C. R. (2014), '*In situ* near infrared spectroscopy monitoring of cyprosin production by recombinant *Saccharomyces cerevisiae* strains', *Journal of Biotechnology* **188**, 148–157.

Scarff, M., Arnold, S., Harvey, L. M. and McNeil, B. (2006), 'Near infrared spectroscopy for bioprocess monitoring and control: current status and future trends', *Critical Reviews in Biotechnology* pp. 17–39.

Schenk, J., Marison, I. W. and von Stockar, U. (2007), 'A simple method to monitor and control methanol feeding of Pichia pastoris fermentations using mid-IR spectroscopy.', *Journal of Biotechnology* **128**(2), 344–53.

Schügerl, K. (2001), 'Progress in monitoring, modeling and control of bioprocesses during the last 20 years.', *Journal of Biotechnology* **85**(2), 149–73.

Sekhon, B. S. (2010), 'Biopharmaceuticals, an overview', *Thai Journal of Pharmaceutical Sciences* **34**, 1–19.

Siesler, H. W., Ozaki, Y., Kawata, S. and Heise, H. M. (2002), *Near-infrared spectroscopy. Principles, instruments, applications*, Vol. 16, Wiley-VCH, Weinheim.

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbuhler, K., Baiker, A., Tummala, S., Glennon, B., Kuentz, M., Steele, G., Kramer, H. J. M., Rydzak, J. W., Chen, Z., Morris, J., Kjell, F., Singh, R., Gani, R., Gernaey, K. V., Louhi-Kultanen, M., Oreilly, J., Sandler, N., Antikainen, O., Yliruusi, J., Frohberg, P., Ulrich, J., Braatz, R. D., Leyssens, T., Von Stosch, M., Oliveira, R., Tan, R. B. H., Wu, H., Khan, M., Ogrady, D., Pandey, A., Westra, R., Delle-Case, E., Pape, D., Angelosante, D., Maret, Y., Steiger, O., Lenner, M., Abbou-Oucherif, K., Nagy, Z. K., Litster, J. D., Kamaraju, V. K. and Chiu, M. S. (2015), 'Assessment of recent process analytical technology (PAT) trends: A multiauthor review', *Organic Process Research and Development* **19**(1), 3–62.

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

Sreekrishna, K., Brankamp, R. G., Kropp, K. E., Blankenship, D. T., Tsay, J.-T., Smith, P. L., Wierschke, J. D., Subramaniam, A. and Birkenberger, L. A. (1997), 'Strategies for optimal synthesis and secretion of heterologous proteins in the methylotrophic yeast *Pichia pastoris*', *Gene* **190**(1), 55–62.

Steinwandter, V., Borchert, D. and Herwig, C. (2019), 'Data science tools and applications on the way to Pharma 4.0', *Drug Discovery Today* **24**(9), 1795–1805.

Suresh, P. and Basu, P. K. (2008), 'Improving Pharmaceutical Product Development and Manufacturing: Impact on Cost of Drug Development and Cost of Goods Sold of Pharmaceuticals', *Journal of Pharmaceutical Innovation* **3**(3), 175–187.

Tamburini, E., Vaccari, G., Tosi, S. and Trilli, a. (2003), 'Near-infrared spectroscopy: a tool for monitoring submerged fermentation processes using an immersion optical-fiber probe.', *Applied spectroscopy* **57**(2), 132–8.

Thompson, B., Kole, M. and Gerson, D. F. (1985), 'Control of ammonium concentration in *Escherichia coli* fermentations', *Biotechnology and Bioengineering* **27**, 818–824.

Tran, A. M., Nguyen, T. T., Nguyen, C. T., Mai, X., Thi, H. and Nguyen, C. T. (2017), 'Pichia pastoris versus Saccharomyces cerevisiae: a case study on the recombinant production of human granulocyte - macrophage colony - stimulating factor', *BMC Research Notes* **10**, 6–13.

Trunfio, N., Lee, H., Starkey, J., Agarabi, C. and Liu, J. (2017), 'Characterization of Mammalian Cell Culture Raw Materials by Combining Spectroscopy and Chemometrics', *Biotechnology Progress* **33**(4), 1127–1138.

Ündey, C., Ertunç, S., Mistretta, T. and Looze, B. (2010), 'Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control', *Journal of Process Control* **20**(9), 1009–1018.

Vaidyanathan, S., Arnold, A., Matheson, L., Mohan, P., Macaloney, G., McNeil, B. and Harvey, L. M. (2000), 'Critical evaluation of models developed for monitoring an industrial submerged bioprocess for antibiotic production using near-infrared spectroscopy.', *Biotechnology Progress* **16**(6), 1098–105.

Vaidyanathan, S., Macaloney, G., Harvey, L. M. and McNeil, B. (2001), 'Assessment of the Structure and Predictive Ability of Models Developed for Monitoring Key Analytes in a Submerged Fungal Bioprocess Using Near-Infrared Spectroscopy', *Applied Spectroscopy* **55**(4), 444–453.

Vaidyanathan, S., Macaloney, G. and McNeil, B. (1999), 'Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring', *The Analyst* **124**(2), 157–162.

Velez-Suberbie, M. L., Betts, J. P. J., Walker, K. L., Robinson, C., Zoro, B. and Keshavarz-Moore, E. (2017), 'High Throughput Automated Microbial Bioreactor System Used for Clone Selection and Rapid Scale-down Process Optimization', *Biotechnology Progress* **34**(1), 58–68.

Vojinović, V., Cabral, J. and Fonseca, L. (2006), 'Real-time bioprocess monitoring', *Sensors and Actuators B: Chemical* **114**(2), 1083–1091.

Walsh, G. (2010), 'Biopharmaceutical benchmarks 2010.', *Nature Biotechnology* **28**(9), 917–24.

Walsh, G. (2014), 'Biopharmaceutical benchmarks 2014', *Nature biotechnology* **32**(7), 992–1000.

Warnes, M., Glassey, J., Montague, G. and Kara, B. (1996), 'On data-based modelling techniques for fermentation processes', *Process Biochemistry* **31**(2), 147–155.

Wentz, A. E. and Shusta, E. V. (2007), 'A novel high-throughput screen reveals yeast genes that increase secretion of heterologous proteins.', *Applied and Environmental Microbiology* **73**(4), 1189–98.

Weusthuis, R. A., Pronk, J. T., van den Broek, P. J. and van Dijken, J. P. (1994), 'Chemostat Cultivation as a Tool for Studies on Sugar Transport in Yeasts', *Microbiology Reviews* **58**(4), 616–630.

Wold, S., Sjöström, M. and Eriksson, L. (2001), 'PLS-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems* **58**(2), 109–130.

Workman, J. J. (2008), NIR Spectroscopy Calibration Basics, *in* D. A. Burns and E. W, eds, 'Handbook of Near-Infrared Analysis', pp. 123–150.

Wu, Z., Du, M., Xu, B., Lin, Z., Shi, X. and Qiao, Y. (2012), 'Absorption characteristics and quantitative contribution of overtones and combination of NIR: Method development and validation', *Journal of Molecular Structure* **1019**, 97–102.

Wülfert, F., Kok, W. T. and Smilde, a. K. (1998), 'Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models.', *Analytical chemistry* **70**(9), 1761–7.

Wurm, F. M. (2004), 'Production of recombinant protein therapeutics in cultivated mammalian cells.', *Nature Biotechnology* **22**(11), 1393–8.

Xu, P., Clark, C., Ryder, T., Sparks, C., Zhou, J., Wang, M., Russell, R. and Scott, C. (2017), 'Characterization of TAP Ambr 250 Disposable Bioreactors as a Reliable Scale-Down Model for Biologics Process Development', *Biotechnology Progress* **33**(2), 478–289.

Xu, X., Nagarajan, H., Lewis, N. E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M. R., Neff, N., Passarelli, B., Koh, W., Fan, H. C., Wang, J., Gui, Y., Lee, K. H., Betenbaugh, M. J., Quake, S. R., Famili, I., Palsson, B. O. and Wang, J. (2011), 'The genomic sequence of the Chinese hamster ovary (CHO) K1 cell line.', *Nature Biotechnology* **29**(8), 735–41.

Yu, F., Wadsworth, W. J. and Knight, J. C. (2012), 'Low loss silica hollow core fibers for 3–4 μm spectral region', **20**(10), 11153–11158.

Zhang, W., Inan, M. and Meagher, M. M. (2000), 'Fermentation strategies for recombinant protein expression in the methylotrophic yeast *Pichia pastoris*', *Biotechnology and Bioprocess Engineering* **5**(4), 275–287.